Chapter 3
Finite Element Methods for Parabolic Problems

The finite element methods are an alternative to the finite difference discretization of partial differential equations. The advantage of finite elements is that they give convergent deterministic approximations of option prices under realistic, low smoothness assumptions on the payoff function as, e.g. for binary contracts. The basis for finite element discretization of the pricing PDE is a *variational formulation* of the equation. Therefore, we introduce the Sobolev spaces needed in the variational formulation and give an abstract setting for the parabolic PDEs. All pricing equations satisfied by the price u of plain vanilla contracts which we encounter in these notes have the general form

$$\partial_t u - \mathcal{A}u = f, \quad \text{in } J \times G, \tag{3.1}$$

with the initial condition $u(0,x) = u_0(x)$, $x \in G$, a linear second order partial (integro-) differential operator $\mathcal{A}$, the forcing term f, the space domain $G \subseteq \mathbb{R}^d$ and the time interval $J = (0,T)$. The finite element approximation of the operators $\mathcal{A}$ fits into an abstract parabolic framework which we present next.

3.1 Sobolev Spaces

We now introduce some particular Hilbert spaces which are natural to use in the study of partial differential equations. These spaces consist of functions which are square integrable together with their partial derivatives up to a certain order.

Let $G = (a,b) \subset \mathbb{R}$ be an open, possibly unbounded domain and let first $u \in C^1(\overline{G})$. Integration by parts yields

$$\int_G u'\varphi \,\mathrm{d}x = -\int_G u\varphi' \,\mathrm{d}x, \quad \forall \varphi \in C_0^1(G).$$

N. Hilber et al., *Computational Methods for Quantitative Finance*, Springer Finance,
DOI 10.1007/978-3-642-35401-4_3,

If $u \in L^2(G)$, then u' does not necessarily exist in the classical sense, but we may define u' to be the linear functional

$$u^*(\varphi) = -\int_G u\varphi' \,\mathrm{d}x, \quad \forall \varphi \in C_0^1(G).$$

This functional is said to be a *generalized* or *weak derivative* of u. When u^* is bounded in $L^2(G)$, it follows from Riesz representation theorem (see Theorem 2.1.1) that there exists a unique function $w \in L^2(G)$ such that $u^*(\varphi) = (w, \varphi)$ for all $\varphi \in L^2(G)$, in particular

$$-\int_G u\varphi' \,\mathrm{d}x = \int_G w\varphi \,\mathrm{d}x, \quad \forall \varphi \in C_0^1(G).$$

We then say that the weak derivative belongs to $L^2(G)$ and write $u' = w$. In particular, if $u \in C^1(\overline{G})$, the generalized derivative u' coincides with the classical derivative u'. In a similar way, we can define weak derivatives $D^n u$ of higher order $n \in \mathbb{N}$.

Definition 3.1.1 The linear functional $D^n u$, $n \in \mathbb{N}$ is a *weak derivative* of u if

$$\int_G D^n u\varphi \,\mathrm{d}x = (-1)^n \int_G u D^n \varphi \,\mathrm{d}x, \quad \forall \varphi \in C_0^n(G).$$

We can now define the spaces $H^m(G)$.

Definition 3.1.2 Let $m \in \mathbb{N}$. $H^m(G)$ is the space of all functions whose weak partial derivatives of order $\leq m$ belong to $L^2(G)$, i.e.

$$H^m(G) = \{u \in L^2(G) : D^n u \in L^2(G) \text{ for } n \leq m\}.$$

We equip $H^m(G)$ with the inner product

$$(u, v)_{H^m(G)} = \sum_{n=0}^{m} (D^n u, D^n v)_{L^2(G)},$$

and the corresponding norm

$$\|u\|^2_{H^m(G)} = (u, u)_{H^m(G)} = \sum_{n=0}^{m} \|D^n u\|^2_{L^2(G)}.$$

We sometimes omit the (G) if the domain is clear from the context. $H^m(G)$ is complete and thus a Hilbert space. The space $H^m(G)$ is an example of a more general class of function spaces, called *Sobolev spaces*.

Definition 3.1.3 Let $p \in \mathbb{N} \cup \{\infty\}$. $W^{m,p}(G)$ is the space of all functions whose weak partial derivatives of order $\leq m$ belong to $L^p(G)$, i.e.

$$W^{m,p}(G) = \{u \in L^p(G) \ : \ D^n u \in L^p(G) \text{ for } n \leq m\}.$$

We equip $W^{m,p}(G)$ with the norm

$$\|u\|_{W^{m,p}(G)}^p = \sum_{n=0}^{m} \|D^n u\|_{L^p(G)}^p.$$

The normed space $W^{m,p}(G)$ is complete and hence a Banach space for $1 \le p \le \infty$. Functions $u \in W^{1,p}(G)$ are "essentially" continuous.

Theorem 3.1.4 *Let G be bounded and $u \in W^{1,p}(G)$. Then, there exists a continuous function $\widetilde{u} \in C^0(\overline{G})$ such that $u = \widetilde{u}$ a.e. on G and for all $x_1, x_2 \in \overline{G}$ there holds*

$$\widetilde{u}(x_2) - \widetilde{u}(x_1) = \int_{x_1}^{x_2} u'(\xi)\,\mathrm{d}\xi. \tag{3.2}$$

Proof Fix $y_0 \in G$ and set for any $g \in L^p(G)$,

$$v(x) := \int_{y_0}^{x} g(t)\,\mathrm{d}t, \quad x \in G.$$

Then, $v \in C^0(\overline{G})$ and

$$\begin{aligned}\int_G v\varphi'\,\mathrm{d}x &= \int_G \left(\int_{y_0}^{x} g(t)\,\mathrm{d}t\right)\varphi'(x)\,\mathrm{d}x \\ &= -\int_a^{y_0}\int_x^{y_0} g(t)\varphi'(x)\,\mathrm{d}t\,\mathrm{d}x + \int_{y_0}^{b}\int_{y_0}^{x} g(t)\varphi'(x)\,\mathrm{d}t\,\mathrm{d}x.\end{aligned}$$

Fubini's theorem implies, $\forall \varphi \in C_0^1(G)$,

$$\begin{aligned}\int_G v\varphi'\,\mathrm{d}x &= -\int_a^{y_0} g(t)\int_a^{t}\varphi'(x)\,\mathrm{d}x\,\mathrm{d}t + \int_{y_0}^{b} g(t)\int_t^{b}\varphi'(x)\,\mathrm{d}x\,\mathrm{d}t \\ &= -\int_G g(t)\varphi(t)\,\mathrm{d}t. \end{aligned}\tag{3.3}$$

We set $\overline{u}(x) := \int_{y_0}^{x} u'(\xi)\,\mathrm{d}\xi$. With (3.3) we obtain

$$\int_G \overline{u}\varphi'\,\mathrm{d}x = -\int_G u'\varphi\,\mathrm{d}x, \quad \forall \varphi \in C_0^1(G),$$

and hence with the definition of the weak derivative,

$$\int_G (u - \overline{u})\varphi'\,\mathrm{d}x = 0, \quad \forall \varphi \in C_0^1(G).$$

Therefore, it follows that for a.e. $x \in G$, we have $u(x) - \overline{u}(x) = C$. Putting $\widetilde{u} := \overline{u} + C$, we obtain the result. □

We will also need spaces with boundary conditions where we impose $u = 0$ on ∂G.

Definition 3.1.5 Let $1 \le p < \infty$. Then, $W_0^{1,p}$ is the closure of C_0^1 in the $W^{1,p}$-norm,

$$W_0^{1,p}(G) = \overline{C_0^1(G)}^{\|\cdot\|_{W^{1,p}(G)}}.$$

The space $W_0^{1,p}(G) \subset W^{1,p}(G)$ is a closed linear subspace. In particular, $H_0^1(G) := W_0^{1,2}(G)$ is again a Hilbert space with the norm $\|\cdot\|_{H^1(G)}$. We have the important *Poincaré inequality*.

Theorem 3.1.6 (Poincaré inequality) *Assume that $G \subset \mathbb{R}$ bounded. Then,*

(i) *There exists a constant $C(|G|, p) > 0$ such that*

$$\|u\|_{L^p(G)} \le C \|u'\|_{L^p(G)}, \quad \forall u \in W_0^{1,p}(G). \tag{3.4}$$

(ii) *Define*

$$W_*^{1,p}(G) := \left\{ u \in W^{1,p}(G) : \int_G u \,\mathrm{d}x = 0 \right\}. \tag{3.5}$$

Then, (3.4) holds also for all $u \in W_*^{1,p}(G)$, with different C.

Proof

(i) Let $u \in W_0^{1,p}(G)$, $G = (x_1, x_2)$ be arbitrary, but fixed. By Theorem 3.1.4, there exists $\widetilde{u} \in C^0(\overline{G})$ such that $u = \widetilde{u}$ for a.e. $x \in \overline{G}$ and such that $\widetilde{u}(x_1) = 0$. Therefore, using Hölder's inequality,

$$|\widetilde{u}(x)| = |\widetilde{u}(x) - \widetilde{u}(x_1)| = \left| \int_{x_1}^{x} u'(\xi) \,\mathrm{d}\xi \right| \le |x - x_1|^{\frac{1}{q}} \, \|u'\|_{L^p(G)},$$

where $\frac{1}{q} + \frac{1}{p} = 1$. Hence, the result follows with $C = (\int_G |x - x_1|^{\frac{p}{q}} \,\mathrm{d}x)^{\frac{1}{p}}$.

(ii) Let $u \in W_*^{1,p}(G)$. Then, there exists $\widetilde{u} \in C^0(\overline{G})$ such that $u = \widetilde{u}$ for a.e. $x \in \overline{G}$ and such that $\int_G \widetilde{u} \,\mathrm{d}x = 0$. Therefore, there is $x^* \in \overline{G}$ such that $\widetilde{u}(x^*) = 0$. We may repeat therefore the proof of (i) with $u(x_1)$ replaced by $u(x^*)$. Taking the supremum over all possible values of x^* gives the result. □

In Chap. 2, we have already introduced the Bochner spaces $L^p(J; \mathcal{H})$ which consist of functions $u : J \to \mathcal{H}$ such that the $L^p(J; \mathcal{H})$-norm is finite. For the theory of parabolic PDEs, it will prove essential to consider maps $u : J \to \mathcal{H}$ which are also differentiable (in time). We call u' the *weak derivative* of u if

$$\int_J u'(t)\varphi(t) \,\mathrm{d}t = -\int_J u(t)\varphi'(t) \,\mathrm{d}t, \quad \forall \varphi \in C_0^1(J).$$

Definition 3.1.7 Let $\mathcal{H}$ be a real Hilbert space with the norm $\|\cdot\|_{\mathcal{H}}$. For $J = (0, T)$ with $T > 0$, and $1 \le p \le \infty$, the space $W^{1,p}(J; \mathcal{H})$ is defined by

$$W^{1,p}(J; \mathcal{H}) := \{u \in L^p(J; \mathcal{H}) : u' \in L^p(J; \mathcal{H})\},$$

with the norm

$$\|u\|_{W^{1,p}(J;\mathcal{H})} := \begin{cases} (\int_J \|u(t)\|_{\mathcal{H}}^p + \|u'(t)\|_{\mathcal{H}}^p \, \mathrm{d}t)^{1/p} & \text{if } 1 \le p < \infty, \\ \operatorname{ess\,sup}_J (\|u(t)\|_{\mathcal{H}} + \|u'(t)\|_{\mathcal{H}}) & \text{if } p = \infty. \end{cases}$$

We again denote by $H^1(J; \mathcal{H}) := W^{1,2}(J; \mathcal{H})$.

3.2 Variational Parabolic Framework

Let $\mathcal{V} \subset \mathcal{H}$ be Hilbert spaces with continuous, dense embedding. We identify $\mathcal{H}$ with its dual $\mathcal{H}^*$ and obtain the triplet

$$\mathcal{V} \subset \mathcal{H} \equiv \mathcal{H}^* \subset \mathcal{V}^*. \tag{3.6}$$

Denote by $(\cdot,\cdot)_{\mathcal{H}}$ the inner product on $\mathcal{H}$, and let $\|\cdot\|_{\mathcal{V}}$, $\|\cdot\|_{\mathcal{H}}$ be the norms on $\mathcal{V}$ and $\mathcal{H}$, respectively. Furthermore, let $J = (0, T)$ with $T > 0$, $f \in L^2(J; \mathcal{V}^*)$ and $u_0 \in \mathcal{H}$. Consider the variational setting of (3.1):

$$\begin{aligned} &\text{Find } u \in L^2(J; \mathcal{V}) \cap H^1(J; \mathcal{V}^*) \text{ such that} \\ &\frac{\mathrm{d}}{\mathrm{d}t}\langle u, v\rangle_{\mathcal{V}^*,\mathcal{V}} + a(u, v) = \langle f, v\rangle_{\mathcal{V}^*,\mathcal{V}}, \ \forall v \in \mathcal{V}, \ \text{ a.e. in } J, \\ &u(0) = u_0, \end{aligned} \tag{3.7}$$

where $\langle\cdot,\cdot\rangle_{\mathcal{V}^*,\mathcal{V}}$ denotes the extension of the $\mathcal{H}$-inner product as *duality pairing* in $\mathcal{V}^* \times \mathcal{V}$. In particular, by Riesz representation theorem, we have $\langle u, v\rangle_{\mathcal{V}^*,\mathcal{V}} = (u, v)_{\mathcal{H}}$, for all $u \in \mathcal{H}$, $v \in \mathcal{V}$. The *bilinear form* $a(\cdot,\cdot) : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is associated with the operator $\mathcal{A} \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$ in (3.1) via

$$a(u, v) := -\langle \mathcal{A}u, v\rangle_{\mathcal{V}^*,\mathcal{V}}, \quad \forall u, v \in \mathcal{V},$$

where we denote by $\mathcal{L}(\mathcal{V}, \mathcal{W})$ the vector space of linear and continuous operators $\mathcal{A} : \mathcal{V} \to \mathcal{W}$.

Remark 3.2.1

(i) $\frac{\mathrm{d}}{\mathrm{d}t}$ in (3.7) is understood as a weak derivative.
(ii) The choice of the space $\mathcal{V}$ depends on the operator $\mathcal{A}$. Often we can choose $\mathcal{H} = L^2$. The choice of $\mathcal{V}$ is usually the closure of a dense subspace of smooth functions, such as C_0^∞, with respect to the 'energy' norm induced by $\mathcal{A}$.

(iii) We only require $u(0) = u_0$ in $\mathcal{H}$. In particular, for well-posedness of the equation, it is only required that $u_0 \in L^2$, *not* that $u_0 \in \mathcal{V}$. This is important, e.g. for binary contracts, where the payoff u_0 is discontinuous (and, therefore, does not belong to $\mathcal{V}$ in general).
(iv) The bilinear form $a(\cdot,\cdot)$ is, in general, *not* symmetric due to the presence of a drift term in the operator $\mathcal{A}$.

We have the following general result for the existence of weak solutions of the abstract parabolic problem (3.7). A proof is given in Appendix B, Theorem B.2.2. See also [64, 65, 115].

Theorem 3.2.2 *Assume that the bilinear form $a(\cdot,\cdot)$ in (3.7) is continuous, i.e. there is $C_1 > 0$ such that*

$$|a(u,v)| \le C_1 \|u\|_{\mathcal{V}} \|v\|_{\mathcal{V}}, \quad \forall u, v \in \mathcal{V}, \tag{3.8}$$

and satisfies the "Gårding inequality", i.e. there are $C_2 > 0$, $C_3 \ge 0$ such that

$$a(u,u) \ge C_2 \|u\|_{\mathcal{V}}^2 - C_3 \|u\|_{\mathcal{H}}^2, \quad \forall u \in \mathcal{V}. \tag{3.9}$$

Then, problem (3.7) admits a unique solution $u \in L^2(J;\mathcal{V}) \cap H^1(J;\mathcal{V}^*)$. Moreover, $u \in C^0(\overline{J};\mathcal{H})$, and there holds the a priori estimate

$$\|u\|_{C^0(J;\mathcal{H})} + \|u\|_{L^2(J;\mathcal{V})} + \|u\|_{H^1(J;\mathcal{V}^*)} \le C\left(\|u_0\|_{\mathcal{H}} + \|f\|_{L^2(J;\mathcal{V}^*)}\right). \tag{3.10}$$

We give a simple example for the weak formulation (3.7).

Example 3.2.3 Consider the heat equation as in (2.4). Then, the spaces $\mathcal{H}, \mathcal{V}, \mathcal{V}^*$ in (3.6) are $\mathcal{H} = L^2(G)$, $\mathcal{V} = H_0^1(G)$, $\mathcal{V}^* = H^{-1}(G)$, and the bilinear form $a(\cdot,\cdot)$ is given by

$$a(u,v) = \int_G u'(x)v'(x)\,\mathrm{d}x, \quad u, v \in H_0^1(G).$$

The bilinear form is continuous on $\mathcal{V}$, since by the Hölder inequality, for all $u, v \in H_0^1(G)$

$$|a(u,v)| \le \int_G |u'v'|\,\mathrm{d}x \le \|u'\|_{L^2(G)} \|v'\|_{L^2(G)} \le \|u\|_{H^1(G)} \|v\|_{H^1(G)}.$$

Furthermore, we have for all $u \in H_0^1(G)$, by the Poincaré inequality (3.4)

$$\begin{aligned} a(u,u) &= \int_G |u'(x)|^2\,\mathrm{d}x = \frac{1}{2}\|u'\|_{L^2(G)}^2 + \frac{1}{2}\|u'\|_{L^2(G)}^2 \\ &\ge \frac{1}{2C}\|u\|_{L^2(G)}^2 + \frac{1}{2}\|u'\|_{L^2(G)}^2 \end{aligned}$$

$$\geq \frac{1}{2}\min\{C^{-1},1\}\big(\|u\|^2_{L^2(G)} + \|u'\|^2_{L^2(G)}\big) = C_2\|u\|^2_{H^1(G)},$$

i.e. (3.9) holds with $C_3 = 0$. Hence, according to Theorem 3.2.2, the variational formulation of the heat equation admits, for $u_0 \in L^2(G)$, $f \in L^2(J; H^{-1}(G))$ a unique weak solution $u \in L^2(J; H_0^1(G)) \cap H^1(J; H^{-1}(G))$.

We can always achieve $C_3 = 0$ in (3.9). If we substitute in (3.1) $v = e^{-\lambda t}u$ with suitably chosen λ and multiply (3.1) by $e^{-\lambda t}$, we find that v satisfies the problem

$$\partial_t v + \mathcal{A}v + \lambda v = e^{-\lambda t} f, \quad \text{in } J \times G,$$

with $v(0,x) = u_0(x)$ in G. Choosing $\lambda > 0$ large enough, the bilinear form $a(u,v) + \lambda(u,v)$ satisfies (3.9) with $C_3 = 0$.

3.3 Discretization

For the discretization we use the method of lines where first (3.7) is only discretized in space to obtain a system of coupled ODEs which are solved in a second step.

Let V_N be a one-parameter family of subspaces $V_N \subset \mathcal{V}$ with finite dimension $N = \dim V_N < \infty$. For each fixed $t \in J$ we approximate the solution $u(t,x)$ of (3.7) by a function $u_N(t) \in V_N$. Furthermore, let $u_{N,0} \in V_N$ be an approximation of u_0. Then, the semidiscrete form of (3.7) is the initial value problem,

$$\begin{aligned} &\text{Find } u_N \in C^1(J; V_N) \text{ such that for } t \in J \\ &(\partial_t u_N, v_N)_{\mathcal{H}} + a(u_N, v_N) = \langle f, v_N\rangle_{\mathcal{V}^*,\mathcal{V}}\,, \ \forall v_N \in V_N, \\ &u_N(0) = u_{N,0}\,, \end{aligned} \tag{3.11}$$

for the approximate solution function $u_N(t) : J \to V_N$. Let V_N be generated by a finite element basis, $V_N = \operatorname{span}\{b_i(x) : 1 \le i \le N\}$. We write $u_N \in V_N$ in terms of the basis functions, $u_N(t,x) = \sum_{j=1}^N u_{N,j}(t)\, b_j(x)$, and obtain the matrix form of the semidiscretization (3.11)

$$\begin{aligned} &\text{Find } \underline{u}_N \in C^1(J; \mathbb{R}^N) \text{ such that for } t \in J, \\ &\mathbf{M}\dot{\underline{u}}_N(t) + \mathbf{A}\underline{u}_N(t) = \underline{f}(t)\,, \\ &\underline{u}_N(0) = \underline{u}_0\,, \end{aligned} \tag{3.12}$$

where $\underline{u}_0$ denotes the coefficient vector of $u_{N,0}$. The mass and stiffness matrices and the load vector with respect to the basis of V_N are given by

$$\mathbf{M}_{ij} = (b_j, b_i)_{\mathcal{H}}, \quad \mathbf{A}_{ij} = a(b_j, b_i), \quad f_i(t) = \langle f, b_i\rangle_{\mathcal{V}^*,\mathcal{V}}, \tag{3.13}$$

where $i, j = 1, \ldots, N$. Let k_m, $m = 1, \ldots, M$, be a sequence of (not necessarily equal sized) time steps and set $t_0 := 0$, $t_m := \sum_{i=1}^{m} k_i$ such that $t_M = T$. Applying the θ-scheme, we obtain the fully discrete form

$$\begin{aligned}
&\text{Find } u_N^m \in V_N \text{ such that for } m = 1, \ldots, M,\\
&k_m^{-1}(u_N^m - u_N^{m-1}, v_N)_{\mathcal{H}} + a(u_N^{m-1+\theta}, v_N) = \langle f^{m-1+\theta}, v_N\rangle_{\mathcal{V}^*,\mathcal{V}}\,,\ \forall v_N \in V_N,\\
&u_N^0 = u_{N,0},
\end{aligned} \tag{3.14}$$

where $u_N^{m+\theta} = \theta u_N(t_{m+1}) + (1-\theta)u_N(t_m)$ and $f^{m+\theta} = \theta f(t_{m+1}) + (1-\theta) f(t_m)$. We can again write (3.14) in matrix notation,

$$\begin{aligned}
&\text{Find } \underline{u}_N^m \in \mathbb{R}^N \text{ such that for } m = 1, \ldots, M,\\
&(\mathbf{M} + k_m\theta\mathbf{A})\underline{u}_N^m = (\mathbf{M} - k_m(1-\theta)\mathbf{A})\underline{u}_N^{m-1} + k_m(\theta\,\underline{f}^m + (1-\theta)\,\underline{f}^{m-1}),\\
&\underline{u}_N^0 = \underline{u}_0\,.
\end{aligned} \tag{3.15}$$

In the next section, we discuss the implementation of the matrix form (3.15).

3.4 Implementation of the Matrix Form

Let $G = (a, b)$. We describe a scheme to calculate the stiffness matrix $\mathbf{A}$ in case the corresponding bilinear form $a(\cdot,\cdot)$ has the form

$$a(\varphi, \phi) = \int_G \big(\alpha(x)\varphi'(x)\phi'(x) + \beta(x)\varphi'(x)\phi(x) + \gamma(x)\varphi(x)\phi(x)\big)\,\mathrm{d}x, \tag{3.16}$$

and the finite element subspace V_N consists of continuous, piecewise linear functions. Let $\mathcal{T} = \{a = x_0 < x_1 < x_2 < \cdots < x_{N+1} = b\}$ be an arbitrary mesh on G. Setting $K_l := (x_{l-1}, x_l)$, $h_l := |K_l| = x_l - x_{l-1}$, $l = 1, \ldots, N+1$, we can also write $\mathcal{T} = \{K_l\}_{l=1}^{N+1}$. Define

$$S_{\mathcal{T}}^1 := \big\{u(x) \in C^0(G) : u|_{K_l} \text{ is linear on } K_l \in \mathcal{T}\big\}. \tag{3.17}$$

A basis for $S_{\mathcal{T}}^1 = \operatorname{span}\{b_i(x) : i = 0, \ldots, N+1\}$ is given by the so-called *hat-functions* where, for $1 \le i \le N$,

$$b_i(x) := \begin{cases} (x - x_{i-1})/h_i & \text{if } x \in (x_{i-1}, x_i],\\ (x_{i+1} - x)/h_{i+1} & \text{if } x \in (x_i, x_{i+1}),\\ 0 & \text{else,}\end{cases} \tag{3.18}$$

and

$$b_0(x) := \begin{cases} (x_1 - x)/h_0 & \text{if } x \in (x_0, x_1),\\ 0 & \text{else,}\end{cases} \tag{3.19}$$

$$b_{N+1}(x) := \begin{cases} (x - x_N)/h_{N+1} & \text{if } x \in (x_N, x_{N+1}), \\ 0 & \text{else.} \end{cases} \tag{3.20}$$

If the mesh $\mathcal{T}$ is equidistant, i.e. $h_i = h = (b-a)/(N+1)$, we can write

$$b_i(x) = \max\{0, 1 - h^{-1}|x - x_i|\}, \ i = 0, \ldots, N+1.$$

Note that for a given subspace $S^1_{\mathcal{T}}$, there are many different possible choices of basis functions. The choice (3.18)–(3.20) are the basis functions with *smallest support*. The FE subspace to approximate functions with homogeneous Dirichlet boundary conditions is

$$S^1_{\mathcal{T},0} := S^1_{\mathcal{T}} \cap H^1_0(G) = \operatorname{span}\{b_i(x) : i = 1, \ldots, N\}, \tag{3.21}$$

with $\dim S^1_{\mathcal{T},0} = N$.

3.4.1 Elemental Forms and Assembly

We decompose $a(\cdot,\cdot)$ (3.16) into *elemental bilinear forms* $a_l(\cdot,\cdot)$, $l = 1, \ldots, N+1$,

$$\begin{aligned} a(b_j, b_i) &= \int_G \big(\alpha(x) b_j'(x) b_i'(x) + \beta(x) b_j'(x) b_i(x) + \gamma(x) b_j(x) b_i(x)\big)\,\mathrm{d}x \\ &= \sum_{l=1}^{N+1} \int_{K_l} \big(\alpha(x) b_j'(x) b_i'(x) + \beta(x) b_j'(x) b_i(x) + \gamma(x) b_j(x) b_i(x)\big)\,\mathrm{d}x \\ &=: \sum_{l=1}^{N+1} a_l(b_j, b_i). \end{aligned}$$

The restrictions $b_i|_{K_l}$, $i = l-1, l$, are linear and given by the element shape functions $N^1_{K_l} := b_{l-1}(x)|_{K_l}$, $N^2_{K_l} := b_l(x)|_{K_l}$, $l = 1, \ldots, N+1$. The element stiffness matrix $\mathbf{A}_l$ associated to $a_l(\cdot,\cdot)$ can be computed for each element independently. We transform each element $K_l \in \mathcal{T}$ to the so-called *reference element* $\widehat{K} := (-1, 1)$ via an element mapping

$$K_l \ni x = F_{K_l}(\xi) := \frac{1}{2}(x_{l-1} + x_l) + \frac{1}{2} h_l \xi, \quad \xi \in \widehat{K},$$

with derivative $F'_{K_l}(\xi) = h_l/2$. Using the reference element shape functions

$$\widehat{N}_1(\xi) = \frac{1}{2}(1 - \xi), \quad \widehat{N}_2(\xi) = \frac{1}{2}(1 + \xi),$$

which are independent of the element K_l, we have for all $K_l \in \mathcal{T}$,

$$N^1_{K_l}(F_{K_l}(\xi)) = \widehat{N}_1(\xi), \quad N^2_{K_l}(F_{K_l}(\xi)) = \widehat{N}_2(\xi).$$

The reference shape functions are shown in Fig. 3.1.

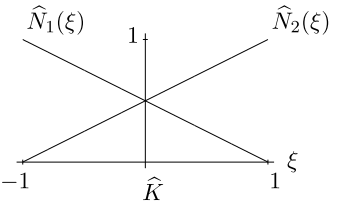

Fig. 3.1 Reference element shape functions on the reference element $\widehat{K}$

Furthermore, for $i, j = 1, 2$ there holds

$$\begin{aligned}(\mathbf{A}_l)_{ij} &= a_l(N^j_{K_l}, N^i_{K_l}) \\ &= \int_{K_l} \left(\alpha(x)\partial_x N^j_{K_l}\partial_x N^i_{K_l} + \beta(x)\partial_x N^j_{K_l} N^i_{K_l} + \gamma(x) N^j_{K_l} N^i_{K_l}\right) \mathrm{d}x \\ &= \int_{\widehat{K}} \left(\widehat{\alpha}_{K_l}(\xi)\frac{4}{h_l^2}\widehat{N}'_j\widehat{N}'_i + \widehat{\beta}_{K_l}(\xi)\frac{2}{h_l}\widehat{N}'_j\widehat{N}_i + \widehat{\gamma}_{K_l}(\xi)\widehat{N}_j\widehat{N}_i\right)\frac{h_l}{2}\,\mathrm{d}\xi,\end{aligned}$$

where

$$\widehat{\alpha}_{K_l}(\xi) := \alpha\big(F_{K_l}(\xi)\big), \quad \widehat{\beta}_{K_l}(\xi) := \beta\big(F_{K_j}(\xi)\big), \quad \widehat{\gamma}_{K_l}(\xi) := \gamma\big(F_{K_l}(\xi)\big).$$

For later purpose, it is useful to split the integral into three parts

$$(\mathbf{A}_l)_{ij} = (\mathbf{S}_l)_{ij} + (\mathbf{B}_l)_{ij} + (\mathbf{M}_l)_{ij}, \tag{3.22}$$

where

$$(\mathbf{S}_l)_{ij} = \frac{2}{h_l}\int_{\widehat{K}} \widehat{\alpha}_{K_l}\widehat{N}'_j\widehat{N}'_i\,\mathrm{d}\xi, \quad (\mathbf{B}_l)_{ij} = \int_{\widehat{K}} \widehat{\beta}_{K_l}\widehat{N}'_j\widehat{N}_i\,\mathrm{d}\xi,$$
$$(\mathbf{M}_l)_{ij} = \frac{h_l}{2}\int_{\widehat{K}} \widehat{\gamma}_{K_l}\widehat{N}_j\widehat{N}_i\,\mathrm{d}\xi.$$

For general coefficients $\alpha(x)$, $\beta(x)$ and $\gamma(x)$, these integrals cannot be computed exactly. Hence, they have to be approximated by a numerical quadrature

$$\int_{-1}^{1} f(\xi)\,\mathrm{d}\xi \approx \sum_{j=1}^{p} w_j^{(p)} f(\xi_j^{(p)}),$$

using a p-point quadrature rule with quadrature weights $w_j^{(p)}$ and quadrature points $\xi_j^{(p)} \in (-1, 1)$, $j = 1, \ldots, p$.

It remains to construct the stiffness matrix $\mathbf{A}$. The idea is to express the global basis functions $b_i(x)$ in terms of the local shape functions. We have for $i = 1, \ldots, N$,

$$b_i(x) = N^2_{K_i} + N^1_{K_{i+1}}, \quad b_0(x) = N^1_{K_1}, \quad b_{N+1}(x) = N^2_{K_{N+1}}.$$

Hence, we can assemble the matrix $\mathbf{A} \in \mathbb{R}^{(N+2)\times(N+2)}$ using the following summation

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{K_1} & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix} + \begin{pmatrix} 0 & & & \\ & \mathbf{A}_{K_2} & & \\ & & \ddots & \\ & & & 0 \end{pmatrix} + \cdots + \begin{pmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & & \mathbf{A}_{K_{N+1}} \end{pmatrix}. \tag{3.23}$$

As for the stiffness matrix, we also decompose the load vector $\underline{f}$ into elemental loads. For $f(t) \in L^2(G)$, we have

$$(f(t), b_i) = \int_G f(t,x) b_i(x)\,\mathrm{d}x = \sum_{l=1}^{N+1} \int_{K_l} f(t,x) b_i(x)\,\mathrm{d}x =: \sum_{l=1}^{N+1} (f_l(t), b_i).$$

Using again the local shape function, we obtain for $i = 1, 2$,

$$(f_l(t), N^i_{K_l}) = \int_{K_l} f(t,x) N^i_{K_l}\,\mathrm{d}x = \int_{\widehat{K}} \widehat{f}_{K_l}(t,\xi) \widehat{N}_i \frac{h_l}{2}\,\mathrm{d}\xi,$$

where

$$\widehat{f}_{K_l}(t,\xi) := f\big(t, F_{K_l}(\xi)\big).$$

For general functions f, these integrals cannot be computed exactly and have to be approximated by a numerical quadrature rule. To assemble the global load vector $\underline{f}$, we use

$$\underline{f} = \begin{pmatrix} \underline{f}_{K_1} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \underline{f}_{K_2} \\ \vdots \\ 0 \end{pmatrix} + \cdots + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \underline{f}_{K_{N+1}} \end{pmatrix}.$$

Remark 3.4.1 For $\mathcal{V} = H_0^1(G)$ we have $V_N = \operatorname{span}\{b_i(x) : i = 1, \ldots, N\}$, since $u(x_0) = u(x_{N+1}) = 0$. Hence, we get N degrees of freedom and the reduced matrices $\widetilde{\mathbf{M}}$, $\widetilde{\mathbf{A}} \in \mathbb{R}^{N\times N}$ where the first and last rows and columns of $\mathbf{A}$ in (3.23) are omitted. Similar considerations can be made for the vector $\underline{f}$. If we have non-homogeneous Dirichlet boundary condition, $u(t,a) = \phi_1(t)$, $u(t,b) = \phi_2(t)$ for given functions ϕ_1, ϕ_2, we write $u(t,x) = w(t,x) + \phi(t,x)$, where the function $\phi(t,x)$ satisfies the boundary conditions, $\phi(t,a) = \phi_1(t)$, $\phi(t,b) = \phi_2(t)$. Now, w has again homogeneous Dirichlet boundary conditions and is the solution of

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle w, v\rangle_{\mathcal{V}^*,\mathcal{V}} + a(w,v) = \langle f, v\rangle_{\mathcal{V}^*,\mathcal{V}} - \langle \partial_t \phi, v\rangle_{\mathcal{V}^*,\mathcal{V}} - a(\phi, v).$$

Choosing $\phi(t,x) = \phi_1(t) b_0(x) + \phi_2(t) b_{N+1}(x)$, we obtain the matrix form

$$\widetilde{\mathbf{M}} \dot{\underline{w}}_N(t) + \widetilde{\mathbf{A}} \underline{w}_N(t) = \widetilde{\underline{l}}(t),$$

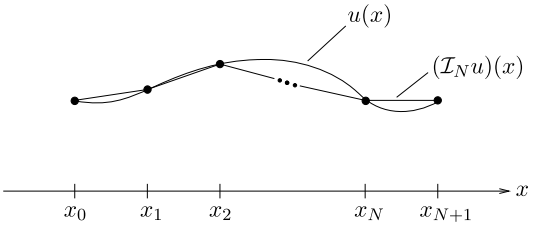

Fig. 3.2 Function $u(x)$ and its piecewise linear approximation $(\mathcal{I}_N u)(x)$

where the load vector $\underline{l}$ is given by

$$\underline{l}(t) = \underline{f}(t) - \mathbf{M}\dot{\underline{\phi}} - \mathbf{A}\underline{\phi},$$

with $\underline{\phi} = (\phi_1(t), 0, \ldots, 0, \phi_2(t))^\top \in \mathbb{R}^{N+2}$ denoting the coefficient vector of ϕ. For Neumann boundary conditions, i.e. $\partial_x u(t,a) = \phi_1(t)$, $\partial_x u(t,b) = \phi_2(t)$, the boundary degrees of freedom must be kept.

3.4.2 Initial Data

In the discretization (3.14), we need an approximation of u_0 in V_N. If $u_0 \in H^1(G)$ (e.g. u_0 being the payoff of a put or call option), we can use *nodal interpolation*, i.e. $u_{N,0} = \mathcal{I}_N u_0$. For $u \in H^1(G)$, define the nodal interpolant $\mathcal{I}_N u \in S^1_{\mathcal{T}}$ by

$$(\mathcal{I}_N u)(x) := \sum_{i=0}^{N+1} u(x_i) b_i(x), \tag{3.24}$$

as illustrated in Fig. 3.2. Since $b_i(x_j) = \delta_{ij}$, we get $(\mathcal{I}_N u)(x_i) = u(x_i)$.

For $u \in H^1(G)$, the nodal value $u(x_i)$ is well defined due to Theorem 3.1.4.

Lemma 3.4.2 *The interpolation operator $\mathcal{I}_N : H^1(G) \to S^1_{\mathcal{T}}$ defined in (3.24) is bounded, i.e. there exists a constant C (which is independent of N) such that*

$$\|\mathcal{I}_N u\|_{H^1(G)} \le C \|u\|_{H^1(G)}, \quad \forall u \in H^1(G). \tag{3.25}$$

If we only have $u_0 \in L^2(G)$ (e.g. u_0 being the payoff of a digital option), we cannot use interpolation, but need the L^2-projection of u_0 on V_N, i.e. $u_{N,0} = \mathcal{P}_N u_0$. For $u \in L^2(G)$, the L^2-projection of u on V_N is defined as the solution of

$$\int_G \mathcal{P}_N u v_N \,\mathrm{d}x = \int_G u v_N \,\mathrm{d}x, \quad \forall v_N \in V_N. \tag{3.26}$$

Using the basis b_i, $i = 0, \ldots, N+1$, of V_N we can obtain the coefficient vector $\underline{u}_N$ of $\mathcal{P}_N u_0$ as the solution of the linear system

$$\mathbf{M}\underline{u}_N = \underline{f}, \quad \text{with } f_i = \int_G u b_i \,\mathrm{d}x, \quad i = 0, \ldots, N+1.$$

We immediately have from (3.26), using Hölder's inequality,

Lemma 3.4.3 *The L^2-projection $\mathcal{P}_N : L^2(G) \to S^1_{\mathcal{T}}$ defined in (3.26) is bounded, i.e.*

$$\|\mathcal{P}_N u\|_{L^2(G)} \le \|u\|_{L^2(G)}, \quad \forall u \in L^2(G). \tag{3.27}$$

3.5 Stability of the θ-Scheme

For the stability of (3.14), we prove that the finite element solutions satisfy an analog of the estimate (3.10). For this section, we assume the uniform mesh width h in space and constant time steps $k = T/M$. We define

$$\|v\|_a := \big(a(v,v)\big)^{\frac{1}{2}}. \tag{3.28}$$

In the analysis, we will use for $f \in V_N^*$ the following notation:

$$\|f\|_* := \sup_{v_N \in V_N} \frac{(f, v_N)}{\|v_N\|_a}. \tag{3.29}$$

We will also need $\lambda_{\mathcal{A}}$ defined by

$$\lambda_{\mathcal{A}} := \sup_{v_N \in V_N} \frac{\|v_N\|^2}{\|v_N\|_*^2}.$$

In the case $\frac{1}{2} \le \theta \le 1$, the θ-scheme is stable for any time step $k > 0$, whereas in the case $0 \le \theta < \frac{1}{2}$ the time step k must be sufficiently small.

Proposition 3.5.1 *In the case $0 \le \theta < \frac{1}{2}$, assume*

$$\sigma := k(1-2\theta)\lambda_{\mathcal{A}} < 2. \tag{3.30}$$

Then, there are constants C_1 and C_2 independent of h and of k such that the sequence $\{u_N^m\}_{m=0}^M$ of solutions of the θ-scheme (3.14) satisfies the stability estimate

$$\|u_N^M\|_{L^2}^2 + C_1 k \sum_{m=0}^{M-1} \|u_N^{m+\theta}\|_a^2 \le \|u_N^0\|_{L^2}^2 + C_2 k \sum_{m=0}^{M-1} \|f^{m+\theta}\|_*^2, \tag{3.31}$$

where C_1, C_2 satisfy in the case of $\frac{1}{2} \le \theta \le 1$,

$$0 < C_1 < 2, \qquad C_2 \ge \frac{1}{2 - C_1}, \tag{3.32}$$

and in the case of $0 \le \theta < \frac{1}{2}$,

$$0 < C_1 < 2 - \sigma, \qquad C_2 \ge \frac{1 + (4 - C_1)\sigma}{2 - \sigma - C_1}. \tag{3.33}$$

Proof Define

$$X^m := \|u_N^m\|_{L^2}^2 - \|u_N^{m+1}\|_{L^2}^2 + C_2 k \|f^{m+\theta}\|_*^2 - C_1 k \|u_N^{m+\theta}\|_a^2.$$

The assertion follows if we show that $X^m \ge 0$. Then, adding these inequalities for $m = 0, \dots, M-1$ will obviously give (3.31).

Let $w := u_N^{m+1} - u_N^m$, then $u_N^{m+\theta} = (u_N^m + u_N^{m+1})/2 + (\theta - \frac{1}{2})w$ and

$$\|u_N^{m+1}\|_{L^2}^2 - \|u_N^m\|_{L^2}^2 = (u_N^{m+1} - u_N^m, u_N^{m+1} + u_N^m) = (w, 2u_N^{m+\theta} - (2\theta - 1)w).$$

By the definition of the θ-scheme, we have

$$\begin{aligned}(w, u_N^{m+\theta}) &= k\big(-\mathcal{A}u_N^{m+\theta} + f^{m+\theta}, u_N^{m+\theta}\big) = k\big[-\|u_N^{m+\theta}\|_a^2 + (f^{m+\theta}, u_N^{m+\theta})\big] \\ &\le k\big[-\|u_N^{m+\theta}\|_a^2 + \|f^{m+\theta}\|_* \|u_N^{m+\theta}\|_a\big].\end{aligned}$$

This gives

$$X^m \ge (2\theta - 1)\|w\|_{L^2}^2 + k\big[(2 - C_1)\|u_N^{m+\theta}\|_a^2 - 2\|f^{m+\theta}\|_* \|u_N^{m+\theta}\|_a + C_2 \|f^{m+\theta}\|_*^2\big].$$

In the case of $\frac{1}{2} \le \theta \le 1$, we now obtain $X^m \ge 0$ if condition (3.32) is satisfied.

In the case $0 \le \theta < \frac{1}{2}$, we have by the definition of the θ-scheme that $(w, v_N) = k(-\mathcal{A}u_N^{m+\theta} + f^{m+\theta}, v_N)$, yielding

$$\begin{aligned}\|w\|_{L^2} &\le \lambda_{\mathcal{A}}^{1/2} \|w\|_* \le \lambda_{\mathcal{A}}^{1/2} k\big(\|\mathcal{A}u_N^{m+\theta}\|_* + \|f^{m+\theta}\|_*\big) \\ &= \lambda_{\mathcal{A}}^{1/2} k\big(\|u_N^{m+\theta}\|_a + \|f^{m+\theta}\|_*\big),\end{aligned}$$

since $(\mathcal{A}u_N^{m+\theta}, v_N) \le \|u_N^{m+\theta}\|_a \|v_N\|_a$ gives $\|\mathcal{A}u_N^{m+\theta}\|_* \le \|u_N^{m+\theta}\|_a$ and choosing $v_N := u_N^{m+\theta}$ gives $\|\mathcal{A}u_N^{m+\theta}\|_* \ge \|u_N^{m+\theta}\|_a$. Hence,

$$k^{-1} X^m \ge (2 - C_1 - \sigma)\|u_N^{m+\theta}\|_a^2 - 2(1 + \sigma)\|f^{m+\theta}\|_* \|u_N^{m+\theta}\|_a + (C_2 - \sigma)\|f^{m+\theta}\|_*^2.$$

Therefore, we have $X^m \ge 0$ if conditions (3.30), (3.33) hold. □

Remark 3.5.2 The conditions (3.30), (3.33) are time-step restrictions of CFL[1]-type. Here, time-step restrictions are formulated in terms of the matrix property $\lambda_{\mathcal{A}}$. If

[1] CFL is an acronym for Courant, Friedrich and Lewy who identified an analogous condition as being necessary for the stability of explicit timestepping schemes for first order, hyperbolic equations.

$\mathcal{V} = H_0^1(G)$, and if $V_N = S_{\mathcal{T}}^1$, we obtain $\lambda_{\mathcal{A}} \sim Ch^{-2}$ as $h \downarrow 0$. Hence, we get stability provided the CFL type stability condition

$$k \le Ch^2/(1-2\theta), \quad 0 \le \theta < \frac{1}{2}$$

holds. For $\frac{1}{2} \le \theta \le 1$, the θ-scheme is stable without any time step restriction.

3.6 Error Estimates

Let $u^m(x) = u(t_m, x)$, u_N^m be as in (3.14) and assume V_N consists of linear finite elements, i.e. $V_N = S_{\mathcal{T}}^1$. For $m = 0, \ldots, M-1$, we want to estimate the error $e_N^m(x) := u^m(x) - u_N^m(x)$. Therefore, we split the error

$$e_N^m = (u^m - \mathcal{I}_N u^m) + (\mathcal{I}_N u^m - u_N^m) =: \eta^m + \xi_N^m, \tag{3.34}$$

where $\mathcal{I}_N : V \to V_N$ is the interpolant as defined in (3.24). For a fixed time point t_m, $\eta^m(x) = u(t_m, x) - (\mathcal{I}_N u)(t_m, x) \in V$ is a consistency error for which we now give an error estimate.

3.6.1 Finite Element Interpolation

We prove error estimates of the *interpolation error* $u - \mathcal{I}_N u$.

Proposition 3.6.1 *Let $\mathcal{I}_N : V \to V_N$ be the interpolant as defined in* (3.24). *Then, the following error estimates hold*:

$$\|(u - \mathcal{I}_N u)^{(n)}\|_{L^2(G)}^2 \le C \sum_{i=1}^{N+1} h_i^{2(\ell-n)} \|u^{(\ell)}\|_{L^2(K_i)}^2, \ n = 0, 1, \ \ell = 1, 2. \tag{3.35}$$

In particular, if the mesh is uniform, i.e. $h_i = h$,

$$\|(u - \mathcal{I}_N u)^{(n)}\|_{L^2(G)} \le Ch^{\ell-n} \|u^{(\ell)}\|_{L^2(G)}, \ n = 0, 1, \ \ell = 1, 2. \tag{3.36}$$

Proof Consider $\widehat{G} = (0, 1)$ and $\widehat{u} \in H^2(\widehat{G})$. Then, $\widehat{u}' - c \in \widetilde{H}^1(\widehat{G})$ for $c = \int_0^1 \widehat{u}'$. By the Poincaré inequality (3.4), which also holds for the space $\widetilde{H}^1(\widehat{G})$ (see (3.5)),

$$\|\widehat{u}' - c\|_{L^2(\widehat{G})} \le \widehat{C} \, \|\widehat{u}''\|_{L^2(\widehat{G})}. \tag{3.37}$$

With

$$\widehat{\mathcal{I}_N \widehat{u}} := \widehat{u}(0) + \int_0^x c \, \mathrm{d}x = \widehat{u}(0) + cx,$$

we have $(\widehat{\mathcal{I}_N \widehat{u}})(1) = \widehat{u}(0) + c = \widehat{u}(1)$ and $\widehat{u} - \widehat{\mathcal{I}_N \widehat{u}} \in H_0^1(\widehat{G}) \cap H^2(\widehat{G})$. Therefore, by (3.4),

$$\|\widehat{u} - \widehat{\mathcal{I}_N \widehat{u}}\|_{L^2(\widehat{G})} \le \widehat{C} \|\widehat{u}' - (\widehat{\mathcal{I}_N \widehat{u}})'\|_{L^2(\widehat{G})} = \widehat{C} \|\widehat{u}' - c\|_{L^2(\widehat{G})} \le \widehat{C}^2 \|\widehat{u}''\|_{L^2(\widehat{G})}. \tag{3.38}$$

If $G = (0, h)$, $h > 0$, we get from (3.37), (3.38) by scaling to the interval $(0, h)$

$$\|u' - (\mathcal{I}_N u)'\|_{L^2(0,h)} \le \widehat{C} h \|u''\|_{L^2(0,h)}, \tag{3.39}$$

$$\|u - \mathcal{I}_N u\|_{L^2(0,h)} \le \widehat{C}^2 h^2 \|u''\|_{L^2(0,h)}, \tag{3.40}$$

and $(\mathcal{I}_N u)(0) = u(0)$, $(\mathcal{I}_N u)(h) = u(h)$. Applying this to each interval K_i, squaring and summing yields (3.35). □

In Proposition 3.6.1, we estimated the mean square error of $u - \mathcal{I}_N u$. We are often also interested in pointwise error estimates. Some (not optimal) bounds on the pointwise error can be deduced from Proposition 3.6.1 with

Proposition 3.6.2 *Let $G = (0, 1)$. For every $u \in H_0^1(G)$, the following holds*:

$$\|u\|_{L^\infty(G)}^2 \le 2 \|u\|_{L^2(G)} \|u'\|_{L^2(G)}. \tag{3.41}$$

Proof If $u \in H_0^1(G)$, $u = \widetilde{u} \in C^0(\overline{G})$. Let $\xi \in \overline{G}$ be such that $\|u\|_{L^\infty(G)} = \max_x |\widetilde{u}(x)| = |\widetilde{u}(\xi)|$. Then

$$\|u\|_{L^\infty(G)}^2 = |\widetilde{u}(\xi)|^2 = |(\widetilde{u}(\xi))^2 - (\widetilde{u}(0))^2| \le \left| \int_0^\xi (\widetilde{u}(\eta)^2)' \, \mathrm{d}\eta \right|$$

$$= 2 \left| \int_0^\xi \widetilde{u}(\eta) \widetilde{u}'(\eta) \, \mathrm{d}\eta \right| \le 2 \|\widetilde{u}\|_{L^2(G)} \|\widetilde{u}'\|_{L^2(G)}. \qquad \square$$

Corollary 3.6.3 *For $G = (0, 1)$, $u \in H^2(G)$ and equidistant mesh width h, one has*

$$\|u - \mathcal{I}_N u\|_{L^\infty(G)} \le C h^{\frac{3}{2}} \|u''\|_{L^2(G)}, \tag{3.42}$$

as $h \to 0$. For a general interval $G = (a, b)$, (3.41), (3.42) also hold with constants that depend on $b - a$.

If u has better regularity than just $u \in H^2(G)$, better convergence rates are possible.

Corollary 3.6.4 *For $G = (0, 1)$, $u \in W^{2,\infty}(G)$ and equidistant mesh width h, one has*

$$\|u - \mathcal{I}_N u\|_{L^\infty(G)} \le C h^2 \|u\|_{W^{2,\infty}(G)}, \tag{3.43}$$

as $h \to 0$.

3.6.2 Convergence of the Finite Element Method

Assume uniform mesh width h in space and constant time steps $k = T/M$ in time. We show now that the computed sequence $\{u_N^m\}$ converges, as $h \to 0$ and $k \to 0$, to the exact solution of (3.7). We have

Theorem 3.6.5 *Assume $u \in C^1(\overline{J}; H^2(G)) \cap C^3(\overline{J}; H^{-1}(G))$. Let $u^m(x) = u(t_m, x)$ and u_N^m be as in (3.14), with $V_N = S_{\mathcal{T}}^1$. Assume for $0 \le \theta < \frac{1}{2}$ also (3.30). Then, the following error bound holds*:

$$\begin{aligned}
&\|u^M - u_N^M\|_{L^2(G)}^2 + k \sum_{m=0}^{M-1} \|u^{m+\theta} - u_N^{m+\theta}\|_a^2 \\
&\quad \le Ch^2 \max_{0\le t\le T} \|u(t)\|_{H^2(G)} + Ch^2 \int_0^T \|\partial_t u(s)\|_{H^1(G)}^2 \,\mathrm{d}s \\
&\quad + C \begin{cases} k^2 \int_0^T \|\partial_{tt} u(s)\|_*^2 \,\mathrm{d}s & \text{if } 0 \le \theta \le 1, \\ k^4 \int_0^T \|\partial_{ttt} u(s)\|_*^2 \,\mathrm{d}s & \text{if } \theta = \frac{1}{2}. \end{cases}
\end{aligned} \tag{3.44}$$

Remark 3.6.6 By the properties (3.8)–(3.9) (with $C_3 = 0$), the norm $\|\cdot\|_a$ in (3.28) is equivalent to the energy-norm $\|\cdot\|_{\mathcal{V}}$. Thus, we see from (3.44) that we have $\|u^M - u_N^M\|_{\mathcal{V}} = \mathcal{O}(h + k)$, i.e. first order convergence in the energy norm, provided the solution $u(t, x)$ is sufficiently smooth. However, one can also prove second order convergence in the L^2-norm, i.e. $\|u^M - u_N^M\|_{L^2(G)} = \mathcal{O}(h^2 + k)$, if $\theta \in [0, 1] \setminus \{1/2\}$, and $\|u^M - u_N^M\|_{L^2(G)} = \mathcal{O}(h^2 + k^2)$ if $\theta = 1/2$. Hence, for continuous, linear finite elements, we obtain the same convergence rates as for the finite difference discretization in Theorem 2.3.8.

The proof of Theorem 3.6.5 will be given in several steps. We define $e_N^m := u^m - u_N^m$ and consider the splitting (3.34), where now $\mathcal{I}_N$ denotes nodal interpolant defined in (3.24). Since we already estimated the consistency error $\eta^m = u^m - \mathcal{I}_N u^m$, we focus on $\xi_N^m \in V_N$.

Lemma 3.6.7 *If $u \in C^1(\overline{J}; H)$, the errors $\{\xi_N^m\}_m$ are solutions of the θ-scheme: Given $\xi_N^0 := \mathcal{I}_N u_0 - u_N^0$, for $m = 0, \ldots, M-1$ find $\xi_N^{m+1} \in V_N$ such that $\forall v_N \in V_N$:*

$$k^{-1}(\xi_N^{m+1} - \xi_N^m, v_N) + a\big(\theta\xi_N^{m+1} + (1-\theta)\xi_N^m, v_N\big) = (r^m, v_N) \tag{3.45}$$

where the residuals $r^m = r_1^m + r_2^m + r_3^m$ are given by

$$\begin{aligned}
(r_1^m, v_N) &= \big(k^{-1}(u^{m+1} - u^m) - \dot{u}^{m+\theta}, v_N\big), \\
(r_2^m, v_N) &= \big(k^{-1}(\mathcal{I}_N u^{m+1} - \mathcal{I}_N u^m) - k^{-1}(u^{m+1} - u^m), v_N\big), \\
(r_3^m, v_N) &= a(\mathcal{I}_N u^{m+\theta} - u^{m+\theta}, v_N)\,.
\end{aligned}$$

The stability of the θ-scheme, Proposition 3.5.1, gives

Corollary 3.6.8 *Under the assumptions of Proposition 3.5.1,*

$$\|\xi_N^M\|_{L^2(G)}^2 + C_1 k \sum_{m=0}^{M-1} \|\xi_N^{m+\theta}\|_a^2 \le \|\xi_N^0\|_{L^2(G)}^2 + C_2 k \sum_{m=0}^{M-1} \|r^m\|_*^2 . \tag{3.46}$$

To prove Theorem 3.6.5, it is therefore sufficient to estimate the residual $\|r^m\|_*$.

Proof of Theorem 3.6.5

(i) Estimate of r_1: for any $v_N \in V_N$, we have

$$|(r_1^m, v_N)| \le \|k^{-1}(u^{m+1} - u^m) - \dot{u}^{m+\theta}\|_* \, \|v_N\|_a .$$

With

$$k^{-1}(u^{m+1} - u^m) - \dot{u}^{m+\theta} = k^{-1} \int_{t_m}^{t_{m+1}} (s - (1-\theta)t_{m+1} - \theta t_m)\, \ddot{u}\, \mathrm{d}s,$$

we get

$$\begin{aligned} \|k^{-1}(u^{m+1} - u^m) - \dot{u}^{m+\theta}\|_* &\le k^{-1} \int_{t_m}^{t_{m+1}} |s - (1-\theta)t_{m+1} - \theta t_m| \, \|\ddot{u}\|_* \, \mathrm{d}s \\ &\le C_\theta k^{\frac{1}{2}} \left(\int_{t_m}^{t_{m+1}} \|\ddot{u}(s)\|_*^2 \, \mathrm{d}s \right)^{\frac{1}{2}}. \end{aligned}$$

(ii) Estimate of r_2: for any $v_N \in V_N$,

$$\begin{aligned} |(r_2^m, v_N)| &\le C \left\| k^{-1} \left[(u^{m+1} - u^m) - \mathcal{I}_N (u^{m+1} - u^m) \right] \right\|_* \, \|v_N\|_a \\ &= C k^{-1} \left\| (I - \mathcal{I}_N) \int_{t_m}^{t_{m+1}} \dot{u}(s)\, \mathrm{d}s \right\|_* \|v_N\|_a \\ &\le C k^{-1} \int_{t_m}^{t_{m+1}} \|\dot{u} - \mathcal{I}_N \dot{u}\|_* \, \mathrm{d}s \, \|v_N\|_a . \end{aligned}$$

(iii) Estimate of r_3: using the continuity of $a(\cdot,\cdot)$,

$$|(r_3^m, v_N)| \le C \, \|u^{m+\theta} - \mathcal{I}_N u^{m+\theta}\|_a \|v_N\|_a .$$

We have proved that for every $m = 0, 1, \ldots, M-1$,

$$\begin{aligned} \|r^m\|_*^2 &\le Ck \int_{t_m}^{t_{m+1}} \|\ddot{u}(x)\|_*^2 \, \mathrm{d}s \\ &\quad + Ck^{-1} \int_{t_m}^{t_{m+1}} \|\dot{u} - \mathcal{I}_N \dot{u}\|_*^2 \, \mathrm{d}s + C \|u^{m+\theta} - \mathcal{I}_N u^{m+\theta}\|_a^2 . \end{aligned}$$

Inserting into (3.46), we get

$$
\begin{aligned}
&\|e_N^M\|_{L^2(G)}^2 + C_1 k \sum_{m=0}^{M-1} \|e_N^{m+\theta}\|_a^2 \\
&\quad \le 2\left\{ \|\eta^M\|_{L^2(G)}^2 + C_1 k \sum_{m=0}^{M-1} \|\eta^{m+\theta}\|_a^2 + \|\xi_N^M\|_{L^2(G)}^2 + C_1 k \sum_{m=0}^{M-1} \|\xi_N^{m+\theta}\|_a^2 \right\} \\
&\quad \le C\left\{ \|\eta^M\|_{L^2(G)}^2 + C_1 k \sum_{m=0}^{M-1} \|\eta^{m+\theta}\|_a^2 + \|\xi_N^0\|_{L^2(G)}^2 + k\, C_\theta k \int_0^T \|\ddot{u}(s)\|_*^2 \,\mathrm{d}s \right. \\
&\qquad \left. + \int_0^T \|\dot{u} - \mathcal{I}_N \dot{u}\|_*^2 \,\mathrm{d}s + C\, k \sum_{m=0}^{M-1} \|u^{m+\theta} - \mathcal{I}_N u^{m+\theta}\|_a^2 \right\} \\
&\quad \le C\left\{ \|\xi_N^0\|_{L^2(G)}^2 + \|\eta^M\|_{L^2(G)}^2 + k \sum_{m=0}^{M-1} \|\eta^{m+\theta}\|_a^2 \right. \\
&\qquad \left. + \int_0^T \|\dot{\eta}\|_*^2 \,\mathrm{d}s + k^2 \int_0^T \|\ddot{u}(s)\|_*^2 \,\mathrm{d}s \right\}.
\end{aligned}
$$

(iv) The terms involving $\eta = u - \mathcal{I}_N u$ are estimated using the interpolation estimates of Proposition 3.6.1. Furthermore, if u_0 is approximated with the L^2-projection, we have $\|\xi_N^0\|_{L^2(G)} = \|\mathcal{I}_N u_0 - u_{N,0}\|_{L^2(G)} \le \|\mathcal{I}_N u_0 - u_0\|_{L^2(G)} + \|u_0 - u_{N,0}\|_{L^2(G)}$. Since $u \in C^1(\overline{J}; H^2(G))$ by assumption, we have $u_0 \in H^2(G)$. The L^2-projection $u_{N,0}$ is a quasi-optimal approximation of u_0, i.e. $\|u_0 - \mathcal{P}_N u_0\|_{L^2(G)} \le Ch^2 \|u_0\|_{H^2(G)}$. We obtain from Proposition 3.6.1 optimal convergence rates with respect to h, and Theorem 3.6.5 is proved. □

3.7 Further Reading

The basic finite element method is, for example, described in Braess [24] and for parabolic problems in detail by Thomée [154]. Error estimates in a very general framework are also given in Ern and Guermond [64]. In this section, we only considered the θ-scheme for the time discretization. It is also possible to apply finite elements for the time discretization as in Schötzau and Schwab [146, 147] where an hp-discontinuous Galerkin method is used. It yields exponential convergence rates instead of only algebraic ones as in the θ-scheme.