Engineering Notes on Concepts of the Finite Element Method for Elliptic Problems

Jörg Schröder

Abstract In this contribution, we discuss some basic mechanical and mathematical features of the finite element technology for elliptic boundary value problems. Originating from an engineering perspective, we will introduce step by step of some basic mathematical concepts in order to set a basis for a deeper discussion of the rigorous mathematical approaches. In this context, we consider the boundedness of functions, the classification of the smoothness of functions, classical and mixed variational formulations as well as the H^{-1}-FEM in linear elasticity. Another focus is on the analysis of saddle point problems occurring in several mixed finite element formulations, especially on the solvability and stability of the associated discretized versions.

1 Introduction

This chapter deals with some fundamental concepts needed for the understanding of the mathematical background of the finite element method (FEM). Starting from a one-dimensional boundary value problem, we motivate the formulation of an abstract minimization problem in order to generalize the problems occurring in the numerical approximation of elliptic boundary value problems. The presented general explanations originate from an engineering point of view and are consulted of the mathematical framework needed for a deeper understanding. Of course, there are a variety of excellent textbooks dealing with this topic, from the engineering as well as from the mathematical point of view. Textbooks with a more mechanical motivation are (amongst many others) e.g. Hughes (1987), Wriggers (2008), Auricchio et al. (2004), Gockenbach (2006), Berdichevsky (2009), Becker et al. (1981), and Oden and Carey

J. Schröder (✉)
Department of Civil Engineering, Faculty of Engineering, Institute of Mechanics, University of Duisburg-Essen, Essen, Germany
e-mail: j.schroeder@uni-due.de

J. Schröder and P. de Mattos Pimenta (eds.), *Novel Finite Element Technologies for Solids and Structures*, CISM International Centre for Mechanical Sciences 597,
https://doi.org/10.1007/978-3-030-33520-5_1

(1983); representatives with a mathematical background are e.g. Braess (1997), Boffi et al. (2013), Oden and Reddy (1976), Brenner and Scott (2002), and Ern and Guermond (2013).

2 Introductory Example and Propaedeutic Remarks

Let $\mathcal{B} \subset \mathbb{R}^d$ be the body of interest parametrized in $x \in \mathbb{R}^d$ with $d = 1,\ 2,\ 3$. The boundary $\partial\mathcal{B}$ of $\mathcal{B}$ is decomposed into $\partial\mathcal{B}_N$ and $\partial\mathcal{B}_D$, where Neumann and Dirichlet boundary conditions are prescribed, respectively. They satisfy

$$\partial\mathcal{B} = \partial\mathcal{B}_N \cup \partial\mathcal{B}_D \quad \text{and} \quad \partial\mathcal{B}_N \cap \partial\mathcal{B}_D = \emptyset\,. \tag{1}$$

The boundary value problem is typically defined by a set of partial differential equations (PDEs) on the open domain $\mathcal{B}$ and boundary conditions.

For simplicity we start with the simple one-dimensional ($d = 1$) boundary value problem

$$-\,(EAu'(x))' + K_s u(x) = f(x) \quad \text{in} \quad x \in \mathcal{B} = (0, l)\ , \tag{2}$$

with Young's modulus $E > 0$, cross section $A > 0$, and continuous elastic support $K_s > 0$, with units $[E] = \mathrm{N/m^2}$, $[A] = \mathrm{m^2}$, $[K_s] = \mathrm{N/m^2}$, $[u] = \mathrm{m}$, $[f] = \mathrm{N/m}$, see Fig. 1. At $x = 0$ a Dirichlet and at $x = l$ a Neumann boundary condition is applied:

$$u(0) = 0 \quad \text{and} \quad EA\,u'(l) = t_l\,, \tag{3}$$

respectively. For the following explanations, we assume that the solution $u(x)$ and the distributed loading $f(x)$ are sufficiently regular.

Analytical solution. The general solution of (2) for constant EA and $f(x) = f_0 + \Delta f\ x/l$ is based on the ansatz

$$u(x) = e^{\lambda x} \quad \rightsquigarrow \quad u'(x) = \lambda e^{\lambda x} \quad \rightsquigarrow \quad u''(x) = \lambda^2 e^{\lambda x}\,. \tag{4}$$

Substituting these expressions into the homogeneous part of (2), denoted by $\tilde{u}(x)$, yields

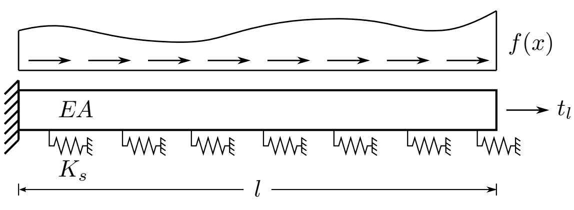


Fig. 1 Bar with continuous elastic support

$$(-EA\,\lambda^2 + K_s)e^{\lambda x} = 0 \quad \rightsquigarrow \quad \lambda_{1,2} = \pm\sqrt{\frac{K_s}{EA}} =: \pm\alpha\,, \tag{5}$$

i.e., the solution is of the form

$$\begin{aligned} \tilde{u}(x) &= \tilde{c}_1\,e^{+\alpha x} + \tilde{c}_2\,e^{-\alpha x} \\ &= \underbrace{(\tilde{c}_1 + \tilde{c}_2)}_{c_1}\,\underbrace{\frac{e^{+\alpha x} + e^{-\alpha x}}{2}}_{\cosh(\alpha x)} + \underbrace{(\tilde{c}_1 - \tilde{c}_2)}_{c_2}\,\underbrace{\frac{e^{+\alpha x} - e^{-\alpha x}}{2}}_{\sinh(\alpha x)}\,. \end{aligned} \tag{6}$$

Adding the particular solution $f_0/K_s + (\Delta f \cdot x)/(K_s\,l)$ yields

$$u(x) = c_1\,\cosh(\alpha x) + c_2\,\sinh(\alpha x) + \frac{f_0}{K_s} + \frac{\Delta f}{K_s}\,\frac{x}{l}\,. \tag{7}$$

Evaluating the boundary conditions yields the analytical expressions for the constants c_1 and c_2:

$$\begin{aligned} u(0) &= c_1 + \frac{f_0}{K_s} = 0 \\ &\to c_1 = \frac{-f_0}{K_s}\,, \\ u'(l) &= c_2\,\alpha\cosh(\alpha\,l) + \frac{\Delta f}{K_s}\,\frac{1}{l} = t_l \\ &\to \quad c_2 = \left(t_l - \frac{\Delta f}{K_s}\,\frac{1}{l}\right)\frac{1}{\alpha}e^{-\alpha l}(1 + \tanh(\alpha\,l))\,. \end{aligned} \tag{8}$$

A **weak formulation** of the boundary value problem is obtained by multiplying (2) with a test function δu and partial integration:

$$\int_0^l (EAu'\delta u' + K_s\,u\;\delta u)\mathrm{d}x = \int_0^l f\;\delta u\,\mathrm{d}x + t_l\,\delta u(l) \quad \forall\,\delta u \in V\,, \tag{9}$$

where V is a suitable space of functions, e.g., a Hilbert space. All test functions $\delta u \in V$ have to vanish at the Dirichlet boundary condition $\delta u(0) = 0$. In the variational formulation u is an element of the class of trial functions V_{trial}. The collection of both functions are denoted as admissible functions; in this simple case, V and V_{trial} coincide and have to satisfy

$$V = V_{trial} = \left\{u(x) : \int_0^l (u^2 + (u')^2)\,\mathrm{d}x < \infty, u(0) = 0\right\}. \tag{10}$$

The idea of approximation methods is to compute an approximate solution $u_h \in V_h$ in the finite-dimensional subspace $V_h \subset V$, based on

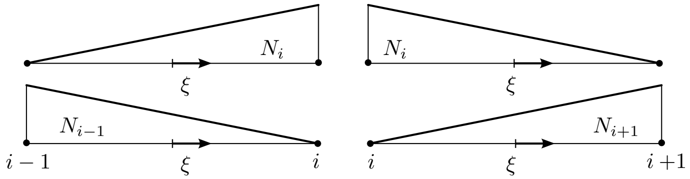


Fig. 2 Linear ansatz functions of neighboring elements, $\xi \in \left[-\frac{l^e}{2}, \frac{l^e}{2}\right]$

$$\int_0^l (EAu_h' \delta u_h' + K_s u_h \, \delta u_h) \mathrm{d}x = \int_0^l f \; \delta u_h \, \mathrm{d}x + t_l \, \delta u_h(l) \quad \forall \, \delta u_h \in V_h \,. \tag{11}$$

In order to do this within the **finite element method**, we have to subdivide the domain in num_{ele} subsections, here we choose individual finite elements with (for simplicity reasons) unit length $l^e = l/\text{num}_{\text{ele}}$. On this individual elements we define a set of ansatz functions, i.e., shape functions $N_i | i = 1, \ldots k$ with local support. We use piecewise polynomial functions which are globally C^0 continuous, as depicted in Fig. 2.

In this case, the continuity can be easily enforced by sharing the degrees of freedom at the interface between two neighboring elements. With this definitions we approximate the individual fields on element level as follows:

$$u_h = \underline{\boldsymbol{N}}^e \, \underline{\boldsymbol{d}}^e , \quad \delta u_h = \underline{\boldsymbol{N}}^e \, \delta \underline{\boldsymbol{d}}^e , \quad u_h' = \underline{\boldsymbol{B}}^e \, \underline{\boldsymbol{d}}^e , \quad \delta u_h' = \underline{\boldsymbol{B}}^e \, \delta \underline{\boldsymbol{d}}^e \tag{12}$$

with the matrix of shape functions $\underline{\boldsymbol{N}}^e$, the matrix containing the derivatives of the shape functions $\underline{\boldsymbol{B}}^e$, and the vectors of nodal (virtual) degrees of freedom ($\delta \underline{\boldsymbol{d}}^e$) $\underline{\boldsymbol{d}}^e$ of the element $e \in \{1, 2, \ldots, \text{num}_{\text{ele}}\}$:

$$\underline{\boldsymbol{d}}^e = \begin{bmatrix} d_1^e \\ d_2^e \end{bmatrix} , \quad \delta \underline{\boldsymbol{d}}^e = \begin{bmatrix} \delta d_1^e \\ \delta d_2^e \end{bmatrix} , \quad \underline{\boldsymbol{N}}^e = \begin{bmatrix} N_1 \\ N_2 \end{bmatrix} , \quad \underline{\boldsymbol{B}}^e = \begin{bmatrix} N_{1,x} \\ N_{2,x} \end{bmatrix} . \tag{13}$$

After substituting these approximations equation (11) is reformulated into

$$\begin{aligned} \sum_{e=1}^{\text{num}_{\text{ele}}} \delta \underline{\boldsymbol{d}}^{eT} \underbrace{\int_{l^e} (EA \, \underline{\boldsymbol{B}}^{eT} \underline{\boldsymbol{B}}^e + K_s \underline{\boldsymbol{N}}^{eT} \, \underline{\boldsymbol{N}}^e) \mathrm{d}x}_{\underline{\boldsymbol{k}}^e} \, \underline{\boldsymbol{d}}^e = & \\ \sum_{e=1}^{\text{num}_{\text{ele}}} \underbrace{\delta \underline{\boldsymbol{d}}^{eT} \int_{l^e} f \; \underline{\boldsymbol{N}}^{eT} \, \mathrm{d}x + t_l \, \delta \underline{\boldsymbol{d}}^e(l)}_{\delta \underline{\boldsymbol{d}}^{eT} \underline{\boldsymbol{r}}^e} & \\ \to \sum_{e=1}^{\text{num}_{\text{ele}}} \delta \underline{\boldsymbol{d}}^{eT} \left\{ \underline{\boldsymbol{k}}^e \, \underline{\boldsymbol{d}}^e - \underline{\boldsymbol{r}}^e \right\} = 0 \, . & \end{aligned} \tag{14}$$

Assembling the element matrices,

$$\underline{\boldsymbol{K}} = \mathop{\mathsf{A}}_{e=1}^{\text{num}_{\text{ele}}} \underline{\boldsymbol{k}}^e , \quad \underline{\boldsymbol{R}} = \mathop{\mathsf{A}}_{e=1}^{\text{num}_{\text{ele}}} \underline{\boldsymbol{r}}^e , \tag{15}$$

yields

$$\delta\underline{\boldsymbol{D}}^T \left\{ \underline{\boldsymbol{K}}\,\underline{\boldsymbol{D}} - \underline{\boldsymbol{R}} \right\} = 0 \quad \forall\, \delta\underline{\boldsymbol{D}} \quad \rightarrow \quad \underline{\boldsymbol{K}}\,\underline{\boldsymbol{D}} = \underline{\boldsymbol{R}} , \tag{16}$$

where $\underline{\boldsymbol{K}}$ is the global element stiffness matrix, $\underline{\boldsymbol{R}}$ the global right-hand side, $\underline{\boldsymbol{D}}$ the global vector of unknowns, and $\delta\underline{\boldsymbol{D}}$ the global vector of virtual node displacements. The FEM solution is depicted for a various number of elements num_{ele} based on a constant K_s in Fig. 3. Figure 4 compares the approximation with $\text{num}_{\text{ele}} = 16$ to the analytical solution considering different values of K_s.

Generalizations: In order to formulate an **abstract minimization problem** we define a quadratic energy functional $J(u)$, e.g., the total potential energy,

$$J(u) := \frac{1}{2} a(u, u) - L(u) , \tag{17}$$

with the symmetric form $a(u, v) = a(v, u)$. We assume that $a(u, u)$ is positive definite, i.e.,

$$a(v, v) > 0 \quad \forall\, v \neq 0 . \tag{18}$$

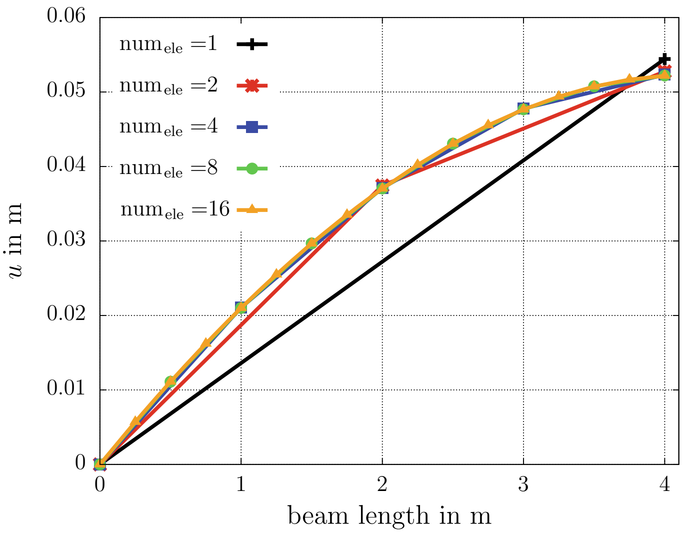


Fig. 3 Approximate FEM solution for $\text{num}_{\text{ele}} = \{1, 2, 4, 8, 16\}$ ($K_s = 10^4$ kN/m^2, $E = 210 \cdot 10^3$ kN/m^2, $l = 4$ m, $t_l = 250$ kN, $f_0 = \Delta f = 10^3$ kN/m)

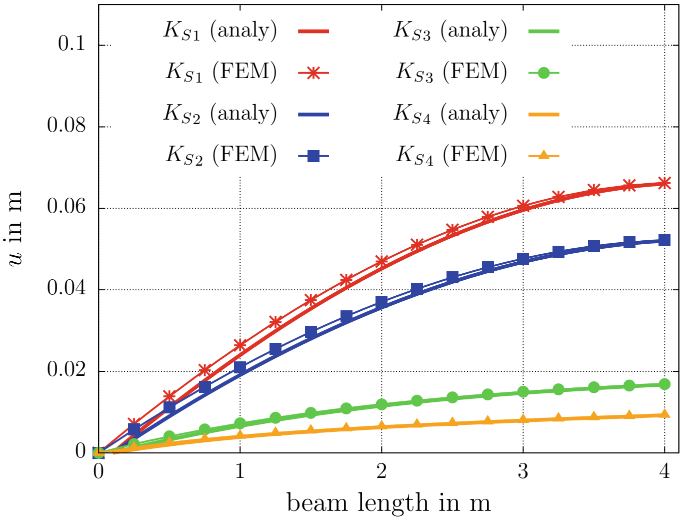


Fig. 4 Comparison of analytical and approximate FEM solution ($\text{num}_{\text{ele}} = 16$) for various values of $K_{Si} = \{10^3, 10^4, 10^5, 210 \cdot 10^3\}$ kN/m^2, $E = 210 \cdot 10^3$ kN/m^2, $t_l = 250$ kN, $f_0 = \Delta f = 10^3$ kN/m, $l = 4$ m

A bilinear form $a : V \times V \rightarrow \mathbb{R}$ is called H-elliptic (or simply elliptic) if there exists a constant $c_\alpha > 0$ such that

$$a(v, v) \geq c_\alpha \, \|v\|_H^2 \quad \forall \, v \in V \, . \tag{19}$$

The H-elliptic bilinear form induces the so-called *energy norm*

$$\|v\|_a := \sqrt{a(v, v)} \, , \tag{20}$$

which is equivalent to a norm of the associated Hilbert space H. Under this assumptions the minimization problem

$$J(u) = \min_{v \in V} J(v) \, , \tag{21}$$

where the minimum is characterized by $\langle J'(u), v \rangle := \int_0^l J'(u) \, v \, \mathrm{d}x = 0$, which is equivalent to the variational problem

$$\text{find} \quad u \in V \quad \text{satisfying} \quad a(u, v) = L(v) \quad \forall \, v \in V \, . \tag{22}$$

Identifying our model problem with the abstract formulation yields the quadratic functional

$$a(u, u) = \int_0^l EA\,(u'(x))^2\,\mathrm{d}x + \int_0^l K_s\,(u(x))^2\,\mathrm{d}x \tag{23}$$

and the linear functional

$$L(u) = \int_0^l u(x)\,f(x)\,\mathrm{d}x + t_l\,u(l) \tag{24}$$

with the Dirichlet boundary condition $u(0) = 0$, to be satisfied by the function $u(x)$, and the Neumann boundary condition $EA\,u'(l) = t_l$, appearing as a natural boundary condition in the functional.

Modus operandi. There are several direct methods for the computation of the approximate solution. Beyond this, there are several mathematical frameworks for the qualitative analysis of the existence and uniqueness of solutions. Beside well-known direct methods for the treatment of established models, described by partial differential equations, this topic is rather important for the derivation of new models in continuum thermodynamics. In this contribution, we want to motivate the main ideas of this scientific branch.

A functional is called **bounded from below** on the space V, if there exist a constant $c \in \mathbb{R}$, such that

$$J(u) \geq c \quad \forall\, u \in V\,. \tag{25}$$

This requirement can be violated if the functional is not bounded by below on V or if it is bounded by below but its minimum is not reached on V, for a physical interpretation see Berdichevsky (2009), Chap. 5.

We assume that in the quadratic functional $a(u, u)$ of our model problem (23) $u(x)$ is a differentiable function and that the integrals exists. We conclude with the meaningful engineering constants

$$EA > 0 \quad \text{and} \quad K_s > 0\,, \tag{26}$$

that $a(u, u)$ is obviously nonnegative and therefore bounded from below. A linear functional $L(u) =< l, u >$ is bounded (from above) if for $C_L > 0$

$$\|L(u)\| = |L(u)| \leq C_L \|u\| \quad \forall\, u \in V\,. \tag{27}$$

Now we have to answer the question, if the functional (17) is bounded from below or not. Let us consider our model problem. Our functional (17) can now be estimated, compare Braess (1997) Chap. 2.5, by

$$
\begin{aligned}
J(u) &\geq \frac{1}{2} c_\alpha \|u\|^2 - \|l\| \, \|u\| \\
&= \frac{1}{2} c_\alpha \left(\|u\|^2 - \frac{2}{c_\alpha} \|l\| \, \|u\| + \frac{\|l\|^2}{c_\alpha^2} - \frac{\|l\|^2}{c_\alpha^2} \right) \\
&= \frac{1}{2} c_\alpha \left(\|u\| - \frac{\|l\|}{c_\alpha} \right)^2 - \frac{\|l\|^2}{2 c_\alpha} \\
&\geq - \frac{\|l\|^2}{2 c_\alpha} \, .
\end{aligned}
\tag{28}
$$

Obviously, the functional is bounded from below.

Lax–Milgram Theorem (existence of classical solutions); Let V' be a Hilbert space, $a : V \times V \to \mathbb{R}$ a *continuous* and *H-elliptic* bilinear form defined on V, $L \in V'$ any *continuous* linear functional. Subject to these conditions there exists a *unique* solution

$$u \in V$$

such that

$$a(u, v) = L(v) \quad \forall \, v \in V \, . \qquad \blacksquare$$

Reminder, the properties of the bilinear and linear form are:

- the bilinear form has to be continuous (bounded from above), i.e., there exists a constant $C_a \in \mathbb{R}^+$[1] such that

$$|a(w, v)| \leq C_a \, \|w\|_V \|v\|_V \quad \forall \, w, v \in V \, ,$$

- the bilinear form has to be H-elliptic, i.e., there exists a constant $c_a \in \mathbb{R}^+$ such that

$$a(v, v) \geq c_a \, \|v\|_V^2 \quad \forall \, v \in V \, ,$$

- the linear functional L is continuous, i.e., there exists a constant $C_L \in \mathbb{R}^+$ such that

$$|L(v)| \leq C_L \, \|v\|_V \quad \forall \, v \in V \, .$$

Note: From the continuity of the bilinear form $a(\cdot, \cdot)$, discussed in exercise 1, we obtain $|a(u, u)| \leq C_a \, \|u\|^2$. The continuity of the linear form yields $|L(u)| \leq C_L \, \|u\|$. From the H-ellipticity of the bilinear form $a(\cdot, \cdot)$, see (19), we deduce

$$c_a \, \|u\|^2 \leq a(u, u) \leq C_a \, \|u\|^2 \, . \tag{29}$$

Exploiting the continuity of the linear form allows with $C \in \mathbb{R}^+$ for

$$c_a \, \|u\|^2 \leq a(u, u) \leq C_a \, \|u\|^2 \leq C \langle l, u \rangle \tag{30}$$

[1] nonnegative real values $\mathbb{R}_0^+$, positive real values $\mathbb{R}^+ = \mathbb{R}_0^+ \backslash 0$.

and we deduce from $\langle l, u\rangle \leq \|l\|\,\|u\|$

$$c_a\,\|u\|^2 \leq C\,\|l\|\,\|u\| \quad \text{and} \quad c_a\,\|u\| \leq C\,\|l\| \tag{31}$$

and

$$\|u\| \leq \frac{C}{c_a}\|l\| \quad \forall\, v \in V\,. \tag{32}$$

In this sense the bounded linear functional is generated by the continuity and H-ellipticity of the bilinear form $a(\cdot,\cdot)$.

Approximate solutions: In general it is cumbersome or even impossible to find exact solutions $u \in V$, therefore we are interested in **approximate solution concepts**. Applying **Ritz method** we seek a solution $u_h \in V_h$ with the discrete subspace $V_h \subset V$, i.e.,

$$J(u_h) = \min_{v_h \in V_h} J(v_h)\,. \tag{33}$$

The Ritz approach, based on our technical assumptions, is equivalent to the **Galerkin method** of the variational counterpart

$$\text{find} \quad u_h \in V_h \quad \text{satisfying} \quad a(u_h, v_h) = l(v_h) \quad \forall\, v_h \in V_h\,, \tag{34}$$

where $a(u_h, v_h)$ is a bilinear functional (linear in both arguments).

Exercise 1 Show that the bilinear form of our model problem in Eq. (9)

$$a(u, v) = \int_0^l EA\,u'v'\,\mathrm{d}x + \int_0^l K_s\,u\,v\,\mathrm{d}x \tag{35}$$

with $EA \in (0,\infty)$ and $K_s \in (0,\infty)$, is continuous!

Remark: Definition of *continuous bilinear forms* $a : U \times V \to \mathbb{R}$ on linear normed spaces U and V: A bilinear form $a(\cdot,\cdot)$ is a continuous bilinear form, if there exists a constant $C_a \in \mathbb{R}^+$ such that

$$|a(u, v)| \leq C_a\|u\|\,\|v\| \quad \forall\, u \in U, v \in V\,. \tag{36}$$

In anticipation of the following chapters we introduce the norm

$$\|u\|^2_{H^1} = (u, u)_{H^1} = \int_0^l (u^2 + (u')^2)\,\mathrm{d}x = \|u\|^2_{L^2} + \|u'\|^2_{L^2}\,; \tag{37}$$

obviously we obtain the inequalities

$$\|u\|^2_{L^2} \leq \|u\|^2_{H^1} \quad \text{and} \quad \|u'\|^2_{L^2} \leq \|u\|^2_{H^1}\,. \tag{38}$$

Solution. In order to show that the bilinear form is continuous, consider

$$\begin{aligned}\tilde{a}(u, v) = \frac{1}{EA} a(u, v) &= \int_0^l u' \, v' \, \mathrm{d}x + \int_0^l \frac{K_s}{EA} \, u \, v \, \mathrm{d}x \\ &\leq \left| \int_0^l u' \, v' \, \mathrm{d}x \right| + \int_0^l \frac{K_s}{EA} \, |u \, v| \, \mathrm{d}x \\ &\leq \left| \int_0^l u' \, v' \, \mathrm{d}x \right| + \frac{K_s}{EA} \int_0^l |u| \, |v| \, \mathrm{d}x \\ &= |(u', v')_{L^2}| + \frac{K_s}{EA} \, (|u|, |v|)_{L^2} \; .\end{aligned} \tag{39}$$

Applying the Cauchy–Schwarz inequality yields

$$|(u', v')_{L^2}| + \frac{K_s}{EA} \, (|u|, |v|)_{L^2} \leq \|u'\|_{L^2} \, \|v'\|_{L^2} + \frac{K_s}{EA} \, \|u\|_{L^2} \, \|v\|_{L^2} \; . \tag{40}$$

Using the inequalities (38) yields the final estimation

$$\begin{aligned}\tilde{a}(u, v) &\leq \|u\|_{H^1} \|v\|_{H^1} + \frac{K_s}{EA} \, \|u\|_{H^1} \, \|v\|_{H^1} \\ &= (1 + \frac{K_s}{EA}) \, \|u\|_{H^1} \, \|v\|_{H^1} \; .\end{aligned} \tag{41}$$

For our bilinear form we write

$$a(u, v) \leq C_a \, \|u\|_{H^1} \|v\|_{H^1} \quad \text{with} \quad C_a = (EA + K_s) \, . \tag{42}$$

Thus, $a(u, v)$ is continuous, or in other words it is bounded by above. ♠

3 Classification of the Smoothness of Functions

In this section we discuss the classification of functions and their derivatives with respect to Hilbert spaces. For this we first set a few notations, a more detailed summary is given in Appendix A.

$$L^2(\mathcal{B}) = \left\{ u : \|u\|^2_{L^2(\mathcal{B})} = \int_{\mathcal{B}} |u|^2 \mathrm{d}v \; < \; +\infty \right\} \tag{43}$$

characterizes the space of square integrable functions on $\mathcal{B}$. At this point, it seems to be meaningful to give some remarks concerning the *Riemann integral* and the *Lebesgue integral*. The Riemann integral has some disqualifications if we would like to use it for a satisfactory theory of PDEs.

In order to obtain a satisfactory theory of PDEs, we must—for technical reasons—integrate certain singular functions. If functions are regular enough to integrate them they are called *Lebesgue measurable*. An introduction to this topic is given in Royden (1968). For $m \in \mathbb{N}_0$[2] we define

$$H^m(\mathcal{B}) = \left\{ u : D^{\boldsymbol{\alpha}} u \in L^2(\mathcal{B}) \quad \forall \quad |\boldsymbol{\alpha}| \leq m \right\} , \tag{44}$$

with the multi-index notation for the derivatives of u, with the 3-tuple of nonnegative integers

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3) \qquad \text{and} \qquad |\boldsymbol{\alpha}| = (\alpha_1 + \alpha_2 + \alpha_3) . \tag{45}$$

Thus the α-th derivative of u with respect to (x_1, x_2, x_3) is defined by

$$D^{\boldsymbol{\alpha}} u = \frac{\partial^{\alpha_1 + \alpha_2 + \alpha_3} u}{\partial x_1^{\alpha_1} \, \partial x_2^{\alpha_2} \ldots \partial x_3^{\alpha_n}} = \frac{\partial^{|\boldsymbol{\alpha}|} u}{\partial x_1^{\alpha_1} \, \partial x_2^{\alpha_2} \ldots \partial x_3^{\alpha_n}} \tag{46}$$

Explanatory examples:

$$\begin{aligned} &D^{(0,0,0)} u = u; \qquad D^{(1,0,0)} u = \frac{\partial u}{\partial x_1}; \qquad D^{(0,0,1)} u = \frac{\partial u}{\partial x_3}; \\ &D^{(1,1,0)} u = \frac{\partial^2 u}{\partial x_1 \, \partial x_2}; \qquad D^{(3,2,1)} u = \frac{\partial^6 u}{\partial x_1^3 \, \partial x_2^2 \, \partial x_3} . \end{aligned} \tag{47}$$

The introduction of the Sobolev spaces $H^m(\mathcal{B})$ allows for the quantification of the smoothness (regularity) of functions. Let $C^m(\mathcal{B})$ be the linear space of functions u with continuous derivatives $D^{|\boldsymbol{\alpha}|} u$ of the order $0 \leq |\boldsymbol{\alpha}| \leq m$. The Sobolev spaces $H^m(\mathcal{B})$ are related with the $C^k(\overline{\mathcal{B}})$ spaces by the **Sobolev embedding theorem:** Let $\overline{\mathcal{B}} = \mathcal{B} \cup \partial \mathcal{B}$ be a bounded domain with a Lipschitz boundary. Every function in $H^m(\mathcal{B})$ belongs to $C^k(\overline{\mathcal{B}})$ if

$$m > k + 1 \quad \text{for } \mathcal{B} \subset \mathbb{R}^2, \qquad m > k + 3/2 \quad \text{for } \mathcal{B} \subset \mathbb{R}^3 . \qquad \blacksquare$$

It should be noted that the embedding is continuous:

$$H^m(\mathcal{B}) \subseteq C^k(\overline{\mathcal{B}}) .$$

Furthermore we introduce the notation

$$H_0^1(\mathcal{B}) := \{u \in H^1(\mathcal{B}), \; u|_{\partial \mathcal{B}} = 0\}, \; H_{0,D}^1(\mathcal{B}) := \{u \in H^1(\mathcal{B}), \; u|_{\partial \mathcal{B}_D} = 0\} . \tag{48}$$

[2] positive integers $\mathbb{N}_+ = \{1, 2, 3, \ldots\}$, nonnegative integers $\mathbb{N}_0 = \{0, 1, 2, 3, \ldots\} = \mathbb{N}_+ \cup \{0\}$.

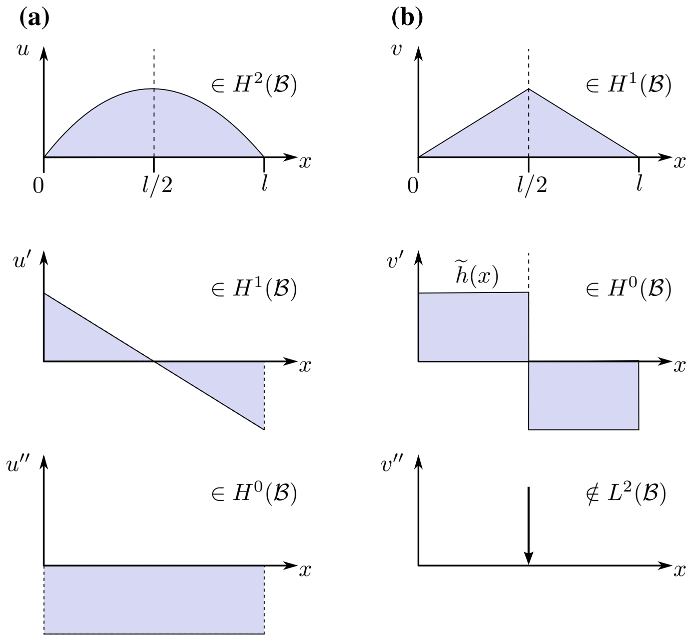


Fig. 5 Regularity of functions u and v and their derivatives on $\mathcal{B} = [0, l]$

3.1 One-Dimensional Example

To discuss the smoothness of functions, we consider the two functions depicted in Fig. 5.
For the function u in Fig. 5a we observe

$$u \in C^2(\mathcal{B}) \,, \tag{49}$$

because u is twice continuously differentiable and $u \in H^2(\mathcal{B})$. In contrast, the function v depicted in Fig. 5b is

$$v \in C^0(\mathcal{B}) \,, \tag{50}$$

because already its first derivative is not continuous. Obviously, $v \in H^1(\mathcal{B})$, due to the fact that its first derivative is square integrable,[3] i.e., $v' \in L^2(\mathcal{B})$. Although the *classical* derivative of $v(x)$ does not exist at $x = l/2$ we can define the weak derivative of v. Consider

[3]The derivatives occurring in $H^m(\mathcal{B})$ have to be interpreted as *weak* or *generalized* derivatives. Classical derivatives are functions defined pointwise on an interval. A weak derivative need only to be locally integrable. If the function is sufficiently smooth, e.g., $v \in C^m(\overline{\mathcal{B}})$, then its weak derivatives $D^\alpha u$ coincide with the classical ones for $|\alpha| \leq m$.

$$\int_0^l v\,\eta'\,\mathrm{d}x = \int_0^{l/2} v\,\eta'\,\mathrm{d}x + \int_{l/2}^l v\,\eta'\,\mathrm{d}x\,, \tag{51}$$

with the infinitely differentiable function η, satisfying $\eta(0) = \eta(l) = 0$. Integration by part, i.e.,

$$\int_a^b v\,\eta'\,\mathrm{d}x = v\,\eta\big|_a^b - \int_a^b v'\,\eta\,\mathrm{d}x\,, \tag{52}$$

yields

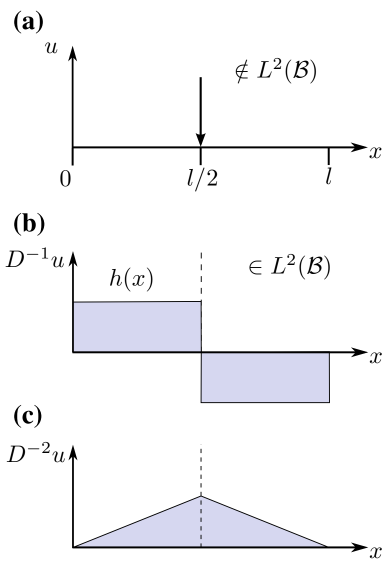
$$\begin{aligned}\int_0^l v\,\eta'\,\mathrm{d}x &= v(l/2)\,\eta(l/2) - \int_0^{l/2} v'\,\eta\,\mathrm{d}x - \eta(l/2)\,v(l/2) - \int_{l/2}^l v'\,\eta\,\mathrm{d}x \\ &= -\left\{\int_0^{l/2} v'\,\eta\,\mathrm{d}x + \int_{l/2}^l v'\,\eta\,\mathrm{d}x\right\}.\end{aligned} \tag{53}$$

The function $v'(x)$ in (53) is denoted as the weak derivative of $v(x)$.

Let us now consider the function in Fig. 6 which is a delta function representing a force acting at a point. This function (distribution) is not square integrable. Before we are able to quantify its smoothness it has to be integrated. In order to generalize the discussion we introduce the *antiderivative* D^{-1}, by means of

$$D(D^{-1}u) = u \qquad \text{with} \qquad D := \frac{\mathrm{d}}{\mathrm{d}x}\,. \tag{54}$$

Fig. 6 Antiderivatives of function u



The meaning of this operator becomes clear if we consider again the function $\widetilde{h}(x)$ in Fig. 5b:

$$v' = Dv = \widetilde{h}(x)\,.$$

The calculation of the antiderivative

$$D^{-1}v' = D^{-1}(Dv) = \int \widetilde{h}(x)\,\mathrm{d}x$$

yields the hat function depicted in Fig. 5b up to a constant. Switching back to our function (distribution) shown in Fig. 6a: Evaluating the antiderivative of the delta function $\delta(^l/_2)$ leads to

$$D^{-1}u = h(x)\,, \tag{55}$$

then we conclude that $h(x)$ is a square integrable function. In other words its first antiderivative, i.e., its "first integral", is in $L^2(\mathcal{B})$. Therefore we define

$$u \in H^{-1}(\mathcal{B})\,. \tag{56}$$

The question is: What are negative Sobolev spaces?
Let m be a positive integer, then the negative Sobolev space $H^{-m}(\mathcal{B})$ is defined as the dual of $H_0^m(\mathcal{B})$, i.e.,

$$H^{-m}(\mathcal{B}) = (H_0^m(\mathcal{B}))'\,. \tag{57}$$

The associated norm, exemplarily for $m = 1$ is

$$\|u\|_{H^{-1}(\mathcal{B})} = \|u\|_{-1,\mathcal{B}} = \min_{v \in H_0^1(\mathcal{B})\setminus 0} \frac{(u, v)_{0,\mathcal{B}}}{\|v\|_{1,B}}\,. \tag{58}$$

Based on the relations

$$H_0^m(\mathcal{B}) \subset H^m(\mathcal{B}) \subset H^0(\mathcal{B}) = L^2(B) \tag{59}$$

we conclude the inclusion properties

$$H^m(\mathcal{B}) \subset H^0(\mathcal{B}) = L^2(\mathcal{B}) \subset H^{-m}(\mathcal{B})\,. \tag{60}$$

3.2 $\boldsymbol{H}(\mathbf{div}, \boldsymbol{\mathcal{B}})$ *Hilbert Spaces*

The introduction of special Hilbert spaces related to vector or tensor valued fields is expedient for the suitable description of many engineering problems. A frequent representative, which is especially of importance in the field of elasticity, heat conduction, flow problems, etc. is the $H(\mathrm{div}, \mathcal{B})$ space which demands a $L^2(\mathcal{B})$-measurable weak divergence. The corresponding space in introduced by

$$H(\mathrm{div}, \mathcal{B}) = \left\{ \boldsymbol{u} \in (L^2(\mathcal{B}))^d \wedge \mathrm{div}\, \boldsymbol{u} \in L^2(\mathcal{B}) \right\} , \tag{61}$$

whereas d denotes the dimension of vector $\boldsymbol{u}$.

4 Variational Formulations of Linear Elasticity

In the following chapters, we concentrate on the formulation of elasticity. Let $\mathcal{B} \subset \mathbb{R}^3$ be the body of interest, parametrized in $\boldsymbol{x} \in \mathbb{R}^3$, $\boldsymbol{\sigma}$ the second-order stress tensor, $\boldsymbol{\varepsilon} = \nabla_s \boldsymbol{u}$ the symmetric second-order strain tensor, $\boldsymbol{u}$ the displacement field, $\boldsymbol{f}$ the given body force per unit volume, $\mathbb{C}$ the fourth-order elasticity tensor and $\bar{\boldsymbol{t}}$ the prescribed Neumann boundary conditions. Then the governing equations in linear elasticity are given by[4]

$$\begin{aligned}
&\text{balance of momentum:} && \mathrm{div}\, \boldsymbol{\sigma} + \boldsymbol{f} = \boldsymbol{0} \\
&\text{constitutive law:} && \boldsymbol{\sigma} = \mathbb{C} : \boldsymbol{\varepsilon} \\
&\text{kinematical condition:} && \boldsymbol{\varepsilon} = \nabla_s \boldsymbol{u} = \tfrac{1}{2}(\mathrm{grad}\, \boldsymbol{u} + \mathrm{grad}^T \boldsymbol{u}) \\
&\text{balance of angular momentum:} && \boldsymbol{\sigma} = \boldsymbol{\sigma}^T \\
&\text{Dirichlet boundary condition:} && \boldsymbol{u} = \boldsymbol{0} \text{ on } \partial\mathcal{B}_D \\
&\text{Neumann boundary condition:} && \boldsymbol{\sigma} \cdot \boldsymbol{n} = \bar{\boldsymbol{t}} \text{ on } \partial\mathcal{B}_N
\end{aligned} \tag{62}$$

4.1 Classical (Bubnov-)Galerkin Formulation

A direct substitution of $(62)_2$–$(62)_4$ into $(62)_1$ leads to a variational formulation where solely the displacements are solved in a weak form. Multiplication with a test function $\delta\boldsymbol{u}$ and integration over the domain leads to the problem of seeking $\boldsymbol{u}$ such that

$$\int_{\mathcal{B}} (\mathrm{div}[\mathbb{C} : \nabla_s \boldsymbol{u}] + \boldsymbol{f}) \cdot \delta\boldsymbol{u} \, \mathrm{d}v = 0 \qquad \forall\, \delta\boldsymbol{u}. \tag{63}$$

Integration by parts and the insertion of the important test function property $\delta\boldsymbol{u} = \boldsymbol{0}$ on $\partial\mathcal{B}_D$ leads to the formulation of seeking $\boldsymbol{u} \in H^1_{0,D}(\mathcal{B})$ of

$$\int_{\mathcal{B}} (\nabla_s \delta\boldsymbol{u} : \mathbb{C} : \nabla_s \boldsymbol{u} - \delta\boldsymbol{u} \cdot \boldsymbol{f}) \, \mathrm{d}v - \int_{\partial\mathcal{B}_N} \delta\boldsymbol{u} \cdot \bar{\boldsymbol{t}} \, \mathrm{d}a = 0 \quad \forall\, \delta\boldsymbol{u} \in H^1_{0,D}(\mathcal{B}). \tag{64}$$

It can be recognized, that for the latter weak formulation, the corresponding function space of the trial function $\boldsymbol{u}$ and the test function $\delta\boldsymbol{u}$ coincide, which is the classical

[4]Note that a restriction to homogeneous Dirichlet boundary conditions is only of technical nature and does not constitute a loss of generality, see, e.g., Braess (1997).

characteristic of the (Bubnov-)Galerkin method. The solution of (64) is equivalent to the minimizer $\boldsymbol{u} \in H^1_{0,D}(\mathcal{B})$ of the potential energy

$$\Pi(\boldsymbol{u}) = \int_{\mathcal{B}} \tfrac{1}{2} \nabla_s \boldsymbol{u} : \mathbb{C} : \nabla_s \boldsymbol{u} \, \mathrm{d}v - \int_{\mathcal{B}} \boldsymbol{u} \cdot \boldsymbol{f} \, \mathrm{d}v - \int_{\mathcal{B}_N} \boldsymbol{u} \cdot \boldsymbol{t} \, \mathrm{d}a \tag{65}$$

and constitutes the basis of the well-known displacement based FEM for linear elasticity.

4.2 Alternative Methods

In the previously discussed approach, Green's theorem has been applied to shift a derivative from the trial to the test functions. Particularly in the framework of finite elements, this is the prevalent approach. However, various alternative approximation techniques are available. In these formulations, the space of the approximative solution is distinct to the space of test functions. A first example is represented by the variational problem in Eq. (63) which is e.g. the basis of collocation methods. In the corresponding discrete formulations, the approximative solution is sought in a subspace of $H^2(\mathcal{B})$ whereas the admissible test space corresponds to $L^2(\mathcal{B})$.

In contrast to these formulations where broken (discontinuous) test spaces are appropriate, both derivatives may be transferred to the test spaces. This is executed by means of successive application of Green‘s theorem. The corresponding formulation is established as the so-called H^{-1}-method, proposed by Rachford et al. (1974). In this H^{-1}-Galerkin method different subspaces for the space of trial functions (approximation functions) U_k and the space of test functions (weighting functions) W_h are used, i.e.,

$$U_k \neq W_h. \tag{66}$$

In Kendall and Wheeler (1976), the authors adopted the procedure to a Crank–Nicolson–H^{-1}-Galerkin procedure and investigates single space variables in parabolic problems. The ansatz was recapitulated in Thomée (2006), Chap. 16. A discussion of negative norm error estimates for semi-discrete Galerkin-type Finite Element formulations for nonheterogenous parabolic equations is given in Thomée (1980). A more recent approach can be found in Goebbels (2015).
The basis of the approach is the theory of distributional differential equations. Let u be a distribution, v a test function and the equation of interest is the second-order ordinary differential equation

$$\mathrm{D}^2 u(x) = f(x) \qquad x \in \mathbb{R} \tag{67}$$

with appropriate boundary conditions.

Based on this assumptions we can set up a family of distributional differential equations, see Oden and Reddy (1976), page 365 ff:

$$
\begin{aligned}
-\ \langle \mathrm{D}^2 u, v\rangle \quad &= \langle f, v\rangle \quad \forall v \in \mathcal{D}(\mathbb{R})\\
\langle \mathrm{D}u, \mathrm{D}v\rangle \quad &= \langle f, v\rangle \quad \forall v \in \mathcal{D}(\mathbb{R})\\
-\ \langle u, \mathrm{D}^2 v\rangle \quad &= \langle f, v\rangle \quad \forall v \in \mathcal{D}(\mathbb{R})\\
\langle \mathrm{D}^{-1}u, \mathrm{D}^3 v\rangle &= \langle f, v\rangle \quad \forall v \in \mathcal{D}(\mathbb{R})
\end{aligned}
\tag{68}
$$

with the space of distribution $\mathcal{D}(\mathbb{R})$. From the distributional point of view all equations can be interpreted as equivalent. Equation $(68)_1$ is the basis for collocation methods, $(68)_2$ for the classical Galerkin method and $(68)_3$ for the H^{-1}-method.

A descriptive explanation of the H^{-1}-method on the basis of a one-dimensional boundary value problem

$$
-\ \langle u, \mathrm{D}^2 v\rangle = \langle f, v\rangle \quad \text{on } x \in (0, l) \tag{69}
$$

with $u(0) = u(l) = 0$ is discussed in Sect. 5.2.

4.3 Mixed Variational Frameworks for Linear Elasticity

Considering again the governing equations in linear elasticity in Eq. (62) it is apparent that the direct substitution of $(62)_2$–$(62)_4$ into $(62)_1$ is not mandatory. Alternatively it is possible to solve another set of equations of (62) in a weak sense. Here, especially two common variants are considered in the following.

Hellinger–Reissner Formulation: The stress–displacement based formulation solves $(62)_1$ and $(62)_2$ in a weak sense, seeking $\boldsymbol{\sigma} \in H(\mathrm{div}, \mathcal{B})$ and $\boldsymbol{u} \in H^1_{0,D}(\mathcal{B})$ such that

$$
\begin{aligned}
&\int_{\mathcal{B}} (\mathrm{div}\,\boldsymbol{\sigma} + \boldsymbol{f}) \cdot \delta\boldsymbol{u}\,\mathrm{d}v = 0 && \forall\,\delta\boldsymbol{u} \in L^2(\mathcal{B})\,,\\
&\int_{\mathcal{B}} \left(\nabla_s \boldsymbol{u} - \boldsymbol{\sigma} : \mathbb{C}^{-1}\right) : \delta\boldsymbol{\sigma}\,\mathrm{d}v = 0 && \forall\,\delta\boldsymbol{\sigma} \in L^2(\mathcal{B})\,.
\end{aligned}
\tag{70}
$$

On this basis, two additional variational formulations can be achieved which differ in their corresponding solution spaces.

Application of integration by parts in $(70)_1$ leads to the so-called *primal* version of the Hellinger–Reissner formulation. This yields a the saddle point problem seeking for $\boldsymbol{\sigma} \in L^2(\mathcal{B})$ and $\boldsymbol{u} \in H^1_{0,D}(\mathcal{B})$ such that

$$
\begin{aligned}
&\int_{\mathcal{B}} (\boldsymbol{\sigma} : \nabla_s \delta\boldsymbol{u} - \boldsymbol{f} \cdot \delta\boldsymbol{u})\,\mathrm{d}v - \int_{\partial\mathcal{B}_N} \delta\boldsymbol{u} \cdot \boldsymbol{t}\,\mathrm{d}a = 0 && \forall\,\delta\boldsymbol{u} \in H^1_{0,D}(\mathcal{B})\,,\\
&\int_{\mathcal{B}} \left(\nabla_s \boldsymbol{u} - \boldsymbol{\sigma} : \mathbb{C}^{-1}\right) : \delta\boldsymbol{\sigma}\,\mathrm{d}v = 0 && \forall\,\delta\boldsymbol{\sigma} \in L^2(\mathcal{B})\,.
\end{aligned}
\tag{71}
$$

Equivalently, this problem can be described by the potential

$$\Pi^{\mathrm{HR}}(\boldsymbol{\sigma}, \boldsymbol{u}) = \int_{\mathcal{B}} \left(-\frac{1}{2}\boldsymbol{\sigma} : \mathbb{C}^{-1} : \boldsymbol{\sigma} + \boldsymbol{\sigma} : \nabla_s \boldsymbol{u} - \boldsymbol{f} \cdot \boldsymbol{u} \right) \mathrm{d}v - \int_{\partial\mathcal{B}_N} \boldsymbol{u} \cdot \bar{\boldsymbol{t}} \, \mathrm{d}A \,. \tag{72}$$

The expression

$$\boldsymbol{\sigma} : \nabla_s \boldsymbol{u} - \frac{1}{2}\boldsymbol{\sigma} : \mathbb{C}^{-1} : \boldsymbol{\sigma} = \boldsymbol{\sigma} : \nabla_s \boldsymbol{u} - \psi^\star(\boldsymbol{\sigma}) \tag{73}$$

represents the free energy $\psi(\boldsymbol{\varepsilon})$ in terms of the complementary potential $\psi^\star(\boldsymbol{\sigma})$, i.e., $\psi(\boldsymbol{\varepsilon}) = \boldsymbol{\sigma} : \nabla_s \boldsymbol{u} - \psi^\star(\boldsymbol{\sigma})$.

In contrast, integration by parts may be applied to $(70)_2$, which yields the so-called dual Hellinger–Reissner formulation, seeking the saddle point $\boldsymbol{\sigma} \in H(\mathrm{div}, \mathcal{B})$ and $\boldsymbol{u} \in L^2(\mathcal{B})$ such that

$$\begin{aligned} &\int_{\mathcal{B}} (\mathrm{div}\, \boldsymbol{\sigma} + \boldsymbol{f}) \cdot \delta \boldsymbol{u} \, \mathrm{d}v = 0 && \forall\, \delta \boldsymbol{u} \in L^2(\mathcal{B}) \,, \\ &\int_{\mathcal{B}} \big(\boldsymbol{\sigma} : \mathbb{C}^{-1} : \delta\boldsymbol{\sigma} + \boldsymbol{u} \cdot \mathrm{div}\, \delta\boldsymbol{\sigma} \big) \, \mathrm{d}v = 0 && \forall\, \delta\boldsymbol{\sigma} \in H(\mathrm{div}, \mathcal{B}) \,. \end{aligned} \tag{74}$$

It should be remarked, that in this formulation the traction boundary condition $(62)_6$ has to be incorporated into the solution space of the stresses, since they do not appear in the weak form. In addition the stress symmetry condition has to be enforced.

Hu-Washizu Functional – Three Field Formulation: A third option is the independent interpolation of all variables entering the elasticity problem. This formulation solves $(62)_1$–$(62)_3$ in a weak sense. The optimization problem is: seek a saddle point $\boldsymbol{\varepsilon} \in L^2(\mathcal{B})$, $\boldsymbol{\sigma} \in L^2(\mathcal{B})$ and $\boldsymbol{u} \in H^1_{0,D}(\mathcal{B})$ such that

$$\begin{aligned} &\int_{\mathcal{B}} (\mathbb{C} : \boldsymbol{\varepsilon} - \boldsymbol{\sigma}) : \delta\boldsymbol{\varepsilon} \, \mathrm{d}v = 0 && \forall\, \delta\boldsymbol{\varepsilon} \in L^2(\mathcal{B}) \,, \\ &\int_{\mathcal{B}} (\nabla_s \boldsymbol{u} - \boldsymbol{\varepsilon}) : \delta\boldsymbol{\sigma} \, \mathrm{d}v = 0 && \forall\, \delta\boldsymbol{\sigma} \in L^2(\mathcal{B}) \,, \\ &\int_{\mathcal{B}} (\boldsymbol{\sigma} : \nabla_s \delta\boldsymbol{u} - \boldsymbol{f} \cdot \delta\boldsymbol{u}) \, \mathrm{d}v - \int_{\partial\mathcal{B}} \boldsymbol{t} \cdot \delta\boldsymbol{u} \, \mathrm{d}a = 0 && \forall\, \delta\boldsymbol{u} \in H^1_0(\mathcal{B}) \,. \end{aligned} \tag{75}$$

An equivalent potential formulation can be given by

$$\Pi^{\mathrm{HW}}(\boldsymbol{\varepsilon}, \boldsymbol{\sigma}, \boldsymbol{u}) = \int_{\mathcal{B}} (\psi(\boldsymbol{\varepsilon}) - \boldsymbol{\sigma} : (\nabla_s \boldsymbol{u} - \boldsymbol{\varepsilon}) - \boldsymbol{u} \cdot \boldsymbol{f}) \, \mathrm{d}v - \int_{\partial\mathcal{B}_N} \boldsymbol{u} \cdot \bar{\boldsymbol{t}} \, \mathrm{d}a. \tag{76}$$

with $\psi(\boldsymbol{\varepsilon}) = {}^1\!/_2\, \boldsymbol{\varepsilon} : \mathbb{C} : \boldsymbol{\varepsilon}$.

5 Finite Element Method

The finite element method constitutes the most prevalent discretization technique for the approximation of boundary value problems in the field of computational mechanics. As discussed in the previous chapters, the solution of the variational equations are in their corresponding Sobolev space. For the numerical treatment this solution space is restrained to a finite-dimensional space, in the following called finite element space and is denoted by the subscript h, e.g., V_h.

5.1 Conforming and Non-conforming Finite Elements

In case of a *conforming* discretization, the finite element space is a discrete subspace of the corresponding Sobolev space. Considering the problem of linear elasticity with the displacements as the only unknown, we seek $\boldsymbol{u} \in H^1_{0,D}(\mathcal{B})$

$$\int_{\mathcal{B}} (\nabla_s \delta\boldsymbol{u} : \mathbb{C} : \nabla_s \boldsymbol{u} - \delta\boldsymbol{u} \cdot \boldsymbol{f})\, \mathrm{d}v - \int_{\partial\mathcal{B}} \delta\boldsymbol{u} \cdot \boldsymbol{t}\, \mathrm{d}a = 0 \qquad \forall\, \delta\boldsymbol{u} \in H^1_{0,D}(\mathcal{B})\,. \tag{77}$$

A conforming discretization of $\boldsymbol{u}_h \in V_h$ demands in this case

$$V_h \subset H^1_{0,D}(\mathcal{B})\,. \tag{78}$$

It can be shown that $\boldsymbol{u}_h$ of a conforming finite element converges monotonically to $\boldsymbol{u}$ with increasing mesh density, if it is in addition able to represent the rigid body displacements and the constant strain states, see Bathe (1996).

Standard $H^1(\mathcal{B})$ conforming finite elements on triangles $\mathbb{P}_k$ or quadrilaterals $\mathbb{Q}_k$ are assigned with $k+1$ nodes on each edge of the element, see Fig. 7. Continuity of the approximated variable is enforced, when these nodes are shared with the adjacent elements.

$H(\mathrm{div}, \mathcal{B})$ conforming elements can be constructed, for example, with help of the Raviart–Thomas functions. In case of triangles the $\mathbb{RT}_k$ elements have $k+1$ vector-valued sampling points on each edge and in addition $k\,(k+1)$ vector-valued sampling points in the interior of the element, as exemplary depicted in Fig. 8.

In contrast, the finite element space of *non-conforming* elements is not a subspace of the appropriate Sobolev solution space and convergence is not obvious. Due to the non-conforming discretization, an additional error is introduced and has to be controlled. However, this leads to additional flexibility in the design of the finite element. The simplest non-conforming element is the Crouzeix–Raviart element, see Crouzeix and Raviart (1973). Here, we only assign k nodes on each edge and generally do not have continuity across inter-element boundaries; thus, these are non-conforming elements. The Crouzeix–Raviart finite elements are depicted in Fig. 9.

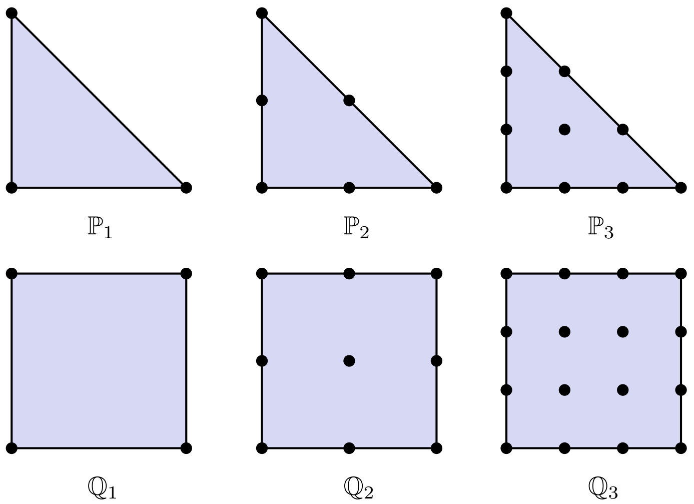


Fig. 7 Examples of $\mathbb{P}_k$ and $\mathbb{Q}_k$ elements

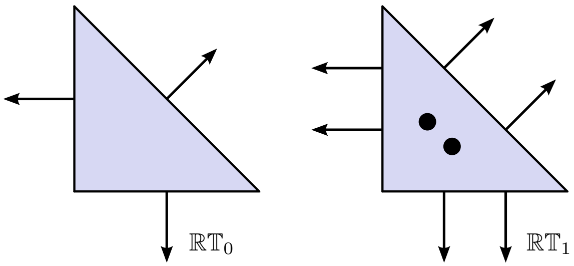


Fig. 8 Examples for $\mathbb{RT}_k$ elements with $k \geq 0$; dim $\mathbb{RT}_k = (k+1)(k+3)$

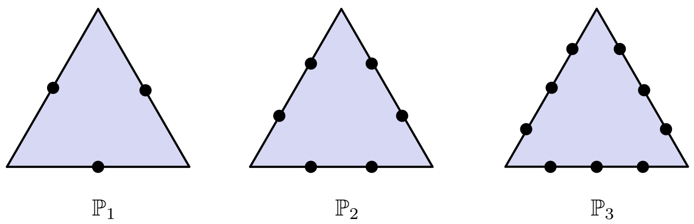


Fig. 9 Non-conforming Crouzeix–Raviart $\mathbb{P}_k$-finite elements

5.2 Example of H^{-1}-FEM for 1D Elliptic Problem

In order to approach the H^{-1}-method we analyze a one-dimensional truss element, in analogy to the one examined in Rachford et al. (1974):

$$L\,u = \left(\mathrm{EA}(x)\,u(x)'\right)' = -f(x), \qquad x \in \mathcal{B} = (0,1)\,. \tag{79}$$

where $EA(x)u(x)$ characterize the normal force in the straight bar with the longitudinal stiffness

$$\mathrm{EA}(x) = \alpha^{-1} + \alpha(x-\bar{x})^2 \quad \text{and} \quad \alpha > 0\,, \tag{80}$$

where α and $\bar{x}$ are constant parameters and the right- hand side is given by

$$f(x) = 2\,(\,1 + \alpha(x-\bar{x})(\,\arctan\alpha(x-\bar{x}) + \arctan\alpha\bar{x}\,)\,)\ . \tag{81}$$

The Dirichlet boundary conditions are defined by

$$u(0) = u(1) = 0\,. \tag{82}$$

The closed-form solution of this problem is given by

$$u(x) = (1-x)(\arctan\alpha(x-\bar{x}) + \arctan\alpha\bar{x}) \tag{83}$$

and is explicitly depicted for two different sets of α and $\bar{x}$ in Fig. 10c and 11c. Considering the plots of the longitudinal stiffness and the applied load in Fig. 10, where the parameter are chosen as $\alpha = 5$ and $\bar{x} = 0.5$. I would like to draw the reader's attention to the low stiffness in the middle of the domain. In the case $\alpha = 1000$ the domain responds to this with a rapid, jump like rising displacement.

Variational approach: The solution in terms of a variational weak form is obtained via

$$\int_{\mathcal{B}} \left(\mathrm{EA}(x)\,u'(x)\right)'\ v\,\mathrm{d}x + \int_{\mathcal{B}} f(x)\,v\,\mathrm{d}x = 0. \tag{84}$$

A reformulation using integration by parts and exploiting $v(0) = v(1) = 0$ yields

$$\int_{\mathcal{B}} \mathrm{EA}(x)\,u'(x)\,v'(x)\,\mathrm{d}x - \int_{\mathcal{B}} f(x)\,v(x)\,\mathrm{d}x = 0. \tag{85}$$

The classical FE discretization with $u_h \in U_h \subset H^1_0$ and $u_h \in \mathbb{P}_1$ or $u_h \in \mathbb{P}_2$, yields an approximation of the displacements as illustrated in Figs. 12 and 13. This standard displacement FEM ansatz even with second-order interpolation is inaccurate in an extreme edge case. It is also worth mentioning that the normal force computed from these element are also very inaccurate in comparison to the analytical solution.

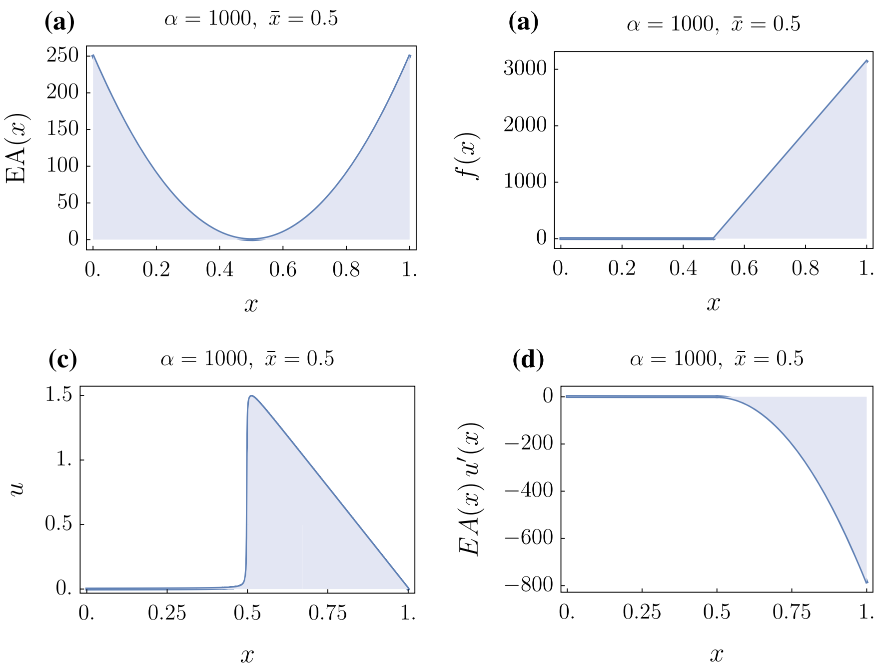


Fig. 10 Distributions of **a** stiffness $\mathrm{EA}(x)$, **b** load $f(x)$, **c** analytical solution for the displacements $u(x)$, and **d** longitudinal force distribution $\mathrm{EA}(x)u'(x)$ over $\mathcal{B}$ for $\alpha = 1000$ and $\bar{x} = 0.5$

Repeated application of integration by parts leads to another weak form, which constitutes the basis of the H^{-1}-FE approach

$$\int_{\mathcal{B}} \mathrm{EA}(x)\, u(x)\, \eta''(x)\, \mathrm{d}x + \int_{\mathcal{B}} f(x)\, \eta(x)\, \mathrm{d}x = 0. \tag{86}$$

In this case the natural discretization is of the form $u_h \in U_h \subset H^0(\mathcal{B})$, i.e., it is possible to choose discontinuous approximations of u_h, denoted by $u_h \in \mathrm{d}\mathbb{P}$. This reduces the coupling between the elements. The associated subspace consists of all piecewise polynomial functions in C^k. Simultaneously the continuity requirements regarding the test space V_h are increased. Let $\mathcal{B}_h$ denote the discretization of $\mathcal{B}$, with

$$\mathcal{B}_h = \bigcup_e \mathcal{B}^e \quad \text{with} \quad B^j = [x_{j-1}, x_j] \quad \text{and} \quad h_j = x_j - x_{j-1}. \tag{87}$$

By setting $r \geq 1$ and $-1 \leq k \leq r-2$ we define the trial space

$$U_h = \left\{ u_h \in C^k(\mathcal{B}_h) : u_h|_{\mathcal{B}^e} \in \mathbb{P}^{r-1} \text{ for } e = 1, \ldots, \mathrm{num}_{\mathrm{ele}} \right\} \tag{88}$$

and the test space

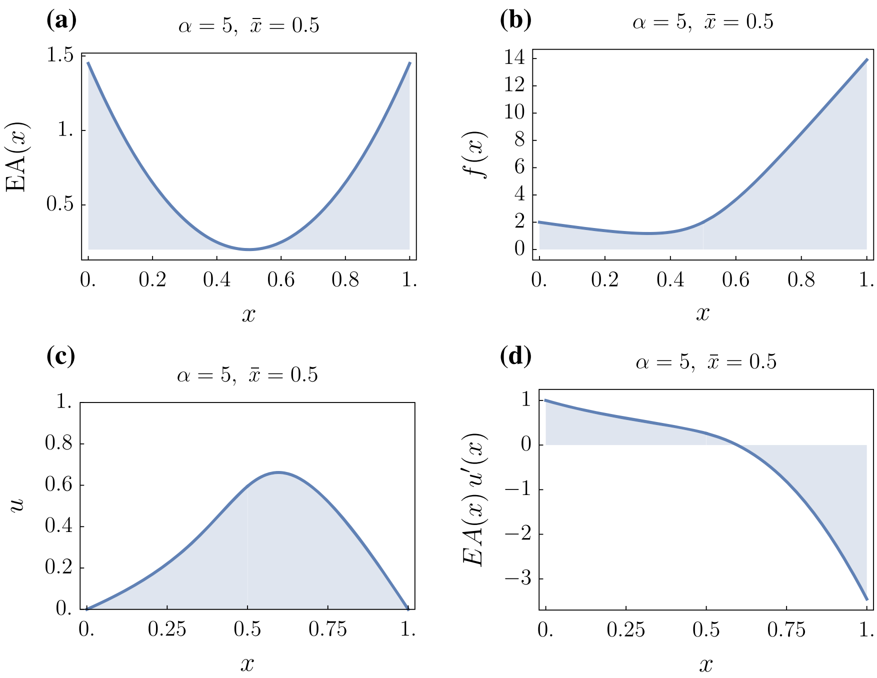


Fig. 11 Distributions of **a** stiffness EA(x), **b** load $f(x)$, **c** analytical solution for the displacements $u(x)$, and **d** longitudinal force distribution EA$(x)u'(x)$ over $\mathcal{B}$ for $\alpha = 5$ and $\bar{x} = 0.5$

$$V_h = \left\{ v_h \in C^{k+2}(\mathcal{B}_h) : v_h|_{\mathcal{B}^e} \in \mathbb{P}^{r+1} \text{ for } e = 1, \ldots, \text{num}_{\text{ele}},\ \eta(0) = \eta(1) = 0 \right\}. \tag{89}$$

For $k = -1$ the trial space U_h exhibits discontinuities at the node of the partition, where the functions in the test space V_h are continuously differentiable. Thus we have

$$U_h = \left\{ u_h \in C^{-1}(\mathcal{B}_h) : u_h|_{\mathcal{B}^e} \in \mathbb{P}^{r-1} \text{ for } e = 1, \ldots, \text{num}_{\text{ele}} \right\}, \tag{90}$$

whereas $C^{-1}(\mathcal{B}_h)$ considers all functions whose antiderivative is in $C^0(\mathcal{B}_h)$, which means we do not require continuity at the nodal points. For convenience we define this space by

$$U_h = \left\{ u_h \in \mathrm{d}\mathbb{P}^{r-1} \text{ for } e = 1, \ldots, \text{num}_{\text{ele}} \right\} \tag{91}$$

to enforce that the trial functions are discontinuous at the exterior nodes.
This leads to the corresponding space for the test functions as

$$V_h = \left\{ \eta \in C^1(\mathcal{B}_h) : \eta|_{\mathcal{B}^e} \in \mathbb{P}^{r+1} \text{ for } e = 1, \ldots, \text{num}_{\text{ele}} \right\}. \tag{92}$$

The numerical results obtained from this discretization are given in Fig. 14 for the displacements and for the stresses. The discontinuity in the displacements is clearly

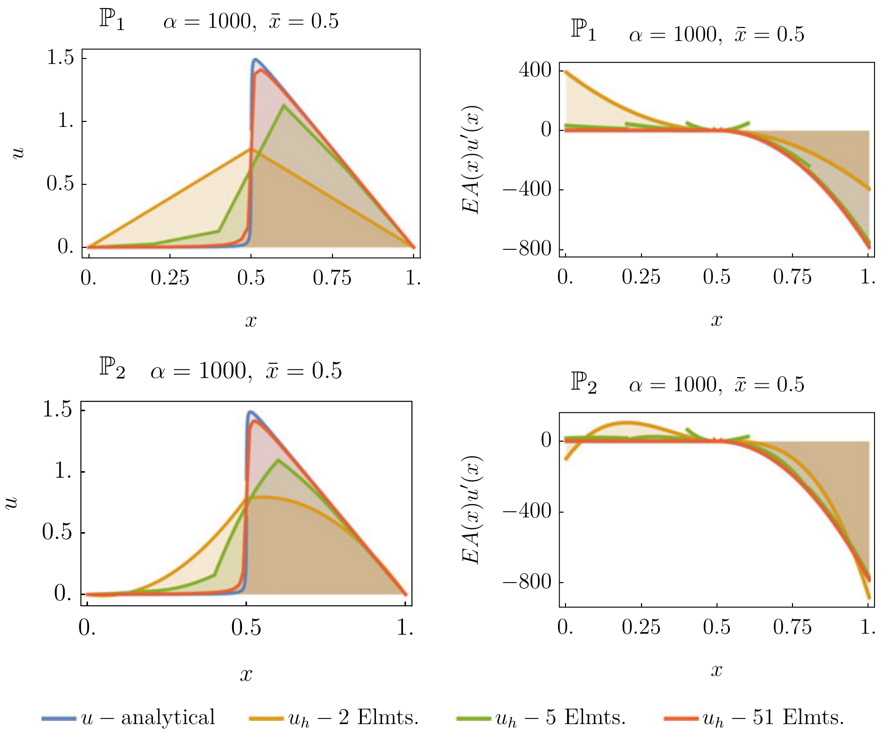


Fig. 12 Illustration of numerical solutions for $u(x)$ and $EA(x)u'(x)$ with $\alpha = 1000$ and $\bar{x} = 0.5$ using classical finite elements with $u \in \mathbb{P}_1$ (top) and $u \in \mathbb{P}_2$ (bottom)

visible, especially for the coarse discretization. The method, however, shows significant advantage for this model problem in comparison to the standard FE method. This is even more significant in terms of the normal force. The interested reader is referred to the error plots of each solution space in Figs. 15, 16, 17 and 18, with respect to both considered loading cases.

6 Analysis of Mixed Finite Elements

For the existence, uniqueness, and approximation of saddle point, problems arise from Lagrangian multipliers see Brezzi (1974). The following explanations are mainly based on the excellent treatises of Auricchio et al. (2004) and Boffi et al. (2013).

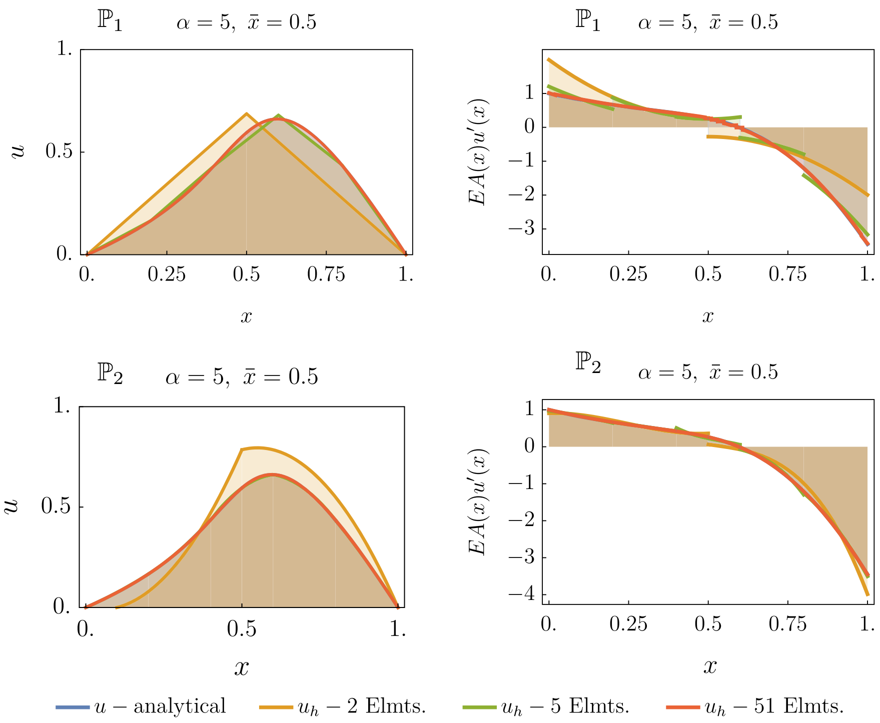


Fig. 13 Illustration of numerical solutions for $u(x)$ and $EA(x)u'(x)$ with $\alpha = 5$ and $\bar{x} = 0.5$ using classical finite elements with $u \in \mathbb{P}_1$ (top) and $u \in \mathbb{P}_2$ (bottom)

6.1 Theoretical Framework

The idea of mixed methods is based on the introduction of Lagrangian multipliers in order to relax several constraints denoted by $constr(v) = 0$, e.g., the incompressibility condition $\operatorname{div} \boldsymbol{u} = 0$. Let's start from the constrained minimization problem

$$\min_{v \in V} \left\{ J(v) \quad \text{subjected to} \quad constr(v) = 0 \right\}. \tag{93}$$

This can be reformulated by means of a Lagrangian functional of the form

$$\begin{aligned} \mathcal{L}(v, q) &= J(v) + b(v, q) \\ &= \tfrac{1}{2} a(v, v) - L(v) + b(v, q) , \end{aligned} \tag{94}$$

with

$$b(v, q) = \int_{\mathcal{B}} q \, constr(v) \, \mathrm{d}v , \tag{95}$$

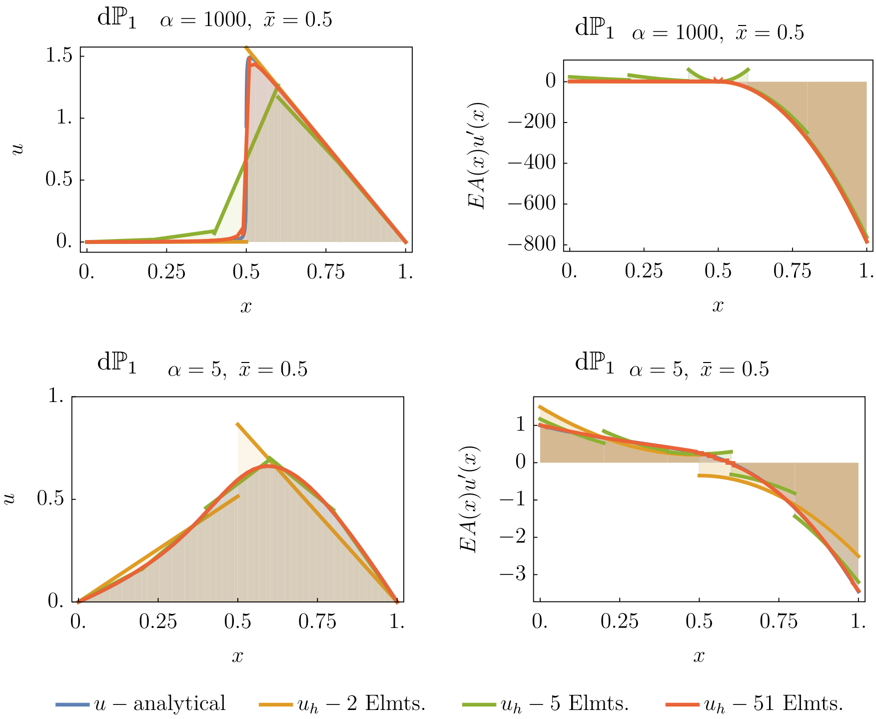


Fig. 14 Numerical solution of $u(x) \in \mathrm{d}\mathbb{P}_1$ and of $EA(x)u'(x)$ for $\alpha = 1000$, $\alpha = 5$ and $\bar{x} = 0.5$

where q denotes the Lagrange multiplier. The solution of this abstract optimization problem is (u, p) if the condition

$$\mathcal{L}(u, q) \leq \mathcal{L}(u, p) \leq \mathcal{L}(v, p) \qquad \forall\, v \in V,\ \forall\, q \in \Sigma \tag{96}$$

is fulfilled, V and Σ are suitable Hilbert spaces.

$$a : V \times V \to \mathbb{R} \qquad \text{and} \qquad b : V \times \Sigma \to \mathbb{R} \tag{97}$$

are continuous bilinear forms, and $L(v) : V \to \mathbb{R}$ is a continuous linear form. It should be noted that the classical Lax–Milgram Lemma cannot be applied. In fact we should apply the so-called Banach–Nečas–Babuška theorem also known as the generalized Lax–Milgram theorem, see, e.g., Ern and Guermond (2013). In summary the variational formulation has a unique solution if

1. The continuous linear form $a(\cdot, \cdot)$ is coercive on

$$K = \left\{ v \in V : b(q, v) = 0 \quad \forall\, q \in \Sigma \right\},$$

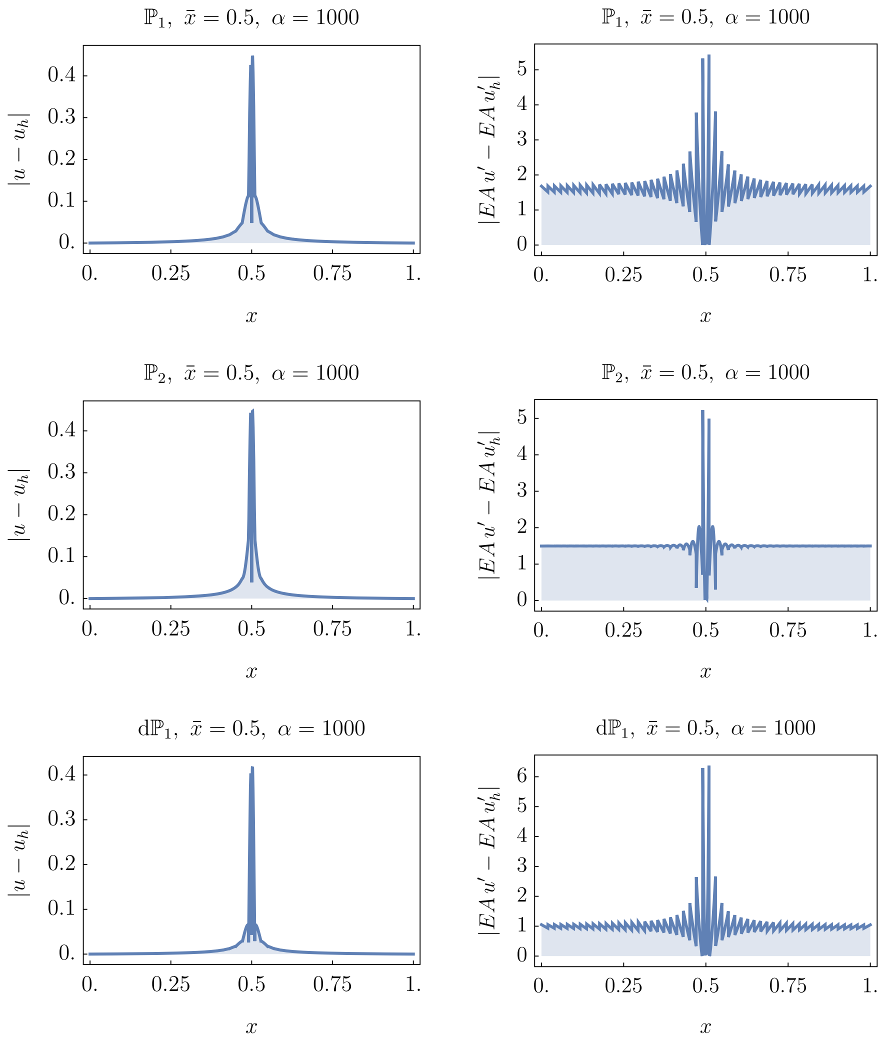


Fig. 15 Error plots for numerical solutions to displacements and longitudinal forces obtained by finite element discretizations of solutions spaces $\mathbb{P}_1$, $\mathbb{P}_2$, and $\mathrm{d}\mathbb{P}_1$ for $\alpha = 1000$, using 51 elements on each

i.e., there exist an $\alpha \in \mathbb{R}_+$, such that

$$a(v, v) \geq \alpha \|v\|_V^2 \qquad \forall\, v \in K\,,$$

and

2. the inf-sup condition, also known as LBB-condition (Ladyzhenskaya–Babuška–Brezzi), is verified, i.e., there exists a $\beta \in \mathbb{R}^+$, such that

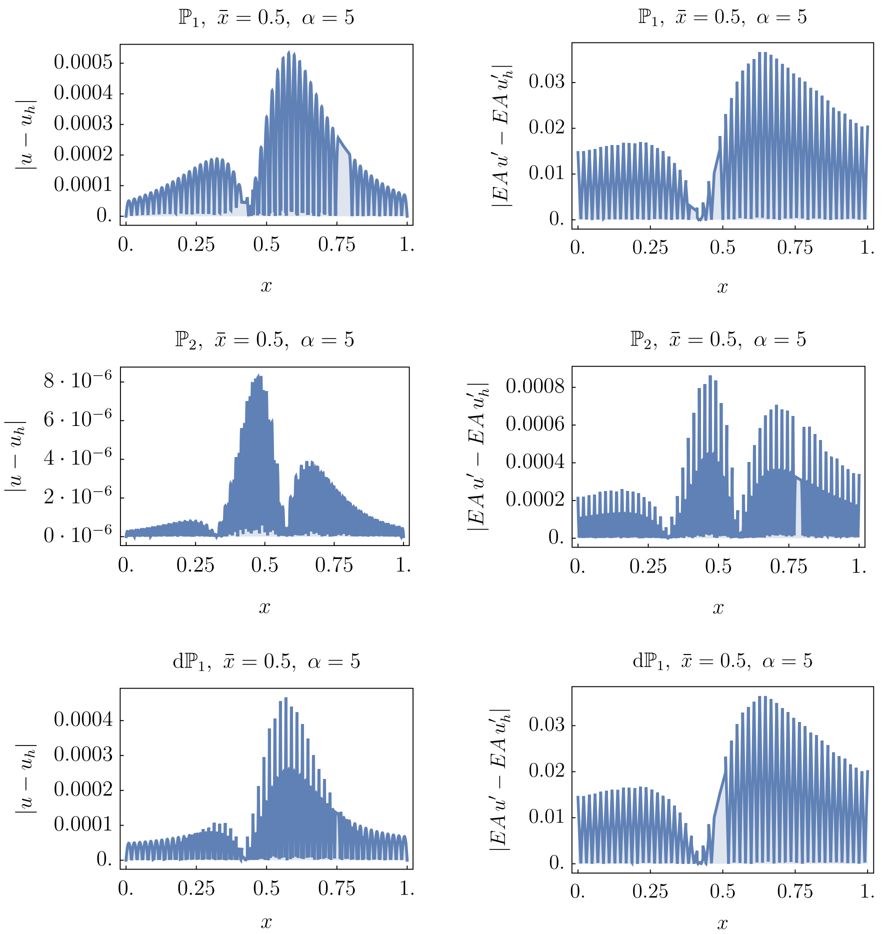


Fig. 16 Error plots for numerical solutions to displacements and longitudinal forces obtained by finite element discretizations of solutions spaces $\mathbb{P}_1$, $\mathbb{P}_2$, and $\mathrm{d}\mathbb{P}_1$ for $\alpha = 5$, using 51 elements on each

$$\inf_{q\in\Sigma\backslash 0}\ \sup_{v\in V\backslash 0} \frac{b(v,q)}{\|q\|_\Sigma\,\|v\|_V} \geq \beta\,.$$

Furthermore, there exists the a priori estimate for the solution

$$\|u\|_V \leq \frac{1}{\alpha}\|f\|_{V'} + \frac{1}{\beta}\left(1+\frac{C}{\alpha}\right)\|q\|_{\Sigma'} \tag{98}$$

and

$$\|p\|_\Sigma \leq \frac{1}{\beta}\left(1+\frac{C}{\alpha}\right)\|f\|_{V'}\,. \tag{99}$$

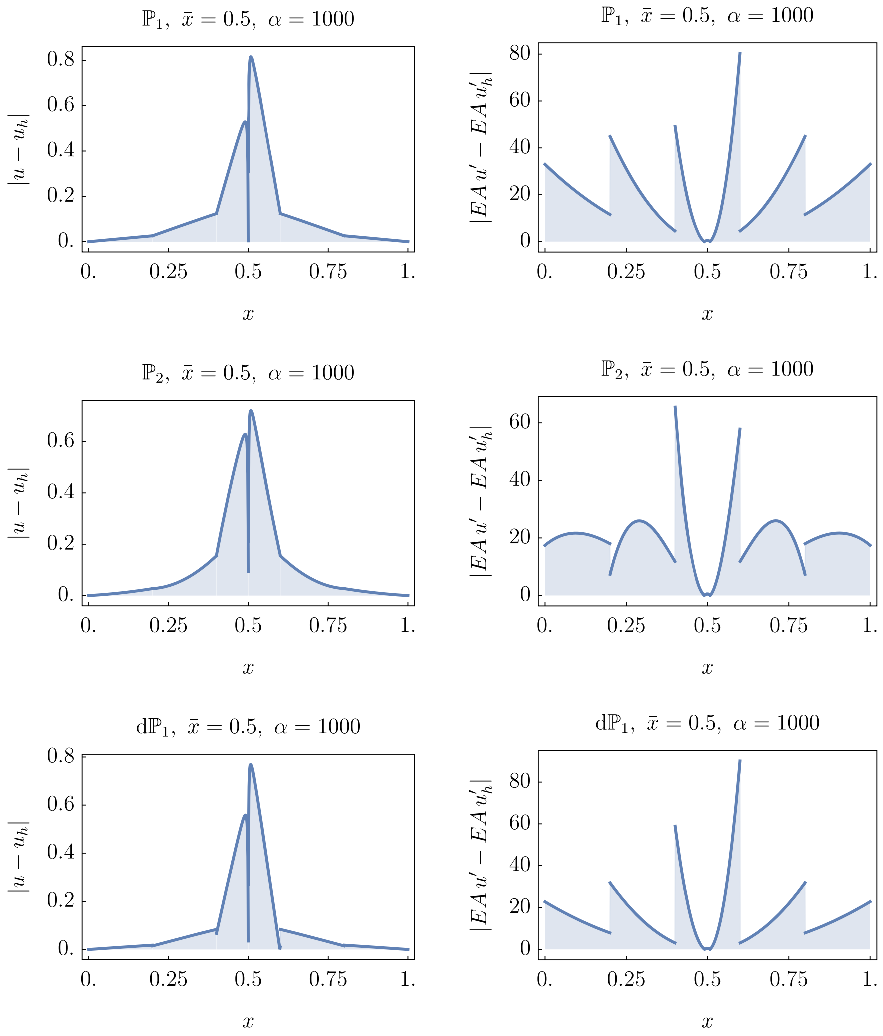


Fig. 17 Error plots for numerical solutions to displacements and longitudinal forces obtained by finite element discretizations of solutions spaces $\mathbb{P}_1$, $\mathbb{P}_2$, and $\mathrm{d}\mathbb{P}_1$ for $\alpha = 1000$, using 5 elements on each

6.2 *Treatment of Saddle Point Problems, Sensitization*

The discrete mixed problem is given by the matrix representation

$$\underbrace{\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B}^T \\ \boldsymbol{B} & \boldsymbol{0} \end{pmatrix}}_{\widehat{\boldsymbol{K}}} \underbrace{\begin{pmatrix} \boldsymbol{d}_u \\ \boldsymbol{d}_p \end{pmatrix}}_{\widehat{\boldsymbol{d}}} = \underbrace{\begin{pmatrix} \boldsymbol{r}_u \\ \boldsymbol{r}_p \end{pmatrix}}_{\widehat{\boldsymbol{r}}}, \tag{100}$$

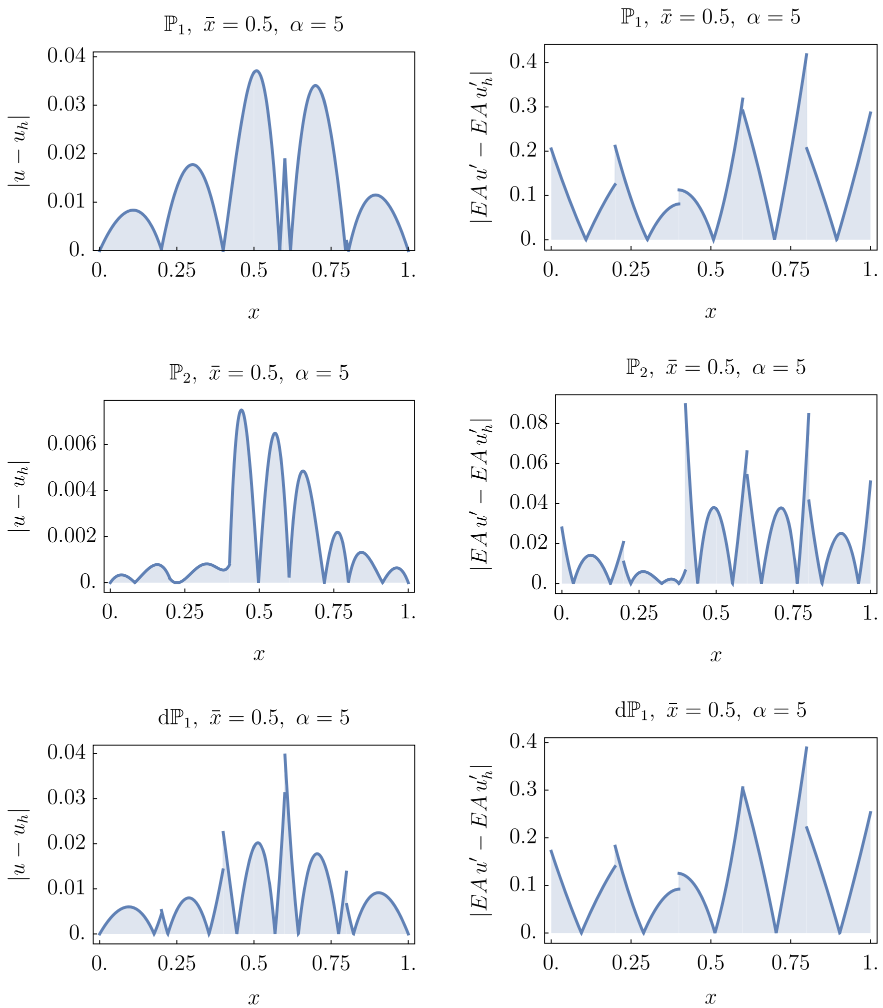


Fig. 18 Error plots for numerical solutions to displacements and longitudinal forces obtained by finite element discretizations of solutions spaces $\mathbb{P}_1$, $\mathbb{P}_2$, and $\mathrm{d}\mathbb{P}_1$ for $\alpha = 5$, using 5 elements on each

with $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{B} \in \mathbb{R}^{m\times n}$, $\boldsymbol{B}^T \in \mathbb{R}^{n\times m}$, $(\boldsymbol{d}_u, \boldsymbol{r}_u) \in \mathbb{R}^n$, $(\boldsymbol{d}_p, \boldsymbol{r}_p) \in \mathbb{R}^m$, $\widehat{\boldsymbol{K}} \in \mathbb{R}^{(n+m)\times(n+m)}$, and $(\widehat{\boldsymbol{d}}, \widehat{\boldsymbol{r}}) \in \mathbb{R}^{n+m}$.

For the **solvability** of (100) we postulate that the system has a unique solution for every right-hand side $\boldsymbol{r}_u$ and $\boldsymbol{r}_p$. Obviously, this condition is fulfilled if $\widehat{\boldsymbol{K}}$ is nonsingular. In other words, we must have a continuous dependency of the solution upon the right-hand side. Therefore, the existence of a constant c, satisfying

$$\|\boldsymbol{d}_u\|_? + \|\boldsymbol{d}_p\|_? \leq c\,(\|\boldsymbol{r}_u\|_? + \|\boldsymbol{r}_p\|_?)\,, \tag{101}$$

is required. However, the existence of c does not depend on the chosen norms, because in finite dimensions all norms are equivalent. Indeed the numerical values will depend on the dimension of the system. As examples consider $\boldsymbol{u} \in \mathbb{R}^n$ with the equivalent norms

$$\|\boldsymbol{u}\|_1 := \sum_i |u_i| \quad \text{and} \quad \|\boldsymbol{u}\|_2 := \sqrt{\sum_i |u_i|^2}\,. \tag{102}$$

For $n < \infty$ there exist the two positive constants c_1 and c_2 satisfying

$$c_1\|\boldsymbol{u}\|_2 \leq \|\boldsymbol{u}\|_1 \leq c_2\|\boldsymbol{u}\|_2 \quad \text{with optimal values} \quad c_1 = 1, c_2 = \sqrt{n}\,. \tag{103}$$

For $n \to \infty$ the latter inequality becomes unbounded from above.

In addition to the solvability condition we are interested in an estimate of the **stability** of (100): In general we consider a sequence of discrete saddle point problems with increasing mesh densities $h \to 0$ and therefore with increasing dimensions. Let $k = 0, 1, 2, 3, ..$ denote a sequence of discretizations with increasing mesh densities, i.e., we consider

$$\begin{pmatrix} \boldsymbol{A}_k & \boldsymbol{B}_k^T \\ \boldsymbol{B}_k & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{d}_u^k \\ \boldsymbol{d}_p^k \end{pmatrix} = \begin{pmatrix} \boldsymbol{r}_u^k \\ \boldsymbol{r}_p^k \end{pmatrix} \tag{104}$$

with $\boldsymbol{A}_k \in \mathbb{R}^{n_k \times n_k}$, $\boldsymbol{B}_k \in \mathbb{R}^{m_k \times n_k}$, ..., where the dimensions n_k, m_k increase with the sequence of k. In addition to the solvability condition we are interested in an estimate

$$\|\boldsymbol{d}_u^k\|_? + \|\boldsymbol{d}_p^k\|_? \leq c\,(\|\boldsymbol{r}_u^k\|_? + \|\boldsymbol{r}_p^k\|_?) \tag{105}$$

with a constant c independent on k, i.e., independent of the increasing mesh densities. For a meaningful analysis we have to specify the norms entering (105) carefully. For the stability requirement this choice is rather important, because constants appearing in the relations between equivalent (discrete) norms depend on the dimension of the problem, which goes to infinity with $h \to 0$. Obviously, the stability is a concept which has to be applied to a sequence of discretized boundary value problems.

Incidental remark: In order to get an impression on the well-known dependency of some characteristic values of a discretized system on the mesh density we perform an **eigenvalue analysis** of a one-dimensional elasticity problem, minimizing the discrete energy potential functional on $\mathcal{B} \in [0, 1] = \mathcal{B}_h = \bigcup_e \mathcal{B}^e$

$$\Pi(u_h) = \sum_{\mathcal{B}^e} \int_{\mathcal{B}^e} \left(\frac{1}{2} EA(u_h')^2 - f\, u_h\right) \mathrm{d}x\,. \tag{106}$$

In order to find the minimum we compute $\delta\Pi(u_h, \delta u_h) = 0$, with

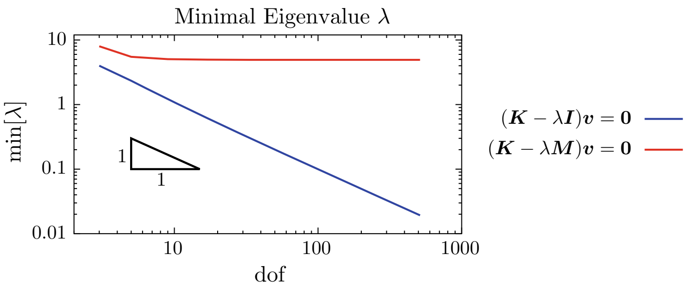


Fig. 19 Minimal eigenvalue depicted over number of unknowns (dof)

$$\delta\Pi(u_h, \delta u_h) = \sum_{\mathcal{B}^e} \int_{\mathcal{B}^e} \left(\delta u_h' \, EA \, u_h' - \delta u_h f\right) \mathrm{d}x \, . \tag{107}$$

We investigate the evolution of the minimal eigenvalue $\lambda_{\min}$ for the eigenvalue problems

$$(\boldsymbol{K}_k - \lambda_k \boldsymbol{I}_k)\boldsymbol{v}^k = \boldsymbol{0} \quad \text{and} \quad (\boldsymbol{K}_k - \lambda_k \boldsymbol{M}_k)\boldsymbol{v}^k = \boldsymbol{0} \, , \tag{108}$$

where $\boldsymbol{K}_k$ is the stiffness matrix, λ_k the associated eigenvalue to the eigenvector $\boldsymbol{v}^k$, $\boldsymbol{I}_k$ the identity and $\boldsymbol{M}_k$ the mass matrix. We consider a truss clamped at the edges, i.e., the Dirichlet boundary conditions $u(0) = u(1) = 0$, a Young's modulus $E = 1$, a cross section $A = 1$. Figure 19 shows the evolution of the minimal eigenvalues with respect to mesh refinement.

Obviously, the eigenvalue problem $(108)_1$ exhibits a decrease of the amplitude of the lowest eigenvalue with increasing mesh density (from $h = 1/2 \to h = 1/500$) whereas the formulation $(108)_2$ seems to offer a lower bound for $\min(\lambda_k)$. ■

6.3 A Saddle Point Problem-Finite-Dimensional Case

Let us concentrate on the saddle point problem for a linear incompressible material behavior. Starting from the general strong form of elasticity given in (62) and substituting the pressure as an additional unknown field as $p = \lambda \, \mathrm{tr}(\boldsymbol{\varepsilon}(\boldsymbol{u}))$ leads for the incompressible case $\lambda \to \infty$ to

$$\begin{aligned} &\mathrm{div}(2\,\mu\,\boldsymbol{\varepsilon}(\boldsymbol{u}) + p\,\boldsymbol{I}) + \boldsymbol{f} = \boldsymbol{0} \, , \\ &\mathrm{tr}(\boldsymbol{\varepsilon}(\boldsymbol{u})) = 0 \, , \\ &\boldsymbol{u} = 0 \quad \text{on} \quad \partial\mathcal{B}_D \, , \\ &\boldsymbol{\sigma} \cdot \boldsymbol{n} = \bar{\boldsymbol{t}} \quad \text{on} \quad \partial\mathcal{B}_N \, . \end{aligned} \tag{109}$$

The solution of this problem is similar to the saddle point of the potential

$$\Pi(\boldsymbol{u},p) = \int_{\mathcal{B}} (\mu\,\nabla_s\boldsymbol{u} : \nabla_s\boldsymbol{u} + p\ \mathrm{div}(\boldsymbol{u}) - \boldsymbol{f}\cdot\boldsymbol{u})\ \mathrm{d}V - \int_{\partial\mathcal{B}_N} \bar{\boldsymbol{t}}\cdot\boldsymbol{u}\,\mathrm{d}A \tag{110}$$

The variational approach and a finite element discretization leads to the finite-dimensional saddle point problem

$$\begin{aligned}
&\sum_{e=1}^{\mathrm{num_{ele}}}\Bigg\{\underbrace{\int_{\mathcal{B}^e} 2\mu\,\nabla_s\delta\boldsymbol{u}_h : \nabla_s\boldsymbol{u}_h\,\mathrm{d}v}_{a(\delta\boldsymbol{u}_h,\boldsymbol{u}_h)} + \underbrace{\int_{\mathcal{B}^e} p_h\,\mathrm{div}(\delta\boldsymbol{u}_h)\mathrm{d}v}_{b(\delta\boldsymbol{u}_h,p_h)}\Bigg\} - \underbrace{\int_{\mathcal{B}^e} \boldsymbol{f}_h\cdot\delta\boldsymbol{u}_h\,\mathrm{d}v}_{f(\boldsymbol{u}_h)} = 0\\
&\qquad\qquad\qquad\qquad\qquad\qquad \sum_{e=1}^{\mathrm{num_{ele}}} \underbrace{\int_{\partial\mathcal{B}^e} \delta p_h\,\mathrm{div}(\boldsymbol{u}_h)\mathrm{d}v}_{b(\boldsymbol{u}_h,\delta p_h)} = 0\,.
\end{aligned} \tag{111}$$

The equivalent problem is the stationarity requirement of the discrete Lagrange functional

$$\mathcal{L}_h(\boldsymbol{d}_u,\boldsymbol{d}_p) = \frac{1}{2}\boldsymbol{d}_u^T\boldsymbol{A}_h\,\boldsymbol{d}_u - \boldsymbol{d}_u^T\boldsymbol{f}_h + \boldsymbol{d}_p^T\boldsymbol{B}_h\,\boldsymbol{d}_u\,, \tag{112}$$

i.e., $\delta_{\boldsymbol{d}_u}\mathcal{L}_h = 0$ and $\delta_{\boldsymbol{d}_p}\mathcal{L}_h = 0$, which leads to

$$\begin{aligned}
\delta_{\boldsymbol{d}_u}\mathcal{L}_h &= \delta\boldsymbol{d}_u^T\{\boldsymbol{A}_h\boldsymbol{d}_u - \boldsymbol{f}_h + \boldsymbol{B}_h^T\boldsymbol{d}_p\} \qquad &\forall\,\delta\boldsymbol{d}_u\,,\\
\delta_{\boldsymbol{d}_p}\mathcal{L}_h &= \delta\boldsymbol{d}_p^T\{\boldsymbol{B}_h\,\boldsymbol{d}_u\} \qquad &\forall\,\delta\boldsymbol{d}_p\,,
\end{aligned} \tag{113}$$

where the field quantities have been substituted by the approximations

$$\boldsymbol{u}_h = \mathbb{N}_u\,\boldsymbol{d}_u\,,\quad \delta\boldsymbol{u}_h = \mathbb{N}_u\,\delta\boldsymbol{d}_u\,,\quad p_h = \mathbb{N}_p\,\boldsymbol{d}_p\,,\quad \delta p_h = \mathbb{N}_p\,\delta\boldsymbol{d}_p\,, \tag{114}$$

with $\mathbb{N}$ and $\boldsymbol{d}$ denoting the shape functions and nodal values corresponding to the displacements, pressure and its virtual counterparts. The solution of this set of algebraic equation follows from

$$\begin{bmatrix}\boldsymbol{A}_h & \boldsymbol{B}_h^T\\ \boldsymbol{B}_h & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{d}_u\\ \boldsymbol{d}_p\end{bmatrix} = \begin{bmatrix}\boldsymbol{f}_h\\ \boldsymbol{0}\end{bmatrix}, \tag{115}$$

with $\boldsymbol{A}_h \in \mathbb{R}^{n\times n}$, $\boldsymbol{A}_h = \boldsymbol{A}_h^T$ and positive definite, $\boldsymbol{B}_h \in \mathbb{R}^{m\times n}$, $\boldsymbol{f}_n \in \mathbb{R}^n$ and $m < n$. The physical interpretation of $m < n$ is obvious, there must be less constraints than "free" variables. Obviously, we have the "identical" structure as we obtain from equation (111). Now we have to ensure that (115) is solvable for all right-hand sides $\boldsymbol{f}_h$, following the remarks of Devendran et al. (2009). This is of course the fact if the whole matrix is invertible, i.e., nonsingular. Let us consider the congruent transformation, known as Sylvester's law of inertia

$$\begin{bmatrix}\boldsymbol{A}_h & \boldsymbol{B}_h^T\\ \boldsymbol{B}_h & \boldsymbol{0}\end{bmatrix} = \begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{B}_h\boldsymbol{A}_h^{-1} & \boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{A}_h & \boldsymbol{0}\\ \boldsymbol{0} & -\boldsymbol{B}_h\boldsymbol{A}_h^{-1}\boldsymbol{B}_h^T\end{bmatrix}\begin{bmatrix}\boldsymbol{I} & \boldsymbol{A}_h^{-1}\boldsymbol{B}_h^T\\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix}, \tag{116}$$

this transformation preserves the number of positive and negative eigenvalues (but not their numerical values). However, our system has full rank if the Schur complement

$$\boldsymbol{S}_h = -\boldsymbol{B}_h \boldsymbol{A}_h^{-1} \boldsymbol{B}_h^T \quad (= \boldsymbol{S}_h^T) \tag{117}$$

is nonsingular. In this case $\boldsymbol{S}_h$ is invertible and we can solve system (115). The full rank requirement is equivalent to

$$\boldsymbol{d}_p^T \boldsymbol{B}_h \boldsymbol{A}_h^{-1} \boldsymbol{B}_h^T \boldsymbol{d}_p > 0 \qquad \forall \boldsymbol{d}_p \in \mathbb{R}^m \backslash 0, \tag{118}$$

i.e., the Schur complement is negative definite. Due to the assumption that $\boldsymbol{A}_h$ and therefore $\boldsymbol{A}_h^{-1}$ is positive definite we argue

$$\boldsymbol{B}_h^T \boldsymbol{d}_p = 0 \qquad \text{iff} \quad \boldsymbol{d}_p = \boldsymbol{0}\,. \tag{119}$$

This means that the kernel of $\boldsymbol{B}^T$, i.e.,

$$\text{Ker}(\boldsymbol{B}_h^T) := \left\{ \boldsymbol{d}_p \in \boldsymbol{R}^m : \boldsymbol{B}_h^T \boldsymbol{d}_p = \boldsymbol{0} \right\}, \tag{120}$$

is trivial, i.e., the image of $\boldsymbol{B}^T$ is

$$\text{Im}(\boldsymbol{B}_h^T) = \mathbb{R}^m, \tag{121}$$

in other words $\boldsymbol{B}_h^T \in \mathbb{R}^{n \times m}$, with $m < n$ has full column rank. If this conditions are fulfilled the system (115) is invertible, i.e.,

$$\begin{bmatrix} \boldsymbol{A}_h & \boldsymbol{B}_h^T \\ \boldsymbol{B}_h & \boldsymbol{0} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{A}_h^{-1}(\boldsymbol{I} - \boldsymbol{B}_h^T \boldsymbol{S}_h^{-1} \boldsymbol{B}_h \boldsymbol{A}_h^{-1}) & \boldsymbol{A}_h^{-1} \boldsymbol{B}_h^T \boldsymbol{S}_h^{-1} \\ \boldsymbol{S}_h^{-1} \boldsymbol{B}_h \boldsymbol{A}_h^{-1} & \boldsymbol{S}_h^{-1} \end{bmatrix}. \tag{122}$$

Let $\beta^2 > 0$ denote the smallest singular value of $\boldsymbol{B}_h$. The condition that the smallest eigenvalue β, is greater than zero is directly related to the inf-sup condition of saddle point problems, which states

$$\inf_{\boldsymbol{d}_p \in \mathbf{R}^m \backslash 0} \; \sup_{\boldsymbol{d}_u \in \mathbf{R}^n \backslash 0} \frac{\boldsymbol{d}_p^T \boldsymbol{B}_h^T \boldsymbol{d}_u}{\|\boldsymbol{d}_p\| \, \|\boldsymbol{d}_u\|} \geq \beta^2 > 0 \tag{123}$$

or equivalently

$$\max_{\boldsymbol{d}_u \in \mathbf{R}^m \backslash 0} \frac{\boldsymbol{d}_p^T \boldsymbol{B}_h^T \boldsymbol{d}_u}{\|\boldsymbol{d}_u\|} \geq \beta^2 \, \|\boldsymbol{d}_p\| \qquad \forall \boldsymbol{d}_p \in \mathbb{R}^m\,. \tag{124}$$

The independency of the mesh size, as discussed for Eq. (105), demands here β to be bounded above zero for $h \to 0$.

Furthermore, we obtain the bounds

$$
\begin{aligned}
\|\boldsymbol{d}_u\|_{A_h} &\leq \|\boldsymbol{f}_h\|_{A_h^{-1}} \leq \frac{1}{\alpha}\|\boldsymbol{f}_h\| \\
\|\boldsymbol{d}_p\| &\leq \frac{1}{\beta}\|\boldsymbol{f}_h\|_{A^{-1}} \leq \frac{1}{\alpha\,\beta}\|\boldsymbol{f}\|\,,
\end{aligned}
\tag{125}
$$

with the energy norm $\|\boldsymbol{d}_u\|_A = \sqrt{\boldsymbol{d}_u^T \boldsymbol{A}\,\boldsymbol{d}_u}$. Obviously, if β is small the bound for $\boldsymbol{d}_p$ gets large.

Numerical Inf-Sup Test The numerical inf-sup test was proposed by Chapelle and Bathe (1993). In order to evaluate the inf-sup constant we use the fact that it is equivalent to the square root of the smallest eigenvalue of

$$
\left(\boldsymbol{B}_h\,\boldsymbol{M}_{u,h}^{-1}\,\boldsymbol{B}_h^T - \Lambda\,\boldsymbol{M}_{p,h}\right)\boldsymbol{d}_p = \boldsymbol{0}\,. \tag{126}
$$

with the global mass matrices $\boldsymbol{M}_{u,h}$, $\boldsymbol{M}_{p,h}$ as

$$
\boldsymbol{M}_{u,h} = \sum_{e=1}^{\text{num}_{\text{ele}}} \int_{\mathcal{B}^e} \mathbb{B}_u^T\,\mathbb{B}_u\,\mathrm{d}v \quad \text{and} \quad \boldsymbol{M}_{p,h} = \sum_{e=1}^{\text{num}_{\text{ele}}} \int_{\mathcal{B}^e} \mathbb{N}_p^T\,\mathbb{N}_p\,\mathrm{d}v, \tag{127}
$$

whereas, $\mathbb{B}_u$ contains spatial derivatives of the shape functions such that it holds $\boldsymbol{\varepsilon}_h = \mathbb{B}_u\,\boldsymbol{d}_u$. For exemplary purposes, the inf-sup stability is investigated by means of an inf-sup test on the example of the well-known Q_1P_0 and T_2P_0, representing elements with a discontinuous pressure approximation, and the T_2P_1, representing an element with a continuous pressure approximation, see Hood and Taylor (1974). The considered boundary value problem is a simple supported rectangle in 2D and brick in 3D, whereas a consecutive number of mesh refinements is considered. The statement on the inf-sup criterium of the considered elements are well known and the formal proofs can be found in the literature, e.g., Boffi et al. (2009). The T_2P_1 element is a famous representative of the Taylor–Hood family, which is well known to be inf-sup stable. In contrast the Q_1P_0 formulation is a text book example for an element which does not satisfy the inf-sup criterium neither in the two-dimensional nor in the three-dimensional case. Interestingly the T_2P_0 formulation fulfills the inf-sup condition in the two-dimensional case but fails in three dimensions. The depicted results in Fig. 20 approve numerically the statements on the inf-sup stability of the elements: In 2D the T_2P_0 and the T_2P_1 elements have an approximately constant $\Lambda > 0$, whereas Λ tends to zero for the Q_1P_0 element. In 3D, only the T_2P_1 seems to have a bounded value for Λ. Of course, a purely numerical check of the LBB condition is not sufficient, but it gives a first impression of the properties of the solution. To be save, a rigorous mathematical proof is needed.

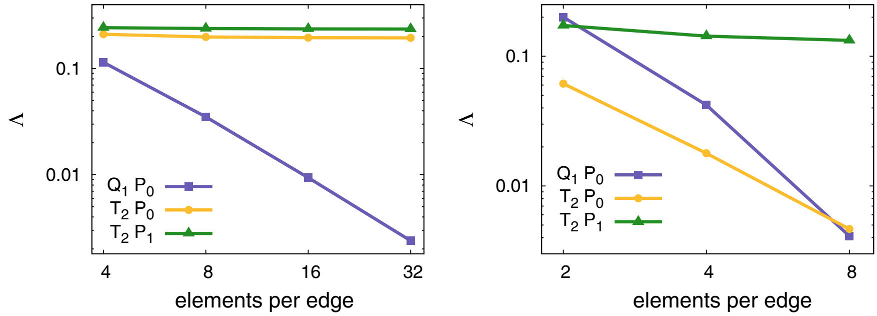


Fig. 20 Inf-Sup test results in 2D (left) and 3D (right). The inf-sup test is satisfied if Λ is bounded above zero, independent of the element size

Acknowledgements We thank the DFG for the financial support within the SPP 1748 *Reliable Simulation Techniques in Solid Mechanics. Development of Non-standard Discretization Methods, Mechanical and Mathematical Analysis*, project *Novel finite elements— Mixed, Hybrid and Virtual Element formulations* (Projectnumber: 255432295) (SCHR 570/23-2). I would also like to thank Nils Viebahn for helpful discussions and his help with the manuscript and Sascha Maassen and Rainer Niekamp for the implementation of the H^{-1}procedure and accompanying discussions.

A Sobolev and Hilbert Spaces

In the following we will use the Sobolev and Hilbert Spaces, they are based on the space of square integrable functions on $\mathcal{B}$:

$$L^2(\mathcal{B}) = \left\{u : \|u\|^2_{L^2(\mathcal{B})} = \int_{\mathcal{B}} |u|^2 \mathrm{d}v < +\infty\right\}. \tag{128}$$

Let $s \geq 0$ be a real number, the standard notation for a Sobolev space is $H^s(\mathcal{B})$ and $H^s(\partial\mathcal{B})$ with the inner products and norm

$$(u, u)_{s,\mathcal{B}}\,, \quad (u, u)_{s,\partial\mathcal{B}} \quad \text{and} \quad \|u\|_{s,\mathcal{B}}\,, \quad \|u\|_{s,\partial\mathcal{B}}\,, \tag{129}$$

respectively. For $s = 0$ the space $H^0(\mathcal{B})$ represents the Hilbert space $L^2(\mathcal{B})$ of all square integrable functions, i.e.,

$$L^2(\mathcal{B}) = H^0(\mathcal{B}) = \{u \in L^2(\mathcal{B})\}. \tag{130}$$

If s is a *positive integer* the spaces $H^s(\mathcal{B})$ consist of all square integrable functions whose derivatives up to the order s are also square integrable, i.e.,

$$H^s(\mathcal{B}) = \left\{u + \sum_{\alpha=1}^{s} \mathrm{D}^\alpha u \in L^2(\mathcal{B})\right\}. \tag{131}$$

Here we shall use the semi-norms

$$|u|_{k,\mathcal{B}} := \sqrt{\sum_{\alpha=k} |\mathrm{D}^\alpha u|^2_{L^2(\mathcal{B})}}, \qquad k = 0, 1, \ldots, s\,, \tag{132}$$

and the norm

$$\|u\|_{s,\mathcal{B}} := \sqrt{\sum_{k \le s} |u|^2_{k,\mathcal{B}}}. \tag{133}$$

Critism: This expression for the norm does not take into account a typical length scale l of the problem, i.e., we are adding, for example, a square integrable function $|u|^2_{L^2(\mathcal{B})}$ and its square integrable derivative $|u'|^2_{L^2(\mathcal{B})}$. Without any physically meaningful parameters these expression is hardly to interpret. This could be avoided by using the expression

$$\|u\|_{s,\mathcal{B}} := \sqrt{\sum_{k \le s} l^{\mathrm{d}k}\ |u|^2_{k,\mathcal{B}}}\,, \tag{134}$$

where d characterizes the dimension of $\mathcal{B} \subset \mathbb{R}^{\mathrm{d}}$, Boffi et al. (2013).

With D^α as the α-st weak differential operator. Thus the often used spaces $H^1(\mathcal{B})$ and $H^1_0(\mathcal{B})$ are defined by

$$H^1(\mathcal{B}) = \left\{u + \mathrm{D}^1 u \in L^2(\mathcal{B})\right\}, \tag{135}$$

and

$$H^1_0(\mathcal{B}) = \left\{u \in H^1(\mathcal{B})\ :\ u = 0 \ \text{ on } \ \partial\mathcal{B}\right\}. \tag{136}$$

For completeness we introduce the spaces $H^2(\mathcal{B})$ and $H^2_0(\mathcal{B})$ defined by

$$H^2(\mathcal{B}) = \left\{u + \mathrm{D}^1 u + \mathrm{D}^2 u \in L^2(\mathcal{B})\right\}, \tag{137}$$

and

$$H^2_0(\mathcal{B}) = \left\{u \in H^2(\mathcal{B})\ :\ u = 0 \ \text{ and } \ \frac{\partial u}{\partial n} = 0 \ \text{ on } \ \partial\mathcal{B}_u\right\}. \tag{138}$$

For negative superscripts, i.e., $H^{-s}(\mathcal{B})$ with $s > 0$, the spaces are identified with the duals of $H^s_0(\mathcal{B})$:

$$H^{-s}(\mathcal{B}) = (H^s_0(\mathcal{B}))'\,. \tag{139}$$

For example, the norm associated to $H^{-1}(\mathcal{B})$, which is the dual of $H^1_0(\mathcal{B})$, is defined as

$$\|u\|_{-1,\mathcal{B}} = \min_{v \in H_0^1(\mathcal{B})\backslash 0} \frac{(u, v)_{0,\mathcal{B}}}{\|v\|_{1,\mathcal{B}}} \,. \tag{140}$$

The norm associated to $H^{-1/2}(\partial\mathcal{B})$, the dual of $H_0^{1/2}(\partial\mathcal{B})$, is defined as

$$\|u\|_{-1/2,\partial\mathcal{B},0} = \min_{v \in H^{1/2}(\partial\mathcal{B})\backslash 0} \frac{(u, v)}{\|v\|_{1/2,\partial\mathcal{B}}} \,. \tag{141}$$

The Hilbert space $H_0^m(\mathcal{B})$ is a closed subspace of $H^m(\mathcal{B})$; furthermore is $H_0^0(\mathcal{B}) = L_2(\mathcal{B})$.

$$\begin{aligned} \ldots H^{-2}(\mathcal{B}) \supseteq H^{-1}(\mathcal{B}) \supseteq L_2(\mathcal{B}) \supseteq H_0^1(\mathcal{B}) \supseteq H_0^2(\mathcal{B}) \ldots \\ \ldots \|u\|_{-2,\mathcal{B}} \leq \|u\|_{-1,\mathcal{B}} \leq \|u\|_{0,\mathcal{B}} \leq \|u\|_{1,\mathcal{B}} \leq \|u\|_{2,\mathcal{B}} \ldots \end{aligned} \tag{142}$$

For tensorial Sobolev spaces, e.g., the three-dimensional tensor product space

$$H^s(\mathcal{B}) \times H^s(\mathcal{B}) \times H^s(\mathcal{B}) \tag{143}$$

we use the abbreviation

$$[H^s(\mathcal{B})]^3 = \prod_{i=1}^{3} H^s(\mathcal{B}) \qquad \text{and analogously} \qquad [L^2(\mathcal{B})]^3 = \prod_{i=1}^{3} L^2(\mathcal{B}) \,. \tag{144}$$

Let $\boldsymbol{u} \in \mathbb{R}^3$ and set the Hilbert space

$$H(\text{div}; \mathcal{B}) = \left\{ \boldsymbol{u} \in [L^2(\mathcal{B})]^3 \;:\; \text{div}\, \boldsymbol{v} \in L^2(\mathcal{B}) \right\}, \tag{145}$$

with the associated norm

$$\|\boldsymbol{v}\|_{H(\text{div};\mathcal{B})} = \left\{ \|\boldsymbol{v}\|^2 + |\,\text{div}\, \boldsymbol{v}|^2 \right\}^{1/2} . \tag{146}$$

References

Auricchio, F., Brezzi, F., & Lovadina, C. (2004). Mixed finite element methods. In E. Stein, R. de Borst, & T. J. R. Hughes (Eds.), *Encyclopedia of computational mechanics* (Chap. 9, pp. 238–277). Wiley and Sons.

Bathe, K.-J. (1996). *Finite element procedures*. New Jersey: Prentice Hall.

Becker, E. B., Carey, G. F. & Oden, J. T. (1981). *Finite elements, an introduction: Volume I.* Prentice-Hall.

Berdichevsky, V. L. (2009). *Variational principles of continuum mechanics*. Springer.

Boffi, D., Brezzi, F., & Fortin, M. (2009). Reduced symmetry elements in linear elasticity. *Communications on Pure and Applied Analysis*, *8*, 95–121.

Boffi, D., Brezzi, F., & Fortin, M. (2013). *Mixed finite element methods and applications*. Heidelberg: Springer.

Braess, D. (1997). *Finite elemente* (2nd ed.). Berlin: Springer.

Brenner, S. C. & Scott, L. R. (2002). The mathematical theory of finite element methods. In *Texts in applied mathematics* (Vol. 15, 2nd edition). New York: Springer.

Brezzi F. (1974). On the existence, uniqueness and approximation of saddle-point problems arising from lagrangian multipliers. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique, 8*(2), 129–151.

Chapelle, D., & Bathe, K.-J. (1993). The inf-sup test. *Computers and Structures*, *47*, 537–545.

Crouzeix, M., & Raviart, P.-A. (1973). Conforming and nonconforming finite element methods for solving the stationary stokes equations i. *Revue francaise d'automatique, informatique, recherche operationnelle*, *7*(3), 33–75.

Devendran, D., May, S., & Corona, E. (2009). Computational fluid dynamics reading group: Finite element methods for stokes and the infamous inf-sup condition.

Ern, A., & Guermond, J.-L. (2013). *Theory and practice of finite elements* (Vol. 159). Springer Science & Business Media.

Gockenbach, M. S. (2006). *Understanding and implementing the finite element method*. SIAM.

Goebbels, S. (2015). An inequality for negative norms with application to errors of finite element methods.

Hood, P., & Taylor, C. (1974). Navier-stokes equations using mixed interpolation. In J. T. Oden, O. C. Zienkiewicz, R. H. Gallagher, & C. Taylor (Eds.), *Finite element methods in flow problems* (pp. 121–132). UAH Press.

Hughes, T. J. R. (1987). *The finite element method*. Englewood Cliffs, New Jersey: Prentice Hall.

Kendall, R. P., & Wheeler, M. F. (1976). A Crank-Nicolson H^{-1}-Galerkin procedure for parabolic problems in a single-space variable. *SIAM Journal on Numerical Analysis*, *13*, 861–876.

Oden, J. T., & Carey, G. F. (1983). *Finite elements. Mathematical aspects. Volume IV*. Prentice-Hall.

Oden, J. T., & Reddy, J. N. (1976). *An introduction to the mathematical theory of finite elements*. Wiley.

Rachford, H. H., Wheeler, J. R., & Wheeler, M. F. (1974). An H^{-1}-Galerkin procedure for the two-point boundary value problem. In C. de Boor (Ed.), *Mathematical Aspects of Finite Elements in Partial Differential Equations, Proceedings of Symposium, Conducted by the Mathematics Research Center, The University of Wisconsin-Madison*. Academic Press.

Royden, H. L. (1968). *Real analysis*. New York: Macmillan.

Thomée, V. (1980). Negative norm estimates and superconvergence in Galerkin methods for parabolic problems. *Mathematics of Computation*, *34*(149), 93–113.

Thomée, V. (2006). *Galerkin finite element methods for parabolic problems*. Berlin, Heidelberg: Springer.

Wriggers, P. (2008). *Nonlinear finite element methods*. Springer.