[1994/12/01]

K. Böhmer

Finite Element and Collocation Methods with Variational Crimes

Preliminary Version

November 15, 2001

Springer-Verlag
Berlin Heidelberg New York
London Paris Tokyo
Hong Kong Barcelona
Budapest

Contents

1. Introduction

In this Booklet we present Finite Element Methods (FEMs) for elliptic equations. We apply FEs with and without variational crimes to linear and nonlinear operators. We start with the linear case. So, let

$$A : \mathcal{U} \to \mathcal{V}', \quad A : \mathcal{U}_b \to \mathcal{V}_b' \text{ bijective}, \ \mathcal{U}_b \subset \mathcal{U}, \quad \mathcal{V}_b \subset \mathcal{V} \text{ Banach spaces}. \tag{1.1}$$

We test with v, $\forall\ v \in \mathcal{V}_b$ to determine the exact solution, u_0,

$$\begin{aligned} u_0 \in \mathcal{U}_b \subset \mathcal{U},\ a(u_0, v) &= < Au_0, v >_{\mathcal{V}' \times \mathcal{V}} \\ &= \ < f, v >_{\mathcal{V}' \times \mathcal{V}} \forall \ \ v \in \mathcal{V}_b \subset \mathcal{V}. \end{aligned} \tag{1.2}$$

In the crime free world this exact variational problem is replaced by the discrete form. Determine the discrete or approximate solution, u_0^h, s.t.

$$u_0^h \in \mathcal{U}_b^h \subset \mathcal{U}_b \subset \mathcal{U} : a(u_0^h, v^h) = f(v^h) \quad \forall v^h \in \mathcal{V}_b^h \subset \mathcal{V}_b \subset \mathcal{V}; \tag{1.3}$$

here the $\mathcal{U}_b, \mathcal{U}_b^h$ and $\mathcal{V}_b, \mathcal{V}_b^h$, represent functions and test functions, resp., satisfying the appropriate boundary, continuity and smoothness conditions; the index and exponent, $_b$ and ^{h}, indicate boundary conditions and finite dimensional subspaces, resp. We call (1.3) a *conforming FEM.*

Non conforming FEMs or, more generally, FEs with *variational crimes* violate the above variational approximate equation (1.3) in different ways: Either the approximating subspaces $\mathcal{U}_b^h, \mathcal{V}_b^h$ violate the boundary conditions, such that $\mathcal{U}_b^h \not\subset \mathcal{U}_b, \mathcal{V}_b^h \not\subset \mathcal{V}_b$. Or they violate continuity or smoothness conditions in the form $\mathcal{U}_b^h \not\subset \mathcal{U}, \mathcal{V}_b^h \not\subset \mathcal{V}$, e.g. $\mathcal{U}_b^h, \mathcal{V}_b^h \not\subset C(\Omega)$, hence $\mathcal{U}_b^h, \mathcal{V}_b^h \not\subset \mathcal{U} = \mathcal{V} = H^1(\Omega)$ for elliptic equations of second order. Or, the exact scalar products or pairings, defining the projectors in the Petrov-Galerkin methods, can be approximated or modified, e.g, by quadrature formulas and related to collocation methods. Finally, the exact A may have to be replaced by approximations $\tilde{A}_h$, e.g.. in difference methods. Hence instead of the above (1.2) we have to modify $a(\cdot, \cdot)$ into $a^h(\cdot, \cdot)$ and solve

$$u^h \in \mathcal{U}_b^h : a^h(u_0^h, v^h) = f^h(v^h) \ \ \forall\ v^h \in \mathcal{V}_b^h. \tag{1.4}$$

The goal in this Booklet is the proof of stability and convergence for *conforming* and *non conforming FEMs* (, a special case of *non conforming Petrov-Galerkin methods*,) with the different forms of variational crimes.

We concentrate on FEs, but give a short presentation for spectral methods as well. Quadrature approximations and collocation methods, specific variational crimes, are important for spectral methods.

We use throughout the Booklet the weak and strong formulation of the problem. For discretizations they usually yield different approaches. But it allows to study variational crimes in a natural way. At the same time it allows to develop conditions for the FEs, which grant only slightly perturbed discrete weak and strong formulations and discrete solutions.

Furthermore, as interesting, we obtain, under specific conditions, a class of *collocation methods on non degenerate subdivisions* which is strongly related to unusual ("super crime") FEMs. To my knowledge, Doedel, [27, 28], is the first to discuss these methods for a special case. Goldlücke has proved in his Diplomarbeit (in preparation), with the tools of this Booklet, stability and convergence for a very special case. The tools for the general case are available in this Booklet, however will need a lot of additional ideas and work. This class includes methods of high order applicable to general elliptic operators. They are important for parameter-dependent problems, where turning points and singularities have to be determined. With the techniques in [15, 16, 6, 7] these results are extended to the case of bifurcation numerics in Chapter 8.

As in [15, 16] we introduce a new, relatively straight forward and simple way to prove stability for a relatively general class of operator equations and discretization methods: We combine compact perturbations of monotone operators with approximation properties(error and inverse estimates). Recently, we succeeded, [10], to prove stability and convergence for all these operators for wavelets. This can again be applied to the bordered systems, required for bifurcation numerics.

As a consequence of the above violations a direct comparison of $a(\cdot,\cdot)$ on $\mathcal{U}_b \times \mathcal{V}_b$ and on $\mathcal{U}_b^h \times \mathcal{V}_b^h$ is not possible. So, in contrast to [51], we have to prove stability *and* estimate the consistency error and finally combine both to get the desired convergence. In fact, we will have to interrelate stability and consistency results on different levels to obtain the final results.

The Booklet is organized as follows: In Chapter 2 we start with a standard presentation of the approximation theory for finite elements (FEs). A new result seems to be existence of an "anti-crime operator" E^h. It transforms an u^h with variational crimes to an $E^h u^h$, s.t. $\|u^h - E^h u^h\|_{H^1(\Omega)} \to 0$.

Chapter 3 starts with a simple example, essentially the Laplacian, motivating the definition of general elliptic differential operators with their strong and weak forms. Then for a $\mathcal{U}_b$ coercive $a(\cdot,\cdot)$, hence $a(u,u) \geq \alpha \|u\|_{\mathcal{U}^2}$ with $\alpha > 0$, the weak solutions u_0^h in (1.3) converge to the solution u_0 in (1.2).

Chapter 4 starts presenting examples for finite element methods with different types of variational crimes as mentioned above. We extend it to finite elements and spectral methods for general elliptic operators. This allows to develop *general concepts for discretization,* including variational crimes. Here

we have to distinguish the *variational consistency errors*, familiar in the finite element community and the *classical consistency errors* as introduced in other general approaches, see below. These general approaches simplify the study of nonlinear problems. It yields the standard result "Consistency and Stability implies Convergence".

Chapter 5 presents generalized Strang Lemmas. It indicates for different variational crimes the consistency terms which measure the influence of the crime.

Chapter 6 presents the estimates for these consistency terms. They are caused by violated boundary conditions, violated continuity, isoparametric approximations, approximate projectors and operators as quadrature and, finally, collocation. In addition, we show discrete coercivity for these cases, if the original $a(\cdot,\cdot)$ is $\mathcal{U}_b$ coercive. This allows the formulation of convergence results for this case.

Chapter 7 deals with stability and convergence for conforming and non conforming methods. We start with the result from the last Chapter: A $\mathcal{U}_b$ -coercive $a(\cdot,\cdot)$, hence a monotone operator, induces a discrete $\mathcal{U}_b^h$ coercive $a^h(\cdot,\cdot)$, hence a stable discretization.. Then we proceed to compact perturbations. Many different methods, e.g., FEM-, spectral- and difference methods, see [16], satisfy the corresponding necessary stability and consistency conditions. So one obtains convergence for all these methods.

Chapter 8 applies the earlier results to the case of *bordered systems*. They have to be used for bifurcation numerics and can be interpreted as compact perturbations of a slightly extended original operator. We show, that the solutions of the numerical Liapunov-Schmidt equations converge to those of the original Liapunov-Schmidt equations for the well known Jepson/Spence conditions, see [40]. The convergence of the bifurcation scenarios is presented, e.g., in [9, 6, 15, 7].

The last Chapter 9 is devoted to the *linearized Navier-Stokes operator*. For the not so interesting case of moderate viscosity, it is a compact perturbation of the well studied Stokes operator. Otherwise, we refer to general stability results. In both cases, the bordered systems are stable and thus bifurcation numerics are safe.

For our approach to treat conforming and non conforming methods, we use a mix of different concepts. It is strongly influenced by many earlier papers. Discrete convergence has been studied by Stummel, [58, 59, 60, 61], Reinhardt [50] and Vainikko, [66], admissible discretization methods by Stetter, [57], (inner and outer) admissible approximation schemes in Petryshyn [43, 44, 45, 46] and nicely presented and extended in Zeidler [68].

Brezzi-Rappaz-Raviart have applied a similar version to finite dimensional approximations of nonlinear problems, including limit and simple bifucation points, however excluding many operator equations and nowadays important discretization methods, [19, 20, 21]. Rappaz and Raugel studied special cases of finite-dimensional approximation of bifurcation problems at a

multiple eigenvalue, [47, 48], Raugel, [49], extended these results to symmetric problems. Crouzeix and Rappaz gave a short survey book on numerical approximation in bifurcation theory, [26], Calos and Rappaz presented numerical approximation in nonlinear and bifurcation problems in [22]. Here, in contrast to the earlier [19, 20, 21, 47, 48], two projectors for the analysis of Petrov-Galerkin methods for nonlinear problems are introduced, but are not employed for the bifurcation analysis. Furthermore, there is no reference to other approaches as in Griewank and Reddien,[32, 35, 34, 36, 33] who prove convergence for general bifurcation scenarios under the conditions of [19, 20, 21] and Jepson/Spence, [40], who study bifurcation for perturbed operators, not covering discretization methods. All the last papers, starting with [19], do not allow to prove convergence for bifurcation of more complicated problems as Navier-Stokes or porous media or nonstandard finite element or spectral methods, see the discussion in [15, 16]. Approaches to treat these more complicated cases use the concepts of consistent differentiability and modified or bordered stability, see [8, 14, 9, 1, 2, 3, ?, 6, 7].

Our presentation is fairly complete for elliptic operators of order 2. A generalization to elliptic operators of order $2m$ works well in Chapters 2- 5 and it will work well in most of Chapter 7 again. However, the construction of the anti crime operator in Chapter 2, and the consistency estimates in Chapter 6 are based on specific $2m = 2-$ techniques. So, only these parts would have to be worked out for the general case of elliptic operators of order $2m$.

Throughout the Booklet, all convergence results require a sufficiently small discretization parameter, h.

2. Approximation Theory for Finite Elements

We indicate a discretization by FEs by a parameter h. In the following we indicate the essential *parameters influencing constants*, C, e.g., $C_{h,\mu}$ depends upon h, μ. Sometimes, we omit parameters to show the *independence of constants*, in particular of the h or as a short notation sometimes. All the results in this Chapter are strictly local. So it would be technically more involved, however possible, to get these results for a strong local refinement of the subdivision $\mathcal{T}^h$ of Ω and use the corresponding local h. This is a consequence of only requiring *non degenerate subdivisions*. Even $h-p-$methods would fit into our frame work. Then the local approximating spaces, here $\mathcal{P}$, and norms and estimates would have to be updated approximately. We give a few results which are basic for the whole finite element approach. We *assume throughout the whole Booklet*

$$\Omega \subset \mathbb{R}^n \text{ to be a bounded domain, with piecewise smooth boundary, (2.1)}$$

hence Ω is open.

2.1 Basic Sobolev Space Results

Important tools for PDEs and approximation properties are the following Theorems. Here (2.1) is sometimes modified.

Theorem 2.1.1. Extension operator, *[68],p305-306, [18]: Let $\Omega \subset \mathbb{R}^n$ have a Lipschitz boundary, let $k \in \mathbb{N}_0$ and $p \in \mathbb{R}$ with $1 \leq p \leq \infty$. Then there exists a mapping $E : W_p^k(\Omega) \to W_p^k(\mathbb{R}^n)$ and a constant, C, independent of h, such that*

$$Ev|_\Omega = v \text{ and } \|Ev\|_{W_p^k(\mathbb{R}^n)} \leq C\|v\|_{W_p^k(\Omega)} \ \forall v \in W_p^k(\Omega). \tag{2.2}$$

This E can be constructed independent of k. Moreover, the function Eu is C^∞ on $\mathbb{R}^h \setminus \overline{\Omega}$.

Theorem 2.1.2. Extension operator,*[29], p 136: Let $\Omega \subset \mathbb{R}^n$ be a $C^{k,\alpha}$ domain, let $k \in \mathbb{N}$ and let Ω' be an open set with $\overline{\Omega} \subset \Omega'$. Then there exists a bounded linear extension $E_D : C_D^{k,\alpha}(\overline{\Omega}) \to C_0^{k,\alpha}(\Omega')$, with $_0$ and*

$_D$ indicating trivial and nontrivial Dirichlet boundary conditions for Ω'. So there is a constant, $C = C(k, \Omega, \Omega')$, such that

$$E_D v|_\Omega = v \text{ and } \|E_D v\|_{C^{k,\alpha}(\Omega')} \le C\|v\|_{C^{k,\alpha}(\overline{\Omega})} \ \forall v \in C^{k,\alpha}(\overline{\Omega}), \quad (2.3)$$

where the $\|\cdot\|_{C^{k,\alpha}(\Omega')}$ denote the maximum norm for C^k functions with with the corresponding bounds for the divided differences of the α-Hoelder continuous $k-th$ derivatives.

Theorem 2.1.3. Trace Theorem: *Let $\Omega \subset \mathbb{R}^n$ have a Lipschitz boundary and let $p \in \mathbb{R}$ with $1 \le p \le \infty$. Then there exists a constant, C, such that*

$$\|v\|_{L^p(\partial\Omega)} \le C\|v\|^{1-1/p}_{L^p(\Omega)} \cdot \|v\|^{1/p}_{W^1_p(\Omega)} \quad \forall v \in W^1_p(\Omega). \quad (2.4)$$

For the following Bramble-Hilbert-Lemma we need the standard *Sobolev norms, semi norms* and notations for *partial derivatives* and *multi-indices*

$$\|v\|_{W^k_p(\Omega)} := \left(\sum_{|\alpha|\le k} \|D^\alpha v\|^p_{L^p(\Omega)} \right)^{1/p}, \ |v|_{W^k_p(\Omega)} := \left(\sum_{|\alpha|= k} \|D^\alpha v\|^p_{L^p(\Omega)} \right)^{1/p}. \quad (2.5)$$

For $w \in W^m_p(\Omega)$ the standard derivatives and hence Taylor polynomials $T^m_y w(x) = \sum\limits_{|\alpha|<m} D^\alpha w(y)(x-y)^\alpha/\alpha!$ are not defined. So we use the $D^\alpha w \in L^2(\Omega)$, $|\alpha| < m$ directly and introduce the *averaged Taylor polynomials* $Q^m w$ for $w \in W^m_p(\Omega)$:

Definition 2.1.4. *Let $\Omega \subset \mathbb{R}^n$ or $G \subset \mathbb{R}^n$ have finite* diam G *and $B := B_\rho(x_0)$ with $\overline{B} \subset G$ be a ball such that for all $x \in G$ the closed convex hull of $\{x\} \cup \overline{B} \subset G$. Then G is called* star-shaped w.r.t. B. *Let*

$$\rho_{\max} := \sup\{\rho : G \text{ is star-shaped w.r.t. a ball } B \text{ of radius } \rho\}, \text{ then } (2.6)$$

$$\gamma = \text{ diam } G/\rho_{\max} \text{ is called the chunkiness parameter of } G. (2.7)$$

Now, let G be star-shaped *w.r.t. B and let ϕ be a* cut-off function *for B, that is (i) supp $\phi = \overline{B}$ and (ii) $\int_{\mathbb{R}^n} \phi(x)dx = 1$. We obtain the well defined* averaged Taylor polynomials, *see [18],*

$$Q^m w(x) := \int_B T^m_y w(x)\phi(y)dy \text{ with} \quad (2.8)$$

$$T^m_y w(x) = \sum_{|\alpha|<m} D^\alpha w(y)(x-y)^\alpha/\alpha!.$$

$Q^m w$ is a polynomial of degree $m-1$.

Obviously, triangles are always star-shaped. Mind that only $D^\alpha u$ with $|\alpha| < m$ are employed. A standard example for a cut-off function is

$\phi(x) = c \exp\,(-(1-(|x-x_0|/\rho)^2)^{-1})$ for $x \in B$ and $\phi(x) \equiv 0$ outside B.

We only consider *star-shaped* G (or $T \in \mathcal{T}^h$, see below) with bounded *chunkiness parameter,* γ, see (2.7).

Theorem 2.1.5. Bramble-Hilbert Lemma: *Let* G *with* diam $G = d$ *be star-shaped w.r.t.* B *of radius* ρ *s.t.* $\rho > (1/2)\rho_{\max}$, G *have the chunkiness parameter* γ, *and* $Q^m w$ *be the averaged Taylor polynomial. Then for all* $w \in W_p^m(G), \ell = 0, 1, \ldots, m,\ 1 \le p \le \infty$, *there exist constants* $C_{m,n,\gamma}$ *s.t.*

$$|w - Q^m w|_{W_p^\ell(G)} \le C_{m,n,\gamma} d^{m-l} |w|_{W_p^m(G)} \quad or$$
$$\|w - Q^m w\|_{W_p^\ell(G)} \quad \le C_{m,n,\gamma} d^{m-\ell} |w|_{W_p^m(G)}$$

An important tool to prove the invertibility of an operator A or stability of its discrete counterpart, A^h, is the so called $\inf - \sup$ condition. Although the following (2.9) would allow $u \in \mathcal{U}, v \in \mathcal{V}$ without boundary conditions, the (2.10) ff. require boundary conditions. Hence we formulate the following Theorem for $\mathcal{U}_b, \mathcal{V}_b$, where the index, $_b$, e.g., $\mathcal{U}_b$ indicates the boundary conditions. Nevertheless, the norms are indexed by the original Banach spaces without boundary conditions, e.g., $\|u\|_{\mathcal{U}}$.

Theorem 2.1.6. Brezzi-Babuska condition: *Let* $\mathcal{U}_b, \mathcal{V}_b$ *be Banach spaces,* $A \in \mathcal{L}(\mathcal{U}_b, \mathcal{V}_b')$, *the set of continuous linear operators, and* $a(\cdot,\cdot) : \mathcal{U}_b \times \mathcal{V}_b \to \mathbb{R}$ *the associated continuous bilinear form, be related by*

$$\text{for fixed } u \in \mathcal{U}_b \ : \ < Au, v >_{\mathcal{V}_b' \times \mathcal{V}_b} = a(u,v) \quad \forall v \in \mathcal{V}_b. \tag{2.9}$$

Then the three statements (2.10), (2.11) and (2.12) are equivalent.

$$A^{-1} \in \mathcal{L}(\mathcal{V}_b', \mathcal{U}_b) \text{ exists,} \tag{2.10}$$

$$\exists \quad \epsilon, \epsilon' > 0 \text{ s.t. } \sup_{0 \ne v \in \mathcal{V}_b} |a(u,v)|/\|v\|_{\mathcal{V}} \ge \epsilon \|u\|_{\mathcal{U}} \quad \forall u \in \mathcal{U}_b \text{ and}$$
$$\sup_{0 \ne u \in \mathcal{U}_b} |a(u,v)|/\|u\|_{\mathcal{U}} \ge \epsilon' \|v\|_{\mathcal{V}} \quad \forall v \in \mathcal{V}_b, \tag{2.11}$$

$$\exists \quad \epsilon > 0 \text{ s.t } \sup_{0 \ne v \in \mathcal{V}_b} |a(u,v)|/\|v\|_{\mathcal{V}} \ge \epsilon \|u\|_{\mathcal{U}} \quad \forall u \in \mathcal{U}_b \text{ and}$$
$$\forall\ v \in \mathcal{V}_b\ \exists\ u \in \mathcal{U}_b : a(u,v) \ne 0. \tag{2.12}$$

If ϵ, ϵ' *in (2.11), (2.12) are chosen as the exact inf–sup– values, then each of the conditions (2.10), (2.11), (2.12) implies* $\|A^{-1}\|_{\mathcal{V}_b' \leftarrow \mathcal{U}_b} = 1/\epsilon = 1/\epsilon'$.

For applications to discrete operators and/or bilinear forms, the above $\epsilon, \epsilon' > 0$ in (2.11), (2.12) have to be independent of h.

Theorem 2.1.7. Brezzi-Babuska condition: *Let* $\mathcal{U}_b^h, \mathcal{V}_b^h$ *be finite dimensional Banach spaces,* $A^h \in \mathcal{L}(\mathcal{U}_b^h, \mathcal{V}_b'^h)$, *and* $a^h(\cdot,\cdot) : \mathcal{U}_b^h \times \mathcal{V}_b^h \to \mathbb{R}$ *the associated continuous bilinear form, defined*

$$\text{for fixed } u^h \in \mathcal{U}_b^h \text{ by } < A^h u^h, v^h >_{\mathcal{V}_b'^h \times \mathcal{V}_b^h} = a^h(u^h, v^h) \quad \forall v^h \in \mathcal{V}_b^h. \tag{2.13}$$

Then the three statements (2.14), (2.15) and (2.16) are equivalent. It is important that for stability arguments the following constants C, ϵ, ϵ' are valid only for sufficiently small $h < h_0$, $h_0 > 0$ *and have to be* independent of h.

$$(A^h)^{-1} \in \mathcal{L}(\mathcal{V}_b'^h, \mathcal{U}_b^h) \quad \text{and } C \quad \text{exist, s.t.} \|(A^h)^{-1}\|_{\mathcal{V}_b'^h \leftarrow \mathcal{U}_b^h} \leq C, \tag{2.14}$$

$$\exists \quad \epsilon, \epsilon' > 0 \ \text{ s.t } \sup_{0 \neq v^h \in \mathcal{V}_b^h} |a^h(u^h, v^h)| / \|v^h\|_{\mathcal{V}}^h \geq \epsilon \|u^h\|_{\mathcal{U}} \quad \forall u^h \in \mathcal{U}_b^h \ \text{ and}$$

$$\sup_{0 \neq u^h \in \mathcal{U}_b^h} |a^h(u^h, v)| / \|u^h\|_{\mathcal{U}}^h \geq \epsilon' \|v^h\|_{\mathcal{V}}^h \quad \forall v^h \in \mathcal{V}_b^h, \tag{2.15}$$

$$\exists \quad \epsilon > 0 \ \text{ s.t. } \sup_{0 \neq v^h \in \mathcal{V}_b^h} |a^h(u^h, v^h)| / \|v^h\|_{\mathcal{V}}^h \geq \epsilon \|u^h\|_{\mathcal{U}}^h \quad \forall u^h \in \mathcal{U}_b^h \ \text{ and}$$

$$\forall \ v^h \in \mathcal{V}_b^h \ \exists \ u^h \in \mathcal{U}_b^h : a^h(u^h, v^h) \neq 0. \tag{2.16}$$

We do not give a proof, but it is straight-forward to show that (2.15) or (2.16) are necessary for (2.14): If a $0 \neq w^h \in \mathcal{R}(A^h)^\perp$ exists, choose $0 \neq v^h \perp w^h$. Then $a^h(w^h, v^h) = 0$ contradicting (2.15) (2.16).

With a subdivision $\mathcal{T}^h$ as introduced in Definition 2.2.1, we often have to modify the *norms* in (2.5), as

$$\|u^h\|_{k,p}^h := \|u^h\|_{W_q^k(\Omega)}^h = \left(\sum_{T \in \mathcal{T}^h} \sum_{|\alpha| \leq k} \|D^\alpha u^h|_T\|_{L_q(T)}^q \right)^{1/q}, \tag{2.17}$$

with the usual $1 \leq q \leq \infty$,

$$\|u^h\|_{W_\infty^k(\Omega)}^h = \operatorname{ess\,sup}\{|D^\alpha u^h(x)| \ : \forall x \ \in \ T \ \forall \ T \ \in \ \mathcal{T}^h\}$$

and corresponding bilinear forms and *semi-norms*

$$|u^h|_{W_q^k(\Omega)}^h = \left(\sum_{T \in \mathcal{T}^h} \sum_{|\alpha| = k} \|D^\alpha u^h|_T\|_{L_q(T)}^q \right)^{1/q}. \tag{2.18}$$

Remark 2.1.8. 1.) To formulate the corresponding condition for the discrete problem, we want to indicate the discrete bilinear forms. In (2.9), based on correct boundary conditions in $\mathcal{U}_b, \mathcal{V}_b$, the $a(\cdot,\cdot) : \mathcal{U}_b \times \mathcal{V}_b \to \mathbb{R}$ had been introduced. In the cases we consider here, these $a(\cdot,\cdot)$ are always realized by integrals over Ω. Similarly as the norms in (2.18) we will have to introduce

$$a^h(\cdot,\cdot) : \mathcal{U}_b^h \times \mathcal{V}_b^h \to \mathbb{R}, \ a^h(u^h, v^h) := \sum_{T \in \mathcal{T}^h} \int_T a|_T(u^h|_T, v^h|_T) d\mu, \tag{2.19}$$

and corresponding $\tilde{a}^h(\cdot,\cdot) : \mathcal{U}_b^h \times \mathcal{V}_b^h \to \mathbb{R}$, obtained, e.g., by approximating these $\int_T$ by quadrature formulas. We will elaborate this term in (2.19) in

Chapter 4.
2.) We want to point out, that the above inf − sup − conditions for $A, a(\cdot,\cdot)$ do not imply the corresponding inf − sup − conditions for $A^h, a^h(\cdot,\cdot)$. The $a^h(\cdot,\cdot)$ has to satisfy these equations *uniformly* w.r.t. h with corresponding $\epsilon^h \geq \epsilon > 0, \epsilon^{'h} \geq \epsilon' > 0$.

2.2 Subdivisions and Finite Elements

The following introduction into the approximation theoretic properties of Finite Element Methods (FEMs) is strongly influenced by the classical [24] and by [38, 17, 18]. We need several Definitions for subdivisions and finite elements:

Definition 2.2.1. Appropriate subdivision: *A set* $\mathcal{T}^h = \{T_1, \ldots, T_M\}$ *of open subsets* $T_i \subset \Omega \subset \mathbb{R}^n$ *is called an* appropriate subdivision *for* Ω *if*

(*i*) $\overline{\Omega} = \bigcup_{i=1}^{M} \overline{T}_i$;
(*ii*) *each* $T \in \mathcal{T}^h$ *has at least* $n+1$ *and at most* $\overline{n} \geq n+1$ *"vertices", e.g. triangles and quadrangles with straight or curved edges;*
(*iii*) $\overline{T}_i \cap \overline{T}_j$ *is either empty or has one vertex or a common edge* (2.20) *on an, in general,* $(n-1)$*-dimensional surface spanned by joint vertices of* T_i *and* T_j, *see Figure* 2.1;
(*iv*) *If the* T_i *are rectangles, the admissibility condition (iii) can be relaxed as follows:* $\overline{T}_i \cap \overline{T}_j$ *may be one half of the larger edge of the larger* T_j, *see Figure*2.13 *below.*

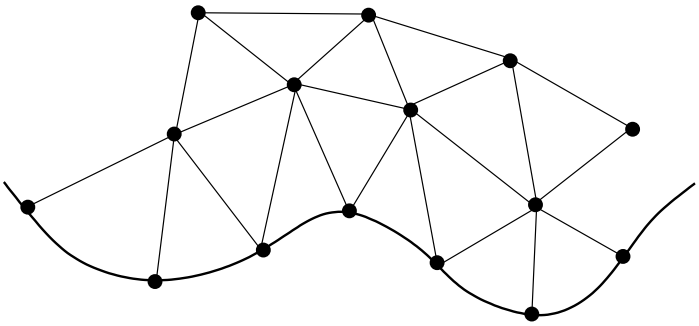

Fig. 2.1. Appropriate triangulation for Ω with curved boundary

We use the term *triangulation* instead of subdivision, whenever we require that the $T \in \mathcal{T}^h$ are triangles, tetrahedrons or n-dimensional simplices. For

$n = 2$ these triangles or rectangles and parallelograms or even more general quadrangles are possible. To simplify the computations we assume each of the $T \in \mathcal{T}^h$ to be affinely (or isoparametrically, see Chapter 2.7), mapped onto the same reference domain, K. We start with strictly locally defined finite elements and introduce, see [24].

Definition 2.2.2. *We assume*

(i) *the* element or reference domain $K \subseteq \mathbb{R}^n$ *to be an open domain with piecewise smooth boundary,*
(ii) *the* space of shape functions $\mathcal{P}$ *to be a finite-dimensional space of functions on* K *and*
(iii) *the span of* nodal variables $\mathcal{N} = \text{span}\{N_1, N_2, \ldots, N_d\}$ *to be a basis* [1] *for* $\mathcal{P}'$,

necessarily of the same dimension $d = \dim \mathcal{P} = \dim \mathcal{P}'$. *Then* $(K, \mathcal{P}, \mathcal{N})$ *is called a* finite element. *There are many possible combinations of* $\mathcal{P}, \mathcal{N}$. *If* $\mathcal{P} = \text{span}\{\phi_1, \phi_2, \ldots, \phi_d\}$ *is a basis for* $\mathcal{P}$ *dual to* $\mathcal{N}$, *hence* $N_i(\phi_j) = \delta_{ij}$, *then* $\phi_1, \ldots, \phi_d$ *is called the* nodal basis *for* $\mathcal{P}$.

It is implicitly assumed that the nodal variables, N_i, lie in the dual space of some larger function space, e.g., a Sobolev space, $\mathcal{U}$. The nodal variables N_i usually denote evaluation of functions or derivatives in given interior and boundary points of K, see the Examples below. A set of nodal variables, $\mathcal{N}$, is called *unisolvent* for $\mathcal{P}$, if

$$\forall (\beta_i)_{i=1}^d \in \mathbb{R}^d \quad \exists_1 \quad \phi \in \mathcal{P} : N_i(\phi) = \beta_i, \quad i = 1, \ldots, d.$$

Then, necessarily $\dim \mathcal{P} = \dim \mathcal{N} = d$, and $\mathcal{N}$ is unisolvent for $\mathcal{P}$ if and only if $\phi \in \mathcal{P}, N_i(\phi) = 0, i = 1, \ldots, d$ implies $\phi \equiv 0$. Figure 2.9 shows a *non unisolvent FE.*

For a given $u \in \mathcal{U}$, a nodal basis of shape functions $\mathcal{P} = \text{span}\{\phi_1, \ldots, \phi_d\}$ and a unisolvent $\mathcal{N} = \text{span}\{N_1, N_2, \ldots, N_d\}$, a (local) interpolation operator is uniquely defined, on a space of functions, $\mathcal{U} : K \to \mathbb{R}$, as

$$I_K := I : \mathcal{U} \to \mathcal{P}, Iu := \sum_{i=1}^{d} N_i(u)\phi_i, N_i(u) = N_i(Iu), i = 1, \ldots, d. \quad (2.21)$$

We use the standard notation of the *Dirac delta function* $\delta(P_i)$ and $\delta(P_i) \in \mathcal{N}$ if $N_i(u) = u(P_i) = \delta(P_i)u$.

[1] a basis of linearly independent elements; this concept will be slightly extended below to *linearly dependent nodal variables*, where $\tilde{\mathcal{N}}$, related to $\mathcal{N}$, satisfies $\dim \tilde{\mathcal{N}} \leq \dim \mathcal{N} = \dim \mathcal{P}$.

2.3 Triangular and Rectangular Polynomial Finite Elements

As most important examples, we start with triangles and quadrangles for $n = 2$ and their generalizations to $n > 2$. To avoid difficulties near the boundary, we start with

$$\text{a polyhedral domain } \Omega, \text{ for } n = 2 \text{ called polygon.} \tag{2.22}$$

The case of curved boundaries for Ω is treated below by interpolation or isoparametric finite elements, see Section 2.7.

We choose the reference domain K in Definition 2.2.2 as a unit rectangular $n-$ triangle or a unit $n-$ cube, hence diam $K = 1$. We give some examples of FE-triples $(K, \mathcal{P}, \mathcal{N})$. The most important choice for $\mathcal{P}$ is, with $\alpha_i \in \mathbb{R}$,

$$\begin{aligned} &\mathcal{P} = \mathcal{P}_m^n := \{u = \sum_{|i| \le m} \alpha_i x^i : x = (x_1, \ldots, x_n) \in \mathbb{R}^n\}, \mathcal{P}_m := \mathcal{P}_m^2 \\ &i = (i_1, \ldots, i_n) \ge 0,\ |i| = i_1 + \cdots + i_n,\ x^i = (x_1, \ldots, x_n)^i = x_1^{i_1} \cdots x_n^{i_n}. \end{aligned} \tag{2.23}$$

Sometimes, we allow even $\mathcal{P} \subset \mathcal{P}_m^n$. We have with $\mathcal{P}_m = \mathcal{P}_m^2$ in (2.23)the following dimensions.

$$\dim \mathcal{P}_m = (m+1)(m+2)/2, \text{ hence } \dim \mathcal{P}_0, \mathcal{P}_1, \mathcal{P}_2, = 1, 3, 6. \tag{2.24}$$

A convenient criterion to check the unisolvence property for $n-$triangles

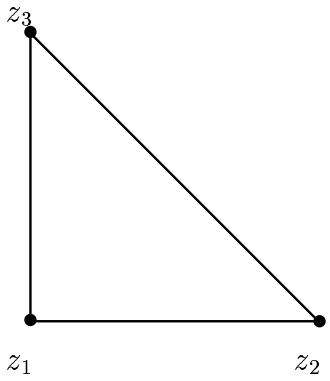

Fig. 2.2. $\mathcal{P}_1$: Linear Lagrange FEs

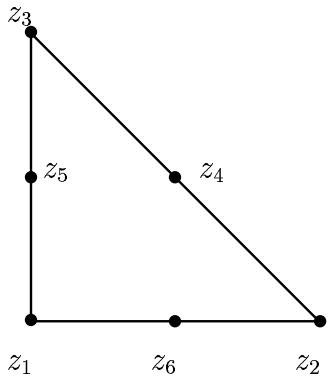

Fig. 2.3. $\mathcal{P}_2$: Quadratic Lagrange FEs

is the following observation, see [18]. Let a *non degenerate hyper plane* L be defined as

$$L := \{x \in \mathbb{R}^n : L(x) = \sum_{|i|=1} \alpha_i x^i - \beta = 0\} \text{ with } \alpha_i \in \mathbb{R}, \sum |\alpha_i| > 0. \tag{2.25}$$

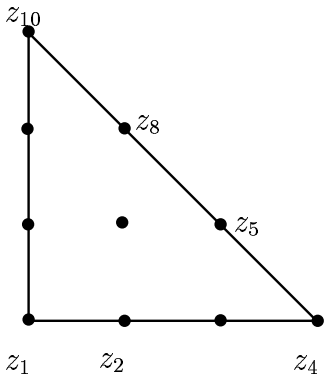

Fig. 2.4. $\mathcal{P}_3$: Cubic Lagrange FEs

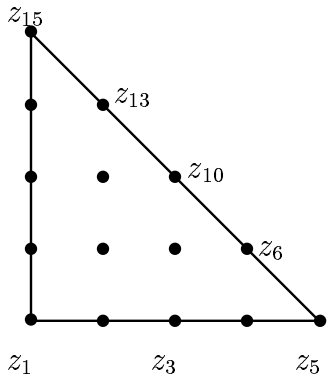

Fig. 2.5. $\mathcal{P}_4$: Quartic Lagrange FEs

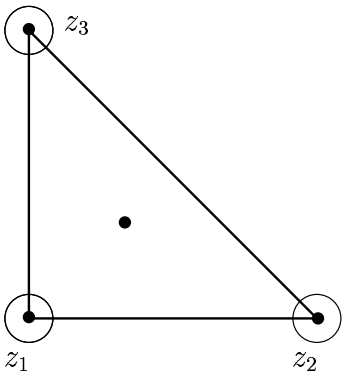

Fig. 2.6. $\mathcal{P}_3$: Cubic Hermite FEs

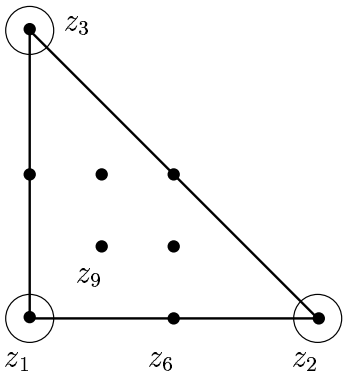

Fig. 2.7. $\mathcal{P}_4$: Quartic Hermite FEs

Proposition 2.3.1. *Let a polynomial $P \in \mathcal{P}_m^n$ vanish on L. Then there exists a $Q \in \mathcal{P}_{m-1}^n$ such that $P(x) = L(x)Q(x)$.*

The edges $L \in \{L_1, L_2, \ldots, L_{n+1}\}$ of the triangle K allow, in combination with Proposition 2.3.1, a factorization of $P \in \mathcal{P}_m^n$. This allows easy proofs for unisolvence, see [18].

Now, we list some combinations of two dimensional triangles K with $\mathcal{P}_m$ and appropriate $\mathcal{N}$, unisolvent for $\mathcal{P}_m$ and show that in the following Figures 2.2- 2.8, 2.10 -2.12. We have included a *non unisolvent combination* for $\mathcal{P}_2$ in Figure 2.9. The different cases are indicated by points z_i in K or on the boundary of K according to

z_i is marked by • if $N_i(\phi) = \phi(z_i)$ for $\phi \in \mathcal{P},\ N_i \in \mathcal{N}$, and by
⊙ if $N_i(\phi) = \phi(z_i)$ and $N_i^g(\phi) = \text{grad } \phi(z_i)$ for
$\phi \in \mathcal{P}_m,\ N_i, N_i^g \in \mathcal{N}$.

In some cases, e.g. for the Argyris element, see Figure (2.17), only normal derivatives are prescribed. If only function values, •, are required, we have a *Langrange finite element*, if function values and first (⊙) or, even additionally, second derivatives, see below, are required we have *Hermite finite elements*. For unisolvence we need the same number of values of the function (and/or its derivative) as $\dim \mathcal{P}_m = (m+1)(m+2)/2$. All Figures show *conforming elements* except the *non conforming elements* in Figure 2.8,2.9. 2.9 si even a *non unisolvent element*.
The case $\mathcal{P}_1$ for the linear conforming and non conforming Lagrange elements shows that the $\mathcal{N}$ is certainly not uniquely determined by $\mathcal{P}$.

Remark 2.3.2. Although it is not explicitely required, all the figures show a high symmetry w.r.t. the boundary (and interior) points. In fact, if we want to allow general nondegenerate subdivisions, see Definition 2.4.5 below, we have to impose symmetric boundary points of the same kind on all edges. This guarantees that the boundary points in neighboring subtriangles coincide. If we discuss highly symmetric subdivisions, e.g., rectangular triangles in a rectangle, we can relax this symmetry of the boundary points.

Klaus Klaus Bleibt das??There are more possible choices of $\mathcal{N}$ for $\mathcal{P}_m$. Obviously, the $\mathcal{P}_m$ Lagrange triangles will do. For $\mathcal{P}_3$ we might choose the midpoint and each vertex of the triangle and two symmetric points on the edges in distance $1/2 \neq a \neq 0$ from each vertex to yield a unisolvent combination. Or we might drop certain monomial terms in $\mathcal{P}_3$, e.g., x^2y, and choose two symmetric points on the edges in distance $1/2 \neq a \neq 0$ from the vertices of each edge and one point on the line from a vertex to the opposite midpoint of an edge, unequal the midpoint of the triangle. Some of these choices impose variational crimes, which have to be paid by the consitency errors below.

For rectangular K we often use instead of $\mathcal{P}_m$ the *tensor products* of *bilinear, biquadratic,.. polynomials*, see Figures 2.10-2.13 and [52, 18]

$$Q_m = \mathcal{P}_m^1 \otimes \mathcal{P}_m^1 = \left\{ \sum_{0 \le j \le m} c_j\ p_j(x) q_j(y) : p_j, q_j \in \mathcal{P}_j^1,\ \le j \right\}, \quad (2.26)$$
$$\dim Q_m = (\dim \mathcal{P}_m^1)^2 = (m+1)^2,$$

Whenever the knots are chosen in a triangle K on $m+1$ parallel lines with altogether $s = 1 + 2 + \cdots (m+1)$ points $z_1, z_2, \ldots, z_s$, there is a *simple inductive computation*, see [17], to determine the interpolating

$$\phi \in \mathcal{P}_m \text{ s.t. } \phi(z_i) = f_i, i = 1, \ldots, s = (m+1)(m+2)/2 = \dim \mathcal{P}_m \quad (2.27)$$

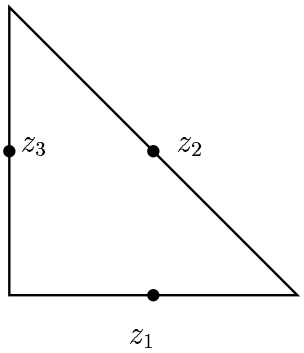

Fig. 2.8. $\mathcal{P}_1$: Non conforming linear Lagrange FEs (Crouzeix-Raviart)

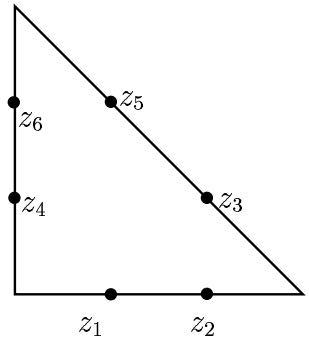

Fig. 2.9. $\mathcal{P}_2$: Non conforming *non unisolevnt* quadratic Lagrange FEs

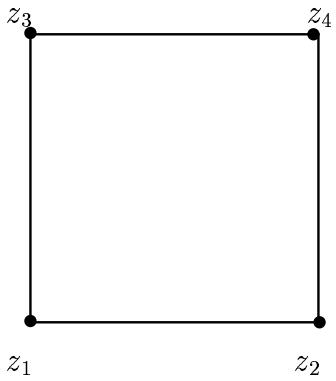

Fig. 2.10. $\mathcal{Q}_1$: Bilinear Lagrange FEs

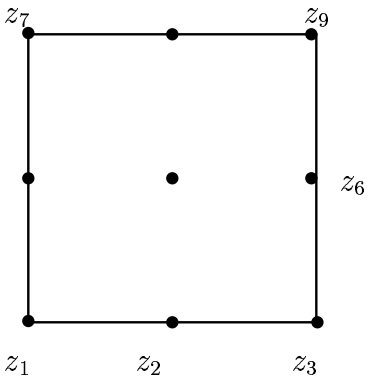

Fig. 2.11. $\mathcal{Q}_2$: Bi quadratic Lagrange FEs

Using the affine equivalence, see Definition 2.4.1, we can choose the parallel lines as $y =$ constant . The ϕ for $m = 1$ line is obvious. So assume we can solve (2.27) for m lines. Let the last line, with the first $m + 1$ knots, $z_1, ..., z_{m+1}$ be located at $y = 0$.

Determine $p_0 \in \mathcal{P}_m^1 : p_0(z_i) = f_i,\ i = 1, \ldots, m+1,$ and, by induction, determine the unique $q \in \mathcal{P}_{m-1}^2 : q(z_i) = \dfrac{1}{y_i}(f_i - p_0(z_i)),\ i = m+2, .., s.$

Then $p \in \mathcal{P}_m^2 = \mathcal{P}_m, p(x,y) := p_0(x) + y\ \ q(x,y)$ solves (2.27) .

A similar approach is possible for some rectangular elements.

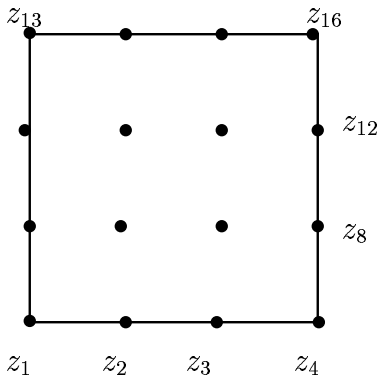

Fig. 2.12. $\mathcal{Q}_3$: Bi cubic Lagrange FEs

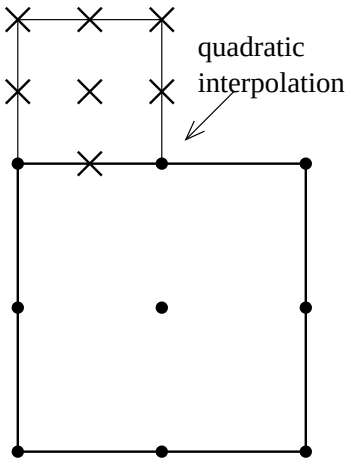

Fig. 2.13. Quadratic subdivision (mesh-refinement), violating Definition 2.2.1(iii), see (iv).

2.4 Finite Element Spaces

Now, we want to handle the general situation. To compute interpolating FEs $\phi \in \mathcal{P}$ we introduce an equivalence.

Definition 2.4.1. *Let $(K, \mathcal{P}, \mathcal{N})$ and $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ be two finite elements, let $F : K \to \hat{K}, F(x) = Ax + b, A \in \mathbb{R}^{n \times n}$ nonsingular, $b \in \mathbb{R}^n$, be an affine (or isoparametric) map such that, see Figure 2.14*

(i) $F(K) = \hat{K}$,
(ii) $F^* \hat{\mathcal{P}} := \hat{\mathcal{P}} \circ F = \mathcal{P}$ *and*
(iii) $\hat{\mathcal{N}}(\hat{f}) = \mathcal{N}(f) = \mathcal{N}(\hat{f} \circ F)$ *for* $f = \hat{f} \circ F$ *or* $\hat{\mathcal{N}} = F_* \mathcal{N}$.

Then $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ is called affine equivalent *to $(K, \mathcal{P}, \mathcal{N})$ and we write $(K, \mathcal{P}, \mathcal{N}) \overset{\simeq}{F} ((\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}}))$.*

Here, we have used the following notation.

Remark 2.4.2. Let $\hat{f} : \hat{K} \to \mathbb{R}$ be given. Then the *pull-back* of $F : K \to \hat{K}$ is defined by $F^* \hat{f} := F^*(\hat{f}) := \hat{f} \circ F$, the *push-forward* F_* as $(F_* N)(\hat{f}) := N(F^* \hat{f}) = N(\hat{f} \circ F)$. We find for $f := \hat{f} \circ F$ that $\hat{f} = f \circ F^{-1}$ and $f = \hat{f} \circ F$ implies $N(\hat{f} \circ F) = N(f) = \hat{N}(\hat{f})$.

Obviously affine equivalence is an equivalence relation. If $\mathcal{P} = \text{span}\{\phi_1, \phi_2, \ldots, \phi_d\}$ is a dual basis for $\mathcal{N}$ with $N_i(\phi_j) = \delta_{ij}$, then $\hat{\mathcal{P}} = \text{span}\{\hat{\phi}_1 = \phi_1 \circ F^{-1}, \ldots, \hat{\phi}_d = \phi_d \circ F^{-1}\}$ is a dual basis for $\hat{\mathcal{N}}$ and $\hat{N}_i \hat{v} = N_i(\hat{v} \circ F)$.
The elements in the Figures 2.15 and 2.16 are inequivalent. Figure 2.17 shows

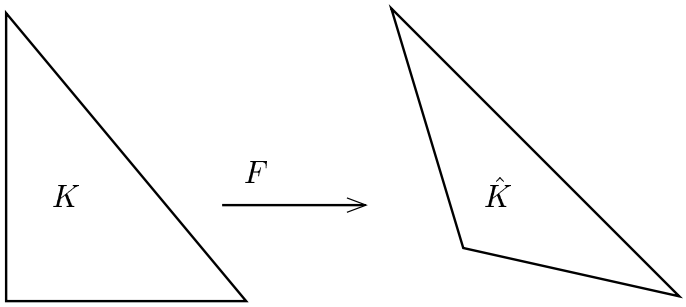

Fig. 2.14. Affine transformation

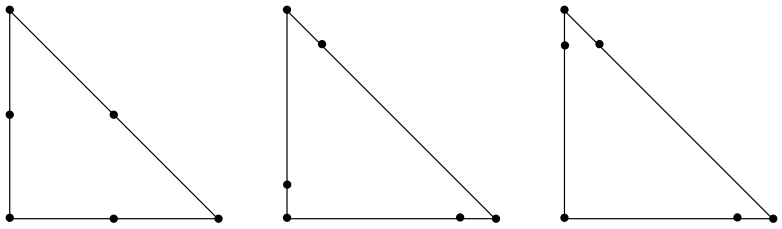

Fig. 2.15. Inequivalent quadratic Lagrange FEs: Affine mappings are not possible

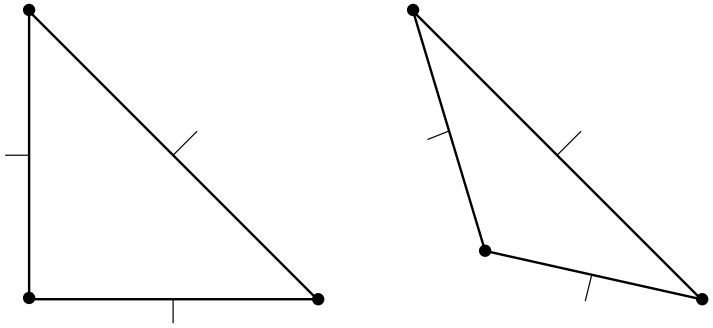

Fig. 2.16. Inequivalent cubic Hermite elements: Incompatible normal derivatives

furthermore, that FEs with prescribed normal derivatives, as e.g. the Argyris elements are not affine equivalent, but only nearly affine equivalent, [24, 17].

Now we can handle the local interpolation in $(K, \mathcal{P}, \mathcal{N})$ and its affine equivalent elements, e.g. the $(T, \mathcal{P}_T, \mathcal{N}_T)$. If (2.22) is satisfied, we can assume in all practical examples the same $(K, \mathcal{P}, \mathcal{N})$, appropriate F_T and $\mathcal{P}_T \circ F_T = \mathcal{P}$. If different reference domains K have to be used, we need ad-

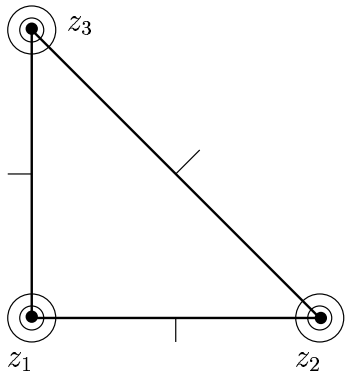

Fig. 2.17. Quintic Argyris FEs

ditional technical tools, e.g., combining triangular and rectangular elements. We introduce the space of Finite Elements. We need $\mathcal{P}, \mathcal{N}$, the subdivision $\mathcal{T}^h$ and the affine (or isoparametric) mappings

$$F = F_T : K \to T,\ Fx = Ax + b,\ A = A_T \in \mathbb{R}^{n\times n},$$
$$b = b_T \in \mathbb{R}^n\ \forall\ T \in \mathcal{T}^h. \tag{2.28}$$

For simplicity we again assume that we have an affine (or isoparametric) map from each $T \in \mathcal{T}^h$ to the same reference element $(K, \mathcal{P}, \mathcal{N})$. A generalization to different K is straight forward.

Definition 2.4.3. *Let $\mathcal{T}^h$ be a subdivision for Ω and let $F_T : K \to T$ be chosen as in (2.28). Then*

$$\mathcal{U}^h := \{u^h \in L^\infty(\Omega) : u^h|_T \in \mathcal{P}_T := \mathcal{P} \circ (F_T)^{-1}\} \tag{2.29}$$

is denoted as (approximating) space of Finite Elements *by $\mathcal{T}^h$ and $\mathcal{P}$, u^h as* Finite Elements, *for short, FEs.*
Since the $T \in \mathcal{T}^h$ are open, (2.29) does not imply transition properties for $u^h|_T$ and $u^h|_{T_1}$ if $\overline{T} \cap \overline{T_1} \neq \emptyset$. For non conforming FEs, $u^h \in \mathcal{U}^h$ and an edge $e \subset \overline{T} \cap \overline{T_1}$ usually $u^h|_{\overline{T}|_e} \neq u^h|_{\overline{T_1}|_e}$. Therefore, in many cases additional properties are imposed such as

$$\mathcal{U}^h\ \cap L^2(\Omega),\ \mathcal{U}^h\ \cap C^0(\Omega),\ \mathcal{U}^h\ \cap H_0^1(\Omega).$$

For interpolation and approximation properties of finite elements the following
local interpolation operator is important. It is essentially defined by $\mathcal{P}$ and $\mathcal{N}$. Let in $\mathcal{P} = \{\phi_1, \ldots, \phi_d\}$ the ϕ_i represent a nodal basis for $\mathcal{N} = \{N_1, \ldots, N_d\}$. Using the bijection F_T, the local interpolation on K, see (2.21)

$$I_K v\ =\ Iv\ := \sum_{i=1}^{1} N_i(v)\phi_i \quad \text{for } v : \overline{K} \to \mathbb{R} \tag{2.30}$$

is combined with Definition 2.4.1. This yields a *global interpolation operator* in the following four steps.

Definition 2.4.4. *For the FE space $\mathcal{U}^h : \Omega \to \mathbb{R}$ with the underlying subdivision $\mathcal{T}^h$, see Definitions 2.2.1 ff, and $(K, \mathcal{P}, \mathcal{N})$:*

(i) Let $(T, \mathcal{P}_T, \mathcal{N}_T)$ and the fixed $(K, \mathcal{P}, \mathcal{N})$ be affine equivalent $\forall T \in \mathcal{T}^h$ and $F_T x = A_T x + b_T$ satisfy $F_T(K) = T$.

(ii) Define the interpolation operator $I^h = I^h_{\mathcal{T}^h} : \mathcal{U} \to \mathcal{U}^h$ with $N_i^T(v)$ $:= N_i(v \circ F_T)(= N_i^T(I^h v \circ F_T))$ and $\phi_i^T = \phi_i \circ F_T^{-1}$, as, see (2.30),

$$I^h v|_T = \sum_{i=1}^{d} N_i(v \circ F_T) \cdot (\phi_i \circ F_T^{-1}) = \sum_{i=1}^{d} N_i^T(v)\phi_i^T \quad \forall \quad T \in \mathcal{T}^h \tag{2.31}$$

(iii) Assume $P \in \overline{T} \cap \overline{T_1}$ to be a vertex or a point on an edge e of $\overline{T}$ with $\delta(P) \in \mathcal{N}_T$. Then $\forall \quad T_j \in \mathcal{T}^h$ with $P \in \overline{T_j}$ the identical values of the function or derivative of v in this point P have to be used $\forall \quad \overline{T_j} \ni P$ to define, e.g., an $N(v \circ F_{T_j}) = \delta(P)v$. This implies continuity of, e.g., the function at P from T to T_j, but not on e and in Ω.

(iv) Let l be the highest derivative required in $\mathcal{N}$. A reference element $(K, \mathcal{P}, \mathcal{N}), K \subset \Omega$ is said to be a C^r element *if r is the largest non-negative integer for which*

$$\mathcal{V}^h = I^h C^l(\overline{\Omega}) \subseteq C^r(\Omega) \cap W_\infty^{r+1}(\Omega).$$

We do not intend to formulate the sharpest results under the weakest possible conditions for the different cases below. Rather we formulate conditions, see (2.34) below, which allow the following implications: The interpolation operator I^h is a bounded operator, a good approximation and the FEs allow inverse estimates. Finally, a convergent anti-crime operator E^h exists, see Section 2.6. This is necessary for the nonconforming finite elements. E.g., discontinuous $u^h \in \mathcal{U}^h$ are not in $\mathcal{U}$ or u^h violate the boundary conditions. We introduce

Definition 2.4.5. *Let $\{\mathcal{T}^h\}, 0 < h \leq 1$, be a family of subdivisions of Ω such that*

$$\max\{ \text{ diam } T : T \in \mathcal{T}^h\} \leq h \quad \text{diam } \Omega,$$

then h is denoted as maximal step size *. The family is said to be* quasi-uniform *if there exists $\chi > 0$ such that*

$$\min\{ \text{ diam } B_T : B_T \subset T \in \mathcal{T}^h\} \geq \chi \, h$$

for all $h \in (0, 1]$; here B_T is the largest ball such that T is star-shaped with respect to B_T. The family is said to be nondegenerate *if there exists $\chi > 0$ such that*

$$\text{diam } B_T \geq \chi \quad \text{diam } T \ \forall \ T \in \mathcal{T}^h, h \in (0, 1]. \tag{2.32}$$

Remark 2.4.6. 1)The sequence of $\{\mathcal{T}^h\}$ is *non degenerate* if and only if the chunkiness parameter is uniformly bounded for all $T \in \mathcal{T}^h$ and for all $h \in (0,1]$. Let all $T \in \mathcal{T}^h$ be triangles. Then $\mathcal{T}^h$ is non degenerate if and only if for every $T \in \mathcal{T}^h$ the interior angles $\alpha \geq \alpha_0 > 0$. This is used as definition for quasi-uniform $\mathcal{T}^h$ e.g., in [38, ?]. A quasi-uniform family is non degenerate, but not conversely.
2) If we start with an arbitrary subdivision in two dimensions and repeatedly subdivide by connecting edge midpoints, we obtain a *non degenerate* family of subdivisions. For a well defined alternative, see [4].
3) For quasi-uniform $\mathcal{T}^h$ each T is *uniformly star-shaped.*

2.5 Interpolation errors and inverse estimates

Since the $u^h \in \mathcal{U}^h$ might not be continuous across common edges $e \subset \overline{T} \cap \overline{T_1}$, and hence $u^h \notin H^1(\Omega)$, [24], the usual Sobolev-norms $\|\cdot\|_{W_p^m(\Omega)}$ might not be defined. However, for all practically important cases the $\|u^h|_T\|_{W_p^m(T)}$ and $\|u|_{\overline{T}}\|_{C^m(\overline{T})}$, and hence $\|u^h\|^h_{W_p^m(\Omega)}$ exist. Correspondingly we have introduced norms and semi norms, $\|u^h\|^h_{W_q^k(\Omega)}$ and $|u^h|^h_{W_q^k(\Omega)}$ in (2.17) and (2.18), resp. They coincide with the original norm $\|u^h\|_{W_q^k(\Omega)}$ in $W_p^m(\Omega)$. Analogous definitions and equalities may have to be used for scalar products and bilinear forms as well. The decision which of the following choices of m, p in the spaces

$$\mathcal{U}^h_{W_p^m(\Omega)} := \{u^h \in \mathcal{U}^h : \|u^h\|^h_{W_p^m(\Omega)} < \infty\} \text{ for } h \to 0 \tag{2.33}$$

is appropriate, depends upon the interplay of ansatz- and test-functions. They are related by the corresponding bilinear forms. Certainly, for each fixed h we want the $\|u^h\|^h_{W_p^m(\Omega)} < \infty$. We assume throughout

Let the reference FE $(K, \mathcal{P}, \mathcal{N})$ and the *subdivision* $\mathcal{T}^h$ satisfy

(i) K is star-shaped w.r.t. some ball,

(ii) choose the smallest $\tau \geq -1$ s.t. $\mathcal{P}_{m-1} \subseteq \mathcal{P} \subseteq \mathcal{P}_{m+\tau}$, $\mathcal{P} \subset W_\infty^m(K)$, mostly $\tau = -1$, $\mathcal{P} \subsetneqq \mathcal{P}_{m+\tau}$ for $\tau \geq 0$, (hence, there exist $v \in \mathcal{P}$ s.t. $\|v\|_{W_\infty^{m+\tau-1}(K)} < \|v\|_{W_\infty^{m+\tau}(K)}$),

(iii) $\mathcal{N} \subset (C^l(\overline{K}))'$, hence the elements of $\mathcal{N}$ are evaluations of u and derivatives of u up to the order $\leq l$ in (different) points P for $P \in \overline{K}$ and additionally let $1 \leq p \leq \infty$ and $\begin{cases} m - l - n/p > 0 & \text{for } p > 1 \text{ or} \\ m - l - n \geq 0 & \text{for } p = 1, \end{cases}$ (2.34)

(iv) alternatively to (iii) we assume $\mathcal{N} \subset (W_p^m(K))'$,

(v) $\{\mathcal{T}^h\}, 0 < h \leq 1$ is a *nondegenerate* family of appropriate subdivisions of a bounded polyhedral domain $\Omega \subset \mathbb{R}^n$ with the parameter χ, see (2.32),

(vi) for all $T \in \mathcal{T}^h, 0 < h \leq 1$, the $(T, \mathcal{P}_T, \mathcal{N}_T)$ is affine equivalent to $(K, \mathcal{P}, \mathcal{N})$,

(vii) the combination of (i) and (v) shows that T is uniformly star-shaped $\forall\, T \in \mathcal{T}^h$ and that the chunkiness parameter γ of all $T \in \mathcal{T}^h$ is bounded.

Then the following results are proved in [18], compare [17, 38] as well. Mind that the conditions, relating m, l, n, p in [17, 38], (4.4.4) and (4.4.20) Theorems, are only imposed for (2.34) (iii) to allow the Sobolev embedding Lemma, but not in (iv).

Theorem 2.5.1. *Let the $\{\mathcal{T}^h\}, 0 < h \leq 1$, and the reference element, $(K, \mathcal{P}, \mathcal{N})$, satisfy (2.34) for some l, m, p. Let I^h be the (local) interpolation operator in (2.31) defined by $(K, \mathcal{P}, \mathcal{N})$. Then there exists a positive constant C, such that for all $0 \leq s \leq m$, the following local and global convergence- and boundedness-results are correct for all $v \in W_p^m(\Omega)$:*

$$\|v - I^h v\|_{W_p^s(T)} \leq C\ (\ \text{diam}\ T)^{m-s} |v|_{W_p^m(T)},$$

$$\|I^h v\|_{W_p^s(T)} \leq \|v\|_{W_p^s(T)} + C(\ \text{diam}\ T)^{m-s} |v|_{W_p^m(T)}. \tag{2.35}$$

$$\|v - I^h v\|^h_{W_p^s(\Omega)} \leq C\ h^{m-s} |v|_{W_p^m(\Omega)}, \tag{2.36}$$

$$\|I^h v\|^h_{W_p^s(\Omega)} \leq \|v\|_{W_p^s(\Omega)} + C\ h^{m-s} |v|_{W_p^m(\Omega)}$$

This C depends on the reference element, and on n, m, p and the number in χ in (2.32). For $0 \leq s \leq l < m$, see (iii) in (2.34), we have

$$\|v - I^h v\|^h_{W_\infty^s(\Omega)} \leq C\ h^{m-s-n/p} |v|_{W_p^m(\Omega)} \quad \forall\ v \in W_p^m(\Omega). \tag{2.37}$$

Inverse estimates relate various norms for $s \geq m$, here for finite element spaces. They are necessary to handle variational crimes. We again present local and global estimates. Note that for the next two results, see [18], we

need $\mathcal{T}^h$ and the affine (isoparametric) maps $F_T : K \to T$. However, no reference to $\mathcal{N}, \mathcal{N}_T$ nor the unisolvence is made.

For the following discussions we often need scaling of T into $\hat{T}$ and of the corresponding Sobolev norms. Let T be a bounded domain in $\mathbb{R}^n$ and v be a function defined on T. Then $\hat{T}$ and $\hat{v}$ are defined as

$$\hat{T} = \{(1/\ \operatorname{diam} T)x : x \in T\} \text{ with } \operatorname{diam}\ \hat{T} = 1 \text{ and}$$
$$\hat{v}(\hat{x}) = v((\ \operatorname{diam} T)\hat{x}) \ \forall\ \hat{x} \in \hat{T}. \tag{2.38}$$

If $\mathcal{P} = \mathcal{P}_T$ is a vector space of functions defined on T, then $\hat{\mathcal{P}} := \hat{\mathcal{P}}_{\hat{T}} = \{\hat{v} : v \in \mathcal{P}\}$ is defined on $\hat{T}$. Often, we have to relate the $\|u^h|_T\|_{W_p^m(T)}$ and $\|u \circ F_T|_{\hat{T}}\|_{W_p^m(\hat{T})}$ or we may choose a general K instead of $\hat{T}$. This *scaling of Sobolev norms* satisfies

Proposition 2.5.2. *Let $f \in W_p^m(T), 1 \le p \le \infty, j := |\alpha| \le m$, and $T, \hat{T} \subset \mathbb{R}^n$ be related, see (2.38), as $T = F_T(\hat{T})$. with $F_T x := Fx := hx + b$. Then, we obtain* [2]

$$\|\partial^\alpha f\|_{L_p(T)} = h^{n/p-|\alpha|}\|\partial^\alpha(f \circ F_T)\|_{L_p(\hat{T})}, \quad \text{with } h = \operatorname{diam} T, \text{ hence} \tag{2.39}$$
$$|f|_{W_p^j(T)} = h^{n/p-j}|f \circ F_T|_{W_p^j(\hat{T})},\ \|f\|_{W_p^m(T)} \le Ch^{n/p-m}\|f \circ F_T\|_{W_p^m(\hat{T})}.$$

Proof For $Fx := F_T x = hx + b$ we find with $F' = hI$, I the identity matrix, that

$$D^{(j)}(f \circ F)(z) = h^j(D^{(j)}f) \circ F(z)$$

hence

$$\partial^\alpha(f \circ F)(z) = h^j(\partial^\alpha f) \circ F(z).$$

This yields

$$\int_K |\partial^\alpha(f \circ F)(z)|^p dz = \int_K |h^{|\alpha|}(\partial^\alpha f) \circ F(z)|^p dz \tag{2.40}$$
$$= h^{p|\alpha|} \int_K |(\partial^\alpha f) \circ F(z)|^p \det(F')|dz\ /\ |\det(F')| \tag{2.41}$$
$$= h^{p|\alpha|-n} \int_T |(\partial^\alpha f)|^p dx. \tag{2.42}$$

This shows the equalities in (2.39), summation yields the last inequality. ■

Note that the condition $\chi h \le \operatorname{diam} T \le h$ see (2.34) (v), in the following Proposition 2.5.3 is strictly local, referring only to T, and thus is satisfied for nondegenerate subdivisions.

[2] the general case of the form $F : \hat{T} \to T, Fx := F_T x = Ax + b =: hA^0 x + b$ can be treated along the lines in [18] essentially using the formula (4.4.23) there. this yields for $T \in \mathcal{T}^h$ with non degenerate $\mathcal{T}^h$ again $\|f\|_{W_p^m(T)} \le Ch^{n/p-m}\|f \circ F\|_{W_p^m(\hat{T})}$

Proposition 2.5.3. *Let* $\chi h \leq \operatorname{diam} T \leq h$, *where* $0 < h \leq 1$ *is any positive number (not necessarily the* h *in* $\mathcal{T}^h$*) and* $\mathcal{P}$ *be a finite- dimensional subspace of* $W_p^j(T) \cap W_q^l(T)$, *where* $1 \leq p \leq \infty, 1 \leq q \leq \infty$ *and* $0 \leq l \leq j$. *Then there exists* $C = C(\hat{\mathcal{P}}, \hat{T}, j, p, q, \chi)$ *such that for all* $v \in \mathcal{P}$, *we have*

$$\|v\|_{W_p^j(T)} \leq C\ h^{l-j+n/p-n/q} \|v\|_{W_q^l(T)}. \tag{2.43}$$

The following theorem is a global version of the last result.

Theorem 2.5.4. *Let* $\{\mathcal{T}^h\}, 0 < h \leq 1$, *and* $(K, \mathcal{P}, \mathcal{N}), \mathcal{P}_T = \mathcal{P}$ *satisfy (2.34),* $\mathcal{P} \subseteq W_p^j(T) \cap W_q^l(T)$ *where* $1 \leq q \leq p \leq \infty, 0 \leq l \leq j \leq m + \tau$. *For* $T \in \mathcal{T}^h$ *let* $\mathcal{V}^h = \{v^h : \Omega \in \mathbb{R}, v^h$ *is measurable and* $v^h|_T \in \mathcal{P}_T \ \ \forall \ \ T \in \mathcal{T}^h\}$. *Then there exists* $C = C(j, p, q, \chi)$ *such that*

$$\|v^h\|_{W_p^j(\Omega)}^h \leq C\ h^{l-j+n/p-n/q} \|v^h\|_{W_q^l(\Omega)}^h \tag{2.44}$$

for all $v^h \in \mathcal{V}^h$. *Mind that (2.44) only makes sense for* $j \leq m + \tau$, *for* τ *see (2.34), since otherwise* $\|v^h\|_{W_p^j(\Omega)}^h = \|v^h\|_{W_p^{m+\tau}(\Omega)}^h$. *For* $p < q$ *the (2.44) remains correct for a quasi-uniform family* $\mathcal{T}^h$ *if the term* $(n/q - n/p)$ *in the exponent is deleted*

Remark 2.5.5. This last Theorem is a slight modification of [18]. Here we only require a non-degenerate $\{\mathcal{T}^h\}$ however for $q \leq p$ instead of $1 \leq p, q \leq \infty$ for a quasi-uniform $\{\mathcal{T}^h\}$ there. A careful control of the [18] proof shows that only for $p < q$ the quasi-uniform $\{\mathcal{T}^h\}$ is needed. Furthermore, for the study of the $C(\hat{\mathcal{P}}, \hat{T}, j, p, q, \chi)$, see (4.5.14) in [18], only norms for the A_T in $F_T x = A_T x + b_T$ and its inverse A_T^{-1} for each single element are needed. There is no reference to (2.34) (i)-(iii).

2.6 Anti-Crime Transformation from $\mathcal{U}^h$ to $\mathcal{U}$

To simplify the technicalities, we *restrict the discussion to two variate FEs* in this and the next Section. The extension to three variate FEs is partially straight forward. For the next Section it is presented in Lenoir, [41]. For the study of variational crimes we require an operator, E^h, which eliminates the variational crimes in the sense that

$$E^h : \mathcal{U}^h \to \mathcal{U} \text{ for } \mathcal{U}^h \not\subset \mathcal{U}, \text{ yields } E^h u^h \in \mathcal{U}. \tag{2.45}$$

This E^h is realized via a local interpolation operator I_e^h, see (2.68). It is defined on an

$$\begin{aligned} &\textit{extended F.E.}\ (K, \mathcal{P}^e, \mathcal{N}^e) \text{ with } \mathcal{P} \subset \mathcal{P}^e,\ \mathcal{N} \subset \mathcal{N}^e \\ &\qquad \text{and a corresponding } \textit{extended subspace}\ \mathcal{U}_e^h. \end{aligned} \tag{2.46}$$

It is not possible to straight forwardly generalize the convergence results with the techniques used for Theorem 2.6.3.

To define E^h, the I_e^h is applied, not to the original u^h, but to an averaged and additionally interpolated u_e^h, indicated as $u_e^h \in \mathcal{U}_e^h \neq \mathcal{U}^h$. However, I_e^h can be applied to the original u^h as well. So, we require that

$$\begin{aligned} &I_e^h : \mathcal{U}_e^h \to \mathcal{U}_e^h \text{ with } I_e^h|_{\mathcal{U}^h} = id_{\mathcal{U}^h}, \text{ hence } I_e^h u^h = u^h \in \mathcal{U}^h \text{ and} \\ &E^h : \mathcal{U}^h \to \mathcal{U}_e^h \subset \mathcal{U}, \; E^h u^h := I_e^h u_e^h \neq I_e^h u^h \text{ with } u_e^h \in \mathcal{U}_e^h \neq \mathcal{U}^h, \end{aligned} \tag{2.47}$$

see Algorithm 2.2. As indicated above, we present the following construction for the case of two variables,

$$n = 2 \text{ and for } \mathcal{U} = W_p^1(\Omega), \; 1 \leq p \leq \infty. \tag{2.48}$$

The above E^h in (2.45) enforces exact boundary conditions, see end of this Section, and deletes, e.g., the discontinuities of an $u^h \in \mathcal{U}^h$ along the edges. In fact, for discontinuities $\mathcal{U}^h$ along an edge $e \subset \overline{T} \cap \overline{T_1}, T \neq T_1$, usually too few interpolation or continuous transition points $P_i \in e$ are required to enforce [3] $u^h|_{T|_e} \equiv u^h|_{T_1|_e}$. We assume for the two vertices of one, and, see (2.49), (2.50), of all edges e : Vertices $Q_1, Q_2 \in e$, missing in $\mathcal{N}_T$, are simultaneously missing for all edges. This excludes some exotic FEs, e.g. subdivisions as in Figure 2.13, or affine equivalent to a hexagon with prescribed function values in every second vertex. However, it is automatically satisfied for subdivisions, affine equivalent to triangles or squares. It implies in general different values in the vertices $Q_i \in e$, hence $u^h|_T(Q_i) \neq u^h|_{T_1}(Q_i), i = 1, 2$. This yields discontinuity and only $u^h_{|_T}$, but not u^h, is defined in Q_1 and Q_2. Altogether these discontinuities of u^h along e imply $u^h \notin \mathcal{U}$, see [24], p 207-208. We study the following most important cases of non conforming elements for $\mathcal{U}^h \not\subset W_p^1(\Omega)$. We construct the operator E^h in (2.68) and study its properties in Theorem 2.6.3.

The basic idea is simple: Instead of $(K, \mathcal{P}, \mathcal{N}_T)$ with [4], say $\mathcal{P}|_e = \mathcal{P}_k^1$ we introduce additional points in $\overline{e} \subset \overline{K}$. Then we extend $\mathcal{P}, \mathcal{N}$ to $\mathcal{P}^e, \mathcal{N}^e$ s.t. $(K, \mathcal{P}^e, \mathcal{N}^e)$ is unisolvent, $\mathcal{P} \subset \mathcal{P}^e$ and it defines a conforming finite element space. To achieve that we might have to delete or add some interior points in T. So we assume[5]

[3] we use here and below the slightly incorrect notation $\mathcal{P}|_{T|_e}$ for $\mathcal{P}|_{\overline{T}|_e}$ or $u^h|_T(Q)$ for $u^h|_{\overline{T}}(Q)$ with $Q \in \overline{T}$.

[4] since k might be different for different edges e, we choose the largest value of k for all edges $e \in \overline{T}$

[5] again excluding only exotic cases or choosing some additional conditions in $\mathcal{N}$. The following derivations $(u^h)^{(j)}(P_i)$ might indicate either all partials of the order j or, as in the case of Argyris FE, only one directional (normal) derivative in P_i.

For $T, T_1 \in \mathcal{T}^h$with an edge $e \subset \overline{T} \cap \overline{T}_1$ let $\mathcal{P}|_{T|e}, \mathcal{P}|_{T_1|e} \in \mathcal{P}_k^1$.
Let the $\mathcal{N}_T, \mathcal{N}_{T_1}$require the same evaluation of functions and partial
derivatives $(u^h)^{(j)}(P_i), j = 0, 1, \cdots, \mu(P_i) - 1,$ in the same points (2.49)
$P_i \in \overline{e}, i = 1, \ldots, \iota(e)$. For the following construction one sometimes only
counts the derivatives$(u^h)^{(j)}(P_i), 0 \le j \le \mu_i \le \mu(P_i),$ in the direction of e.

As indicated and to simplify the technicalities, we exclude some exotic cases, e.g., subdivisions with regular triangles or rectangles and different numbers of conditions on *different edges* (this is indeed possible and admitted in (2.49)!),

we assume the same conditions (2.49) for every edge $e \subset \overline{T}, T \in \mathcal{T}^h$,
for short $e \in \mathcal{T}^h$. Vertices $Q \in e$ are simultaneously missing. (2.50)

If for all $u^h \in \mathcal{U}^h$ the two polynomials $u^h|_{T|e} = u^h|_{T_1|e}$ for one, and by (2.49) - (2.50), for all edges $e \in \mathcal{T}^h$, then we obtain $u^h \in C(\Omega) \cap H^1(\Omega)$, hence conformity. We need the above condition $\mathcal{P} \subset \mathcal{P}_{m^e}$, $\mathcal{N} \subset \mathcal{N}^{m^e}$, see (2.46), to guarantee $I_e^h u^h = u^h \ \forall \ u^h \in \mathcal{U}^h$, see (2.31). To simplify the presentation we assume, see (2.34) with minimal τ,

$$\begin{aligned} &\mathcal{P}_{m-1} \subseteq \mathcal{P} \subseteq \mathcal{P}_{m+\tau} \subseteq \mathcal{P}^e, \text{ now choose the minimal} \\ &\qquad k \ge m + \tau \text{ s.t. } u^h|_{T|e} \in \mathcal{P}_k^1 \ \forall\, u^h \in \mathcal{U}^h, \\ &\qquad \text{with } \mathcal{P}_{m-1} = \mathcal{P}, \ k = m - 1 \text{ for } \tau = -1. \end{aligned} \tag{2.51}$$

Sometimes this k is overestimated in the sense that $u^h|_{T|e}$ has a degree $< k$. Since we only need E^h for theoretical purposes these overestimates in (2.51) do not cost anything and make our proofs easier. For the k in (2.51) the FEs defined by $(K, \mathcal{P}^e, \mathcal{N}^e)$ and satisfying (2.49)- (2.51) have the property [6]

$$\nu(e) := \mu_1 + \ldots \mu_{\iota(e)} \text{ with } \begin{array}{l} k + 1 - \nu(e) = 0 \text{ yields conforming } u^h \\ k + 1 - \nu(e) > 0 \text{ yields non conforming } u^h. \end{array} \tag{2.52}$$

The case $k + 1 - \nu(e) < 0$ would contradict the aimed unisolvability of $(K, \mathcal{P}^e, \mathcal{N}^e)$. Since we consider the non conforming case here, we will define $k + 1 - \nu(e)$ or $k + 2 - \nu(e)$ additional points Q_j on e and their function values, see Algorithm 2.1. The two vertices of $\overline{e}$ are or have to be included anyway. Thus, we finally have $k+1$ or $k+2$ data on $\overline{e}$. We define the extended degree m^e for polynomials on e and use, a possibly updated, m^e on K :

Algorithm 2.1 *Definition of an* appropriate polynomial degree, m^e, *for the extended conforming FEs. This* $m^e \ge m + \tau$ *in (2.34), (2.51) depends upon the transition properties along the edge under the conditions (2.48)- (2.52). We introduce additional points* Q_j on one edge e *and the corresponding new*

[6] mind that every additional directional derivative along e increases k in (2.51) by 1.

values for $u_e^h(Q_j)$. According to (2.50), we choose the same points on every edge e of $\mathcal{T}^h$:

input $K,\ \mathcal{N},\ \mathcal{P}, Q_j$

for *the non conforming case,* $k+1-\nu(e)>0$, *we choose the missing points* Q_j *and define* $u_e^h(Q_j)$ *according to* (2.55), (2.58).

if *the* vertices $Q_1, Q_2 \in \overline{e}$ *are missing (simultaneously!),* (2.53)

include both Q_1, Q_2 : *for* $Q_j = \overline{T}_1 \cap \cdots \cap \overline{T}_{m_j}$, (2.54)

a common vertex of m_j *squares or triangles, let*

$$u_e^h(Q_j) := \frac{1}{m_j}(u^h|_{T_1}(Q_j) + \cdots + u^h|_{T_{m_j}}(Q_j)), \tag{2.55}$$

if $0 \geq k+1-\nu(e)-2$ (≥ -1 *by* (2.52))

define $m^e := k$ *or* $k+1$ *to obtain* $m^e+1-\nu(e)-2=0$, *hence, conforming* m^e

done

else $k+1-\nu(e)-2>0$, *(non conformity), then choose* $k+1-\nu(e)-2$ *additional points* $Q_j \in (int\ e) \cap (\overline{T}_1 \cap \overline{T})$, (2.56) *different from the original* P_i,

define $m^e := k$; **goto (*)**

else *choose* $k+1-\nu(e)$ *different points* $Q_j \in (int\ e) \cap (\overline{T}_1 \cap \overline{T})$, (2.57) *and* Q_j *different from the original* P_i,

define $m^e := k$

(*) **for** Q_j *in (2.56),(2.57) define*

$$u_e^h(Q_j) = \frac{1}{2}(u^h|_T(Q_j) + u^h|_{T_1}(Q_j))\ . \tag{2.58}$$

done

Remark 2.6.1. 1) For non conforming cases we always obtain $m^e > m-1$.
2) The extended degree m^e, the corresponding additional points Q_j and the values $u_e^h(Q_j)$ on one and hence on all edges $e \in \mathcal{T}^h$ are thus well defined.
3) Mind that totally $m^e+1 = k+2$ conditions are only possible,if *exactly* the two vertices Q_1, Q_2 and no other points are added, otherwise $m^e+1 = k+1$.
4) According to (2.50) we define the *"same"* points Q_j on every edge $e \in T$. *However,* mind that for different edges $e_1, e_2 \in \overline{T}$ with $e_1 \cap e_2 = Q_1$, this vertex Q_1 is added only once to the final extended system $\mathcal{N}_T^e$.
5) The same construction as in Algorithms 2.1 and 2.2 can be used for violated (trivial) Dirichlet boundary conditions. To achieve this, the $u^h \in \mathcal{U}_b^h$ is extended as $u^h \equiv 0$ in $\mathbb{R}^2 \setminus \Omega$.

So, we obtain a unique $p \in \mathcal{P}^1_{m^e}|_e$, see (2.50), (2.49), interpolating in the original and new points, P_i and Q_j in Algorithm 2.1:

$$p \in \mathcal{P}^1_{m^e}|_e \text{ with } (p)^{(j)}(P_i), i = 1, \dots, \iota(e), j = 0, \dots, \mu_i, \quad \text{original,}$$
$$\nu(e) = \mu_1 + \cdots \mu_{\iota(e)} \text{ and } p(Q_j), j = 1, \dots, m^e + 1 - \nu(e), \quad \text{new.} \quad (2.59)$$

As next step, we define a unisolvent combination $(K, \mathcal{P}^e, \mathcal{N}^e)$ or $(T, \mathcal{P}^e_T, \mathcal{N}^e_T)$ by the following Algorithm. Since, for $T \in \mathcal{T}^h$ all the $(T, \mathcal{P}_T, \mathcal{N}_T)$ are affine equivalent to $(K, \mathcal{P}, \mathcal{N})$, the following construction is presented for $(K, \mathcal{P}, \mathcal{N})$. The points Q_j in Algorithm 2.1 are understood to be affinely re transformed from T to K. Again we exclude for a simpler presentation some exotic cases by the following condition. This is, due to (2.24), (2.26) and since $m > 0$, certainly satisfied for subdivisions, affinely equivalent to triangles and squares.

$$\begin{aligned} \dim \mathcal{P}_m &= \dim \mathcal{P}_{m-1} + m + 1 \text{ and} \\ &= \dim \mathcal{P}_{m-1} + 2m + 1 \text{ for triangles and squares.} \end{aligned} \quad (2.60)$$

Algorithm 2.2 *For a* non conforming $(K, \mathcal{P}, \mathcal{N})$ *an* extended conforming unisolvent *FE* $(K, \mathcal{P}^e, \mathcal{N}^e)$ *is defined. By Algorithm 2.1 and* $\mathcal{P} \subset \mathcal{P}^e = \mathcal{P}_{m^e}$ *necessarily* $m - 1 < m_e$. *We use the conditions (2.48) -(2.50), Algorithm 2.1, and determine a* [7] *unisolvent combination* $\mathcal{P}^e$ *and* $\mathcal{N}^e$ *s.t.,*

$$\mathcal{P}_{m-1} \subseteq \mathcal{P} \subsetneqq \mathcal{P}_{m^e} = \mathcal{P}^e \;\Rightarrow\; \textit{dim}\, \mathcal{P}_{m-1} \leq \dim \mathcal{P} < \textit{dim}\, \mathcal{P}_{m^e} \textit{ and } \mathcal{N} \subset \mathcal{N}^e.$$

We employ the values $u^h(Q_j)$ *for the new points in Algorithm 2.1, determined in (2.55), (2.58).*

input $\mathcal{N}$, Q_j, $\mathcal{P}$, $\mathcal{P}_{m^e}$, m^e *is the (modified) degree of the boundary polynomial in Algorithm 2.1,*

define $\mathcal{N}^? := \mathcal{N} \cup \{\delta(Q_j) :$ *all new points* Q_j *from Algorith 2.1* $\}$, *see Remark 2.6.1,* (2.61)

if $|\mathcal{N}^?| >$ *dim* $\mathcal{P}_{m^e}$, *increase* $m^e := m^e + \sigma, \sigma > 0$, *minimal, and chose* σ *additional points* Q_j, *in the interior of every edge* e *according to (2.57), s.t., possible by (2.60),* $\mathcal{N}^? := \mathcal{N}^? \cup \{\forall$ *new* $\delta(Q_j)\}$ *satisfies* $|\mathcal{N}^?| \leq$ *dim* $\mathcal{P}_{m^e}$, **goto** (+) (2.62)

else $|\mathcal{N}^?| \leq$ *dim* $\mathcal{P}_{m^e}$, **goto** (+) (2.63)

[7] Now we allow the derivatives included in $\mathcal{N}$, but not in (2.52). It might be, that this causes the following case $|\mathcal{N}^?| > \dim \mathcal{P}_{m^e}$ and $\sigma > 0$ in (2.62.) By (2.49) we have to reduce σ by the number of conditions on e imposed in $\mathcal{N}$ and not counted in (2.59). This is important for (2.65).

(+) **if** $|\mathcal{N}^?| <$ *dim* $\mathcal{P}_{m^e}$, **goto** $(*)$
else $|\mathcal{N}^?| =$ *dim* $\mathcal{P}_{m^e}$, $\mathcal{N}^a := \mathcal{N}^?$, **goto** **(**)**

(*) *choose enough additional new interior points* $S_\iota \in$ *int*(T)
s.t. $\mathcal{N}^a := \mathcal{N}^? \cup \{\forall \delta(S_\iota)\}$ *satisfies* $|\mathcal{N}^a| =$ *dim* $\mathcal{P}_{m^e}$

(**) **if** $\mathcal{N}^a$ *is unisolvent for* $\mathcal{P}_{m^e}$, **define** $\mathcal{N}^e := \mathcal{N}^a$, **goto** **(***)**
else *move the new points* $Q_j \in e$ *in (2.62),* $S_\iota \in K$, *s.t.,*
$\mathcal{N}^e := \mathcal{N} \cup \{\forall$ *new* $\delta(Q_j), \delta(S_\iota)\}$ *is* **unisolvent** *for* $\mathcal{P}_{m^e}$ (2.64)

(***) **define** *the local conforming interpolant for* u_e^h *by*
$\forall$ P_i *with* $\delta(P_i) \in \mathcal{N} \cap \mathcal{N}^e$ *define*
$(u_e^h)^{(j)}(P_i) := (u^h)^{(j)}(P_i), j = 0, 1, \ldots, \mu(P_i) - 1;$ (2.65)
$\forall$ Q_j *from Algorithm 2.1 and (2.62), (2.64) define*
$u_e^h(Q_j)$ *according to* (2.55), (2.58); (2.66)
$\forall$ S_ι *define* $u_e^h(S_\iota) := u^h(S_\iota)$. (2.67)

done

Finally, we are able [8] to define the anti-crime transformation E^h by applying I_e^h to the semi-locally defined u_e^h see Remark 2.6.1, 1), instead of u^h directly. The local FE interpolation operator I_e^h is, defined [9] in (2.31) based upon $(K, \mathcal{P}^e, \mathcal{N}^e)$ instead of $(K, \mathcal{P}, \mathcal{N})$:

$$E^h : \mathcal{U}^h \to \mathcal{U}_e^h \subset \mathcal{U} \text{ for } \mathcal{U}^h \not\subset \mathcal{U} \text{ as } E^h u^h := I_e^h u_e^h \in \mathcal{U} \text{ and}$$
$$I_e^h u_e^h|_T = \sum_{i=1}^{d^e} N_i^e(u_e^h \circ F_T)(\phi_i^e \circ F_T^{-1}) =: \sum_{i=1}^{d^e} N_{T,i}^e(u_e^h)\phi_{T,i}^e, \quad (2.68)$$

with $N_i^e = N_i$ for all the original $i = 1, \ldots, d = |\mathcal{N}|$; here $d^e = |\mathcal{N}^e| \geq d$, the ϕ_i^e represent the new nodal basis for $\mathcal{P}^e$, w.r.t. $\mathcal{N}^e$ and F_T is unchanged compared to $(T, \mathcal{P}_T, \mathcal{N}_T)$. We next apply I_e^h directly to u^h. In this case we replace the $N_{T,i}^e(u_e^h)$ in (2.68) by $N_{T,i}^e(u^h)$, the values of the function and derivatives of the original u^h. So we get, compare (2.65) - (2.67) the

$$(u^h)^{(j)}(P_i) \text{ as in (2.49)} , \quad (2.69)$$
$$u^h(Q_j) := u^h|_{\overline{T}}(Q_j) \text{ for } Q_j \in \overline{e} \subset \overline{T}, \quad (2.70)$$
$$u^h(S_\iota) := u^h(S_\iota) \text{ as in (2.67)} . \quad (2.71)$$

Mind that the $u^h(Q_j) = u^h|_{\overline{T}}(Q_j)$ and $u^h|_{T_1}(Q_j)$ are usually different for $Q_j \in \overline{e} \subset \overline{T} \cap \overline{T}_1$, and $T \neq T_1$. Since $\mathcal{P} \subset \mathcal{P}_{m^e}$ and $\mathcal{N}^e$ (with $\mathcal{N} \subset \mathcal{N}^e$) is unisolvent for $\mathcal{P}_{m^e}$, we obtain

[8] again this condition (2.64) only excludes a few exotic cases
[9] certainly I_e^h can be applied to $\mathcal{U}$ directly as well, if all the $N_i^e(u_e^h \circ F_T) = N_{T,i}^e(u_e^h)$, see (2.68), are defined; again this I_e^h is defined strictly locally

$$I_e^h u^h = u^h \ \ \forall \ \ u^h \in \mathcal{U}^h \not\subset \mathcal{U} \text{ and } E^h u^h - u^h = I_e^h(u_e^h - u^h). \quad (2.72)$$

To prove the following Theorem, we need some results in [18]. For a non degenerate family of subdivision $\mathcal{T}^h$ and $T \in \mathcal{T}^h$, let T be affine equivalent $\forall\ T \in \mathcal{T}^h$ and $F : K \to T,\ Fx := F_T x = \ \text{diam}\ T A_T x + b_T = \ \text{diam}\ T Ax + b$. Let $\hat{T} := \{x/\ \text{diam}\ T : x \in T\}$ with $\hat{F}_T : K \to \hat{T},\ \hat{F}x := \hat{F}_T x = A_T x + \hat{b}_T = Ax + \hat{b}$. Then the local interpolation operators $I_K : C^l(\overline{K}) \to W_p^m(K)$ and $I^h = I_T^h : C^l(\overline{T}) \to W_p^m(T)$ are defined as, see (2.30),

$$\begin{aligned} I_K v \ = \ Iv \ &:= \sum_{i=1}^{d} N_i(v)\phi_i \ \text{ for } v : C^l(\overline{K}) \to W_p^m(K) \text{ and} \\ I^h v|_T = \sum_{i=1}^{d} N_i(v \circ F_T) \cdot (\phi_i \circ F_T^{-1}) &= \sum_{i=1}^{d} N_i^T(v)\phi_i^T, F_T : K \to T, \quad (2.73) \\ F_T x = \ \text{diam}\ T\ Ax + b \ \ \forall\ v \in C^l(\overline{T}) &\to W_p^m(T),\ T \in \mathcal{T}^h. \\ \hat{I}^h \hat{v}|_{\hat{T}} = \sum_{i=1}^{d} N_i(\hat{v} \circ \hat{F}_T) \cdot (\phi_i \circ \hat{F}_T^{-1}) &= \sum_{i=1}^{d} N_i^{\hat{T}}(\hat{v})\phi_i^{\hat{T}},\ \hat{F}_T : K \to \hat{T}, \\ \hat{F}_T x = Ax + \hat{b} \ \ \forall\ \hat{v} \in C^l(\overline{\hat{T}}) &\to W_p^m(\hat{T}),\ T \in \mathcal{T}^h. \end{aligned}$$

In fact, [18] introduce the operator norm

$$\sigma(\hat{T}) := \sup_{0 \neq \hat{v} \in C^l(\overline{\hat{T}})} \|\hat{I}^h \hat{v}|_{\hat{T}}\|_{W_p^m(\hat{T})} / \|\hat{v}\|_{C^l(\overline{\hat{T}})} \quad (2.74)$$

and show, cf (4.4.9)–(4.4.10), (4.4.23):

Proposition 2.6.2. *For a polyhedral domain Ω, a non degenerate family of subdivision $\mathcal{T}^h$, $T \in \mathcal{T}^h$, affine equivalent to a reference element $K, \mathcal{P}, \mathcal{N}$, $\hat{T} := \{x/\ \text{diam}\ T : x \in T\}$ let $0 < h := \max_{T \in \mathcal{T}^h} \text{diam}\ T \leq 1$, χ as in (2.32) and $A := (a_{ij})_{i,j=1}^n$, $\|A\|^\infty := \max\{|a_{ij}|\}_{i,j=1}^n$, $A^{-1} := (a_{ij}^{-1})_{i,j=1}^n$, $\|A^{-1}\|^\infty := \max\{|a_{ij}^{-1}|\}_{i,j=1}^n$. Then there exists a constant $C_{ref} = C(K, \mathcal{P}, \mathcal{N})$ s.t.*

$$\begin{aligned} \sigma(\hat{T}) &\leq C_{ref}\big(1 + \|A\|^\infty\big)^l \big(1 + \|A^{-1}\|^\infty\big)^m |\det A|^{1/p}, \quad (2.75) \\ \sup\{\sigma(\hat{T}) : T \in \mathcal{T}^h, 0 < h \leq 1\} &= C(\chi, n, m, p, K) < \infty. \end{aligned}$$

Now, a simple combination and scaling of (2.75) yields, with $0 < \ \text{diam}\ T \leq 1$ and (2.73),

$$\begin{aligned} \|I^h v|_T\|_{W_p^m(T)} \leq \sigma(T)\|v\|_{C^l(\overline{T})} &\leq C_{ref}(\ \text{diam}\ T)^{-m+n/p}\ \big(1 + \|A\|^\infty\big)^l \\ &\quad \times \big(1 + \|A^{-1}\|^\infty\big)^m |\det A|^{1/p} \|v\|_{C^l(\overline{T})} \\ &\leq C_{ref}^*(\ \text{diam}\ T)^{-m+n/p} T\ \|v\|_{C^l(\overline{T})} \quad (2.76) \end{aligned}$$

In the following estimates we use analog to (2.17) the

$$\|u^h\|^h_{C^l(\Omega)} \text{ and } \|u^h\|^h_{C^l(\overline{T}_N)} := \max\{\|u^h\|_{C^l(\overline{T}_i)} : \overline{T_i} \cap \overline{T} \neq \emptyset\},$$

similarly $\|u^h\|^h_{W_p^m(\overline{T}_N)}$. This indicates, that the definition of u_e^h in E^h requires information from the neighboring T_i. Below we want to apply (2.76) to $I_e^h u_e^h$. Then (2.76) is transformed into

$$\begin{aligned}\|(I_e^h u_e^h)|_T\|_{W_p^m(T)} \leq \sigma(T)\|u^h\|_{C^l(\overline{T}_N)} &\leq C_{ref} \ \text{diam}\ T^{-m+n/p}\ \left(1+\|A\|^\infty\right)^l \\ &\quad \times \left(1+\|A^{-1}\|^\infty\right)^m |\det A|^{1/p}\|u^h\|_{C^l(\overline{T}_N)} \\ &\leq C^*_{ref} \ \text{diam}\ T^{-m+n/p}\ \|u^h\|^h_{C^l(\overline{T}_N)}. \end{aligned} \tag{2.77}$$

In fact, the neighboring $\|u^h\|_{C^\ell(\overline{T}_i)}$ only enter via $(1+(\ \text{diam}\ T_i)\|A\|^\infty)^\ell \leq (1+\|A\|^\infty)^l$. This shows that the $\ \text{diam}\ T_i$ do not matter in (2.77).

In these last estimates the $\|v\|_{C^l(\overline{T}_i)}$ and $\|u^h\|_{C^l(\overline{T}_i)}$ in fact only require upper bounds for the values of functions and derivatives in the few points needed to define the $N_i(v \circ F_T)$ in (2.73) and the $N_i^e(u_e^h \circ F_T)$ in (2.68).

Theorem 2.6.3. *Choose, under the conditions $1 < n,\ 1 < p \leq \infty$, (2.34), (2.48), (2.50), for the original $(K, \mathcal{P}, \mathcal{N})$ a new extended and unisolvent reference element $(K, \mathcal{P}^e, \mathcal{N}^e)$. Define the values for u_e^h as in Algorithms 2.1 and 2.2. Let l be the highest derivative required in the definition of I^h and let the local mapping $(I_e^h)_{|_T} : C^l(\overline{T}_N) \to W_p^1(T)$, be defined by replacing in (2.68) Then the local $E^h|_T : \mathcal{U}^h \to W_p^1(T)$, $(E^h u^h)|_T := I_e^h u_e^h$ and its global (piecewise) extension $E^h : \mathcal{U}^h \to \mathcal{U}_e^h \subset \mathcal{U} = W_p^1(\Omega)$ are bounded. E^h is an approximate identity in the following sense: A constant $C = C(\mathcal{P}, \mathcal{P}^e, \mathcal{N}, \mathcal{N}^e, p, \chi)$ exists, s.t.*

$$\|E^h u^h - u^h\|^h_{W_p^1(T)} \leq C(\ \text{diam}\ T)^{(n-1)/p}\|u^h\|^h_{W_p^1(\overline{T}_N)} \ \forall T \in \mathcal{T}^h \tag{2.78}$$

$$\|E^h u^h - u^h\|^h_{W_p^1(\Omega)} \leq Ch^{(n-1)/p}\|u^h\|^h_{W_p^1(\Omega)}, \tag{2.79}$$

$$\|I^h E^h u^h - u^h\|^h_{W_p^1(\Omega)} \leq Ch^{(n-1)/p}\|u^h\|^h_{W_p^1(\Omega)}, \tag{2.80}$$

$$\|E^h u^h\|^h_{W_\infty^1(\Omega)} \leq C(\|u^h\|^h_{W_p^1(\Omega)} + \|u^h\|^h_{C^l(\Omega)}) \ \forall\ u^h \in \mathcal{U}^h.$$

These results remain valid if the trivial Dirichlet boundary conditions are violated see Remark 2.6.1, 5).

In (2.78) - (2.81) the $h^{(n-1)/p}$ could be replaced by

$$h^{-m+n/p} \cdot h^{1-1/p} = h^{1-m+(n-1)/p} = h^{(1-m)+(n-1)/p}.$$

Since $p \geq 1$ this would yield no convergence for $m = 2$. So we keep the above formulation.

For Chapters 6 ff. we need (2.79) for $W_p^1(\Omega)$. For a polyhedral domain Ω, see (2.22), $E^h : \mathcal{U}^h \to \mathcal{U}_b$, hence the $E^h u^h$ exactly satisfy the trivial Dirichlet boundary condition, for curved boundaries, see Chapter 2.7.

Proof We use the norms in (2.17) and combine the semi-local definition of E^h, I_e^h, I^h based on u_e^h in (2.68) with the facts $\mathcal{P} \subset \mathcal{P}^e, u^h = I_e^h u^h \ \forall u^h \in \mathcal{P}$ and the following obvious equalities. The non-obvious estimate in (2.81), will be proved below.

$$|(u^h - u_e^h)^{(j)}|_T(P_i)| = 0 \quad \forall\, j = 0, \ldots, \mu(P_i) - 1, \ |(u^h - u_e^h)|_T(S_\iota)| = 0$$
$$|(u^h - u_e^h)|_T(Q_j)| \le C(\text{diam T})^{1-1/p} \cdot \|u^h\|^h_{W_p^1(\overline{T}_N)} \ \forall\, P_i,\ S_\iota,\ Q_j \in \overline{T} \quad (2.81)$$

The third estimate (2.80) follows from the second, since I^h is bounded and

$$I^h E^h u^h - u^h = I^h I_e^h u_e^h - u^h = I^h(I_e^h u_e^h - u^h) = I^h(E^h u^h - u^h)$$

For the original and the new points P_i and Q_j, S_ι, resp., (2.81) implies, in particular for the evaluation of new function values in $N_T^e \in \mathcal{N}_T^e$,

$$|N_T^e(u^h) - N_T^e(u_e^h)| \le C\ (\ \text{diam}\ T)^{1-1/p} \|u^h\|^h_{W_p^1(T_N)}.$$

We have applied I_e^h in (2.72) to u^h and estimate $\|I_e^h(u_e^h - u^h)\|^h_{W_p^1(\Omega)}$ or

$$\|I_e^h \hat{u}^h\|^h_{W_p^1(T)} \text{ for } \hat{u}^h \ := \ u_e^h - u^h \text{ for } \hat{u}_i^h := N_{T,i}^e(\hat{u}^h), i = 1, \ldots, d^e.$$

This can be achieved by applying (2.77) to the situation for the new $(K, \mathcal{P}^e, \mathcal{N}^e)$ or its $I_e^h u_e^h$ extension. The $\|u^h\|^h_{C^l(\overline{T}_N)}$ is, hence, by (2.77), estimated by (2.81). So the combination of (2.77) and (2.81) yields, for $m = 1$, the (2.78), (2.79), (2.80).

The missing part of this proof is the inequality in (2.81). Until the end of this proof we use the notation $h := \ \text{diam}\ T$, $x := P_i$ and $y := Q_j \in \overline{e}$ representing old and new points in $\overline{T}$ with $\delta(P_i) \in \mathcal{N}_T, \delta(Q_j) \in \mathcal{N}_T^e$, hence, the $u^h(P_i),\ u^h(Q_j)$ are needed in (2.68). The Taylor formula yields

$$u^h(y) - u^h(x) = \int_0^1 (u^h)'(x + tk)k dt \text{ with fixed } k := y - x \in \mathbb{R}^2 \quad (2.82)$$

and for $u^h \in W_p^2(\Omega)$, see (2.4). With $(u^h)' \in L_p(\gamma), k \in L_q(\gamma), 1 \le p \le \infty, 1/p + 1/q = 1$ and the segment $\gamma := \overline{xy} \subset \overline{e} \subset \overline{T}$ we obtain with the Hoelder inequality

$$|u^h(y) - u^h(x)| \le \int_0^1 |(u^h)'(x + tk)k| dt \le C\|(u^h)'\|_{L_p(\gamma)} \cdot \|k\|_q, \quad (2.83)$$

with the $q-$norm $\|k\|_q \leq h$ in $\mathbb{R}^2$. The trace theorem 2.1.3 yields the following estimate

$$\begin{aligned}\|(u^h)'\|_{L_p(\gamma)} &\leq C\|(u^h)'\|^{1-1/p}_{L_p(T)}\|(u^h)'\|^{1/p}_{W^1_p(T)} \text{ by (2.4) and } u^h \in W^2_p(T)\\ &\leq C\|(u^h)'\|^{1-1/p}_{L_{p(T)}}h^{-1/p}\|(u^h)'\|^{1/p}_{L_p(T)} \text{ by (2.44)}\\ &\leq Ch^{-1/p}\|(u^h)'\|_{L_p(T)} \;\leq\; Ch^{-1/p}\|u^h\|_{W^1_p(T)}.\end{aligned}$$

The combination with (2.83) now yields the estimate

$$|u^h(y) - u^h(x)| \leq Ch^{1-1/p}\|(u^h)'\|_{L_p(T)} \leq Ch^{1-1/p}\|u^h\|^h_{W^1_p(T)}. \quad (2.84)$$

This is applicable to all combinations of $x, y \in \{P_i, Q_j\}$ and via triangle inequality to $|u^h_e(y) - u^h(y)|$ as well for the cases $y = Q_j$ and $y = S_\iota$. Since in (2.55), (2.58) the $u^h_e(Q_j)$ are determined as mean values of different $u^h_{|T_i}(Q_j)$, we have to replace the $\|u^h\|^h_{W^1_p(T)}$ in (2.84) by $\|u^h\|^h_{W^1_p(\overline{T}_N)}$. We still can use the same h since we integrate for the different T_i along the same edge e of length $\leq h$. This yields the necessary (2.81). ■

Corollary 2.6.4. *For a non conforming $(K, \mathcal{P}, \mathcal{N})$ the interpolation projector I^h, see (2.31) and the crime eliminating operator E^h in Theorem 2.6.3 satisfy, for $1 \leq p \leq \infty$, the compatibility property*

$$\|E^h\ I^h\, u - u\|_{W^1_p(\Omega)} \leq Ch^{1-1/p}\|u\|_{W^1_p(\Omega)} \text{ for } u \in W^1_p(\Omega). \quad (2.85)$$

Remark 2.6.5. 1) For natural boundary conditions we do not have to worry about their violation.
2) This shows that, after proving stability in the following Chapters the equivalence for compatible approximations in [43, 44, 45, 46], [68], [53, 55, 54], [56] are valid.

Proof The I^h and I^h_e, defined for $(K, \mathcal{P}, \mathcal{N})$ and $(K, \mathcal{P}_e, \mathcal{N}_e)$ yield convergent approximations, w.r.t. $\|\cdot\|^h_{W^1_q(\Omega)}$. So Theorems 2.5.1 and 2.6.3 yield (2.85). ■

2.7 Curved Boundaries

Until now we have excluded non polygonal boundaries see (2.1). Now, we allow

$$\begin{aligned}\Omega \in \mathbb{R}^2 \text{ and let } \partial\Omega \in C^t_p,\ t \geq 1, \quad &\text{where } {}_p \text{ indicates piecewise}\\ &\text{smooth functions, and let } \mathcal{P} = \mathcal{P}_{m-1} \text{ with } m \geq 1. \quad (2.86)\end{aligned}$$

In a first step we define an approximating Ω^h:

Choose points $P_j \in \partial\Omega$ with distance $\leq h$ for neighbouring points.
Include all "non-smooth" points on $\partial\Omega \in C_p^t$ into these P_j.
Replace $\partial\Omega$ and Ω by $\partial\Omega^h$ and Ω^h: $\partial\Omega^h$ is obtained by
connecting the neighbouring $P_j \in \partial\Omega$ by straight lines, thus (2.87)
defining the new edges $e \subset \partial\Omega^h$. The *polygonal* Ω^h is then the
interiour of the $\partial\Omega^h$.
Choose a *nondegenerate subdivision* $\mathcal{T}^h$ for Ω^h as above.

Then automatically the $T \in \mathcal{T}^h$ are star shaped, for $\mathcal{T}^h$ see Figure 2.18. We have to be prepared that the necessarily modified FEs will satisfy the

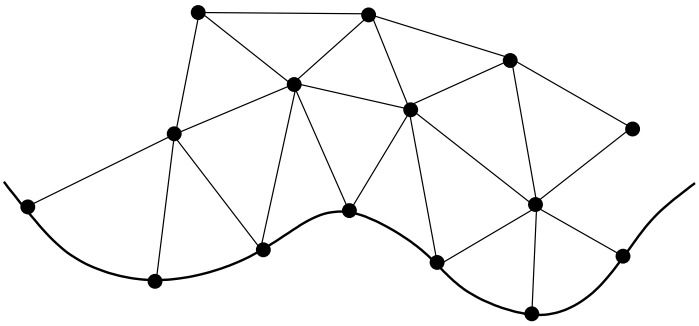

Fig. 2.18. Appropriate triangulation for a curved boundary

boundary conditions only approximately.
The discussion is restricted to (curved) triangulations and to only

$$\begin{aligned} &K \text{ is a triangle, } \mathcal{N} \text{ requires on } \forall\, \overline{e} \subset \overline{K} \text{ only} \\ &m \text{ evaluations of functions and no derivatives on } \overline{e}, \end{aligned} \tag{2.88}$$

this includes both vertices, see (2.92) below. Next, we present two different possibilities to handle this case. For the "boundary FEs" $(T, \mathcal{P}_T, \mathcal{N}_T)$ with $|\overline{T} \cap \partial\Omega| \geq 2$ we change $\mathcal{N}_T$ or $\mathcal{P}_T$: In the first case, see Figure 2.19,

$$\begin{aligned} &\text{replace } \mathcal{N}_T \text{ by interpolation conditions along the curved } \partial\Omega, \\ &\qquad\qquad \text{instead of the straight boundary } \partial\Omega^h. \end{aligned} \tag{2.89}$$

Otherwise, we use F_T defined for T w.r.t. Ω^h, see (2.28), Figure 2.20, however,

$$\begin{aligned} &\text{replace the } \textit{affine } F_T : K \to T \text{ by an } \textit{isoparametric} \\ &\qquad F^h|_T \circ F_T : K \to T_c, \text{ where } F^h|_T \in \mathcal{P}_{m-1}; \end{aligned} \tag{2.90}$$

$\mathcal{P} = \mathcal{P}_{m-1}$ are the original (piecewise) polynomials, here of degree $m-1$ as in the usual FEs and F_c^h, T_c are introduced in (2.93). In this latter case the basic functions $\phi_i \circ (F^h|_T \circ F_T)^{-1} \notin \mathcal{P}$ are more complicated functions.

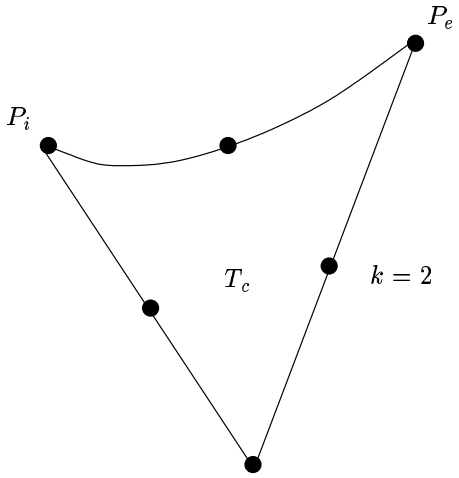

Fig. 2.19. Polynomial interpolation on $\partial\Omega$.

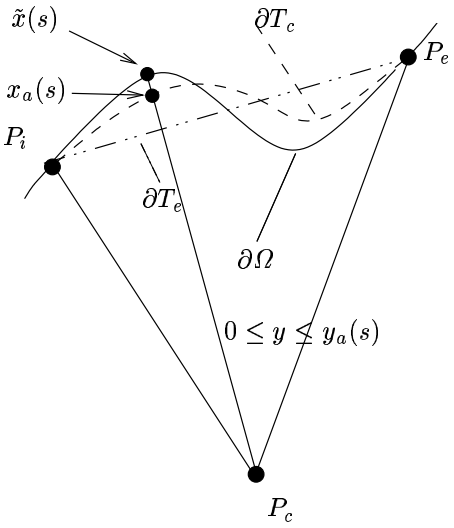

Fig. 2.20. Isoparametric interpolation on $\partial\Omega$.

2.7.1 Polynomial Interpolation in Points of $\partial\Omega$

For the *new triangulation*, $\mathcal{T}^h$, only elements T at the boundary $\partial\Omega$ are changed, see (2.89): For these elements T we admit one or two vertex points in $\overline{T}_v \cap \partial\Omega^h$, with $\overline{T}_v := \{\text{vertices of } T\}$. If $\overline{T}_v \cap \partial\Omega^h = \{P_e\}$, the $(T, \mathcal{P}_T, \mathcal{N}_T)$ is still affine equivalent to $(K, \mathcal{P}, \mathcal{N})$. For two vertex points we modify the Lagrange elements. For this case,

$$\overline{T}_v \cap \partial\Omega^h = \{P_i,\, P_e\}, \quad P_i \neq P_e,\ \overline{T}_v := \{\text{vertices of } T\}, \tag{2.91}$$

we proceed in two steps. Choose a smooth parametrization $x = x(s)$ of $\partial\Omega$ between P_i and P_e w.r.t. the arclength $s, 0 \le s \le h_e$, if possible $s \in C^{2m-1}[0, h_e]$, see (6.26). Next choose m Gauss-Lobatto points $\xi_0, \ldots, \xi_{m-1}, \xi_j := h_e(1+y_j^2)/2,\ \xi_0 = 0, \xi_{m-1} = h_e$ in $[0, h_e]$, see Sections 4.2 and 6.2,in particular (6.14). Determine

$$\begin{aligned} \partial\Omega \ni P_j := x(\xi_j) \approx P_j^s := P_i + \xi_j(P_e - P_i)/\|(P_e - P_i)\|_{\mathbb{R}^2}, \\ j = 0, \ldots, m-1,\ P_i = P_0,\ P_e = P_{m-1}. \end{aligned} \tag{2.92}$$

Replace the m function evaluations in the P_j^s along the straight edge $\overline{P_i P_e}$ by the m function evaluations in the P_j, $j = 0, \ldots, m-1$, thus replacing $\mathcal{N}_T$ by a modified $\mathcal{N}_T^c$ corresponding to one curved edge, see Figure 2.19. For small enough h, the edge $\overline{P_i P_e}$ will be at most $\mathcal{O}(h^2)$ away from the curved part of $\partial\Omega$ between P_i and P_e, see [18], 8 ex.3. Hence $\mathcal{P}_T, \mathcal{N}_T^c$ will be unisolvent simultaneously with $\mathcal{P}_T, \mathcal{N}_T$ for small enough h. These perturbation arguments show that the results in Chapters 2.5, 2.6 remain essentially valid.

2.7.2 Isoparametric Polynomial Approximation

Although the Gauss-Lobatto points decrease, as far as possible, interpolation and quadrature errors along the relevant boundary part of $\partial\Omega$, the accuracy can be improved by a method, established in engineering applications, which allows much more freedom. Particularly efficient is the *isoparametric polynomial approach* in (2.90). As in (2.91) we assume one curved

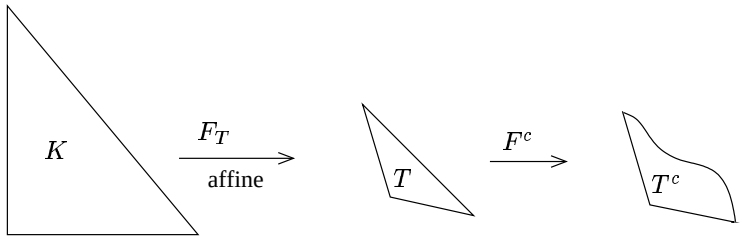

Fig. 2.21. Isoparametric mapping $F = F_T : K \to T, \quad F_c : T \to T_c, \quad F_c \circ F_T : K \to T_c$

and two straight edges P_i, P_e and $P_0\,P_i$, $P_0\,P_e$ for T_c. Well known results of Ciarlet/Raviart,[25], Ciarlet, [24], and Lenoir, [41],allow, for appropriate combinations, Ciarlet/Raviart,[25] the construction of a bijective, piecewise polynomial mapping $F^h : \Omega^h \to \Omega_c^h \approx \Omega \subset \mathbb{R}^2$, such that $F^h = id$ away from the boundary and that [10] the boundary $\partial\Omega$ is well approximated:

[10] for compact B, C, the *distance* of A and B is defined as $\mathrm{dist}\,(B,C) = \max\{\max_{x\in A} : \{\min\{|x-y| : y \in B\}\}, \max_{y\in B} : \{\min\{|x-y| : x \in A\}\}\}$. It represents the maximal distance from any point of one set to the nearest point of the other set.

Definition of the triangulation $\mathcal{T}_c^h$ *and the boundary* $\partial\Omega_c^h := F^h(\partial\Omega^h)$:

Define $F^h : \Omega^h \to \Omega_c^h \approx \Omega$, F^h bijective, and a triangulation $\mathcal{T}^h$ in the following way:

choose Ω^h, $\partial\Omega^h$ and a *triangulation* $\mathcal{T}^h$ for Ω^h as in (2.87), (2.91),

define (simultaneously) $F^h : \Omega^h \to \Omega_c^h := F^h(\Omega^h) \subset \mathbb{R}^2$, $\mathcal{T}_c^h$ and $T_c \in \mathcal{T}_c^h$ such that

(i) let $(F^h)|_T = id|_T \quad \forall T \in \mathcal{T}^h$ with $|\overline{T}_v \cap \partial\Omega| \le 1$ and $\forall \overline{e} \subset \overline{T}$ with $|\overline{e} \cap \partial\Omega| = 1$ we also have $F^h|_{\overline{e}} = id|_{\overline{e}}$,

(ii) let $(F^h)_i|_T \in \mathcal{P}_{m-1} \quad \forall T \in \mathcal{T}^h$ with $|\overline{T}_v \cap \partial\Omega| > 1$, where $(F^h)_i, i = 1, 2$, define the components of F^h,

(iii) $\operatorname{dist}(F^h(\partial\Omega^h), \partial\Omega) = \mathcal{O}(h^m)$, $\operatorname{dist}(F^h(\Omega^h), \Omega) = \mathcal{O}(h^m)$, e.g., realized by the conditions $F^h(P_j^s) = P_j$ for the P_j^s, P_j in(2.92),

(iv) $\|(F^h)'\|_{W_\infty^m(\Omega^h)} \le C$ and $\|((F^h)')^{-1}\|_{W_\infty^m(\Omega_c^h)} \le C$ independent of h,

(v) define the T_c as $T_c = F^h(T) \quad \forall\, T \in \mathcal{T}^h$ and $T = T_c \; \forall\, T \in \mathcal{T}^h$ with $|\overline{T} \cap \partial\Omega| \le 1$ and $T_c \ne T$ with $|\overline{T} \cap \partial\Omega| = 2$,

(vi) define $\mathcal{T}_c^h := \{T_c^h\}$ for the T_c^h in (v), and $\overline{\Omega_c^h} := \bigcup_{T_c \in \mathcal{T}_c^h} \overline{T_c}$. (2.93)

As for non conforming FEs, the F^h is only piecewise continuous with only piecewise defined derivatives and gives rise to violated continuity . This can be treated along the same lines as in Section 2.6, see Theorem 2.6.3.

The (2.93) defines an *isoparametric subdivision*, $\mathcal{T}_c^h$. We proceed by introducing *isoparametric FEs*, $\mathcal{U}^h$, and, later on, an *isoparametric interpolation operator*, I^h. We have to combine the affine mapping $F = F_T : K \to T \in \mathcal{T}^h$ with the polynomials $F^h|_T : T \to T_c$, see Figure 2.21:

Starting with the original $\mathcal{T}^h$ defined for Ω^h in (2.87) and the F^h, $\mathcal{T}_c^h$ and Ω_c^h in (2.93), let

$$\begin{aligned}
\mathcal{U}^h := \{u^h : \Omega_c^h = F^h(\Omega^h) \to \mathbb{R} \;:\; u^h|_{T_c} = \sum_{i=1}^{d} \alpha_i \cdot \phi_i \circ (F^h|_T \circ F_T)^{-1},\\
\alpha_i \in \mathbb{R} \;\; \forall \;\; T_c = F^h(T) \in \mathcal{T}_c^h\}, \text{ and}\\
\|u^h\|_{W_q^k(\Omega_c^h)}^h \text{as in (2.17) with } \Omega,\ \mathcal{T}^h \text{ replaced by} F^h(\Omega^h),\ \mathcal{T}_c^h\\
\mathcal{U}_b^h := \{u^h \in \mathcal{U}^h : u^h(P_i) = 0 \;, i = 0, \ldots, m-1,\\
\text{for all interpolation points along } \partial\Omega_c^h\}. \qquad (2.94)
\end{aligned}$$

The $u^h \in \mathcal{U}_b^h$ violate boundary and continuity conditions.

If $e \subset \overline{K}$ denotes the edge mapped by the original $F_T : K \to T$ onto $\overline{P_i P_e}$ then $\partial\Omega_T^h := F^h(F_T(e))$ only approximates that part of $\partial\Omega$ between P_i and P_e up to $\mathcal{O}(h^m)$. For the following interpolation process this might have the consequence that some of the interpolation points $P_j \in \partial\Omega_T^h$, $j = 1, \ldots, m-2$, are $P_j \notin \Omega \cup \partial\Omega$ and so $u(P_j)$ might not be defined. In this case,

Theorem 2.1.1 yields an appropriate extension u^e of u, to allow the evaluation of $u^e(P_j)$, see [18]. More accurate is an iteratively defined extension, presented in [41]. In fact, we do need this extension for the error estimates in the following Theorem 2.7.1 as well.

For the following analysis we need an auxiliary F_c, which will allow to satisfy the boundary conditions and define the interpolation operator exactly. We use the interpolation operator I^h in (2.31) and choose

$$\begin{aligned} F_c = ((F_c)_i)_{i=1}^2 : \Omega^h \to \Omega \subset \mathbb{R}^2 \text{ such that} \\ F^h = I^h\ F_c := (I^h\ ((F_c)_i)_{i=1}^2) \text{ with} \\ F^h(\Omega^h) = \Omega + \mathcal{O}(h^m) \text{ and, replacing } F^h \text{ by } F_c, \\ F_c \text{ satisfies (2.93) (i)-(iv) and} \partial F_c(\Omega^h) = F_c(\partial\Omega^h) = \partial\Omega; \end{aligned} \tag{2.95}$$

thus F_c has the property that in Figure 2.20 the part of $\partial\Omega$ is the image $F_c(\partial T_{\overline{e}}) \subset$
$\partial\Omega$ of that part $\partial T_{\overline{e}}$ of ∂T_c near $\overline{e}$; again, the nontrivial construction of F_c is presented in [41]. We introduce ϕ^h (and sometimes use its inverse$(\phi^h)^{-1}$)

$$\begin{aligned} &(i) \quad \phi^h : \Omega \to F^h(\Omega^h) \text{ as } \phi^h := F^h \circ F_c^{-1} : t \in \Omega \to x \in F^h(\Omega^h) \approx \Omega, \\ &(ii) \quad (\phi^h)^{-1} : F^h(\Omega^h) \to \Omega \text{ with } t - \phi^h(t) = \mathcal{O}(h^m) \text{ and} \\ &(iii) \quad (\phi^h)' - Id_\Omega = \mathcal{O}(h^{m-1}), ((\phi^h)^{-1})' - Id_{F^h(\Omega^h)} = \mathcal{O}(h^{m-1}). \end{aligned} \tag{2.96}$$

With this ϕ^h and for

$$\begin{aligned} &u^h : F^h(\Omega^h) \to \mathbb{R},\ u^h \in \mathcal{U}^h, \mathcal{V}^h, \text{ we define new } \hat{u}^h \to \mathbb{R} \text{ as} \\ &\hat{u}^h : \Omega \to \mathbb{R},\ \hat{u}^h(t) := (u^h \circ \phi^h)(t), \text{ and } \widehat{\mathcal{U}}_b^h = \{\hat{u}^h : u^h \in \mathcal{U}_b^h\}, \end{aligned} \tag{2.97}$$

analogously, the $\widehat{\mathcal{V}}_b^h$. Note that $\widehat{\mathcal{U}}_b^h \subset \mathcal{U}_b$, that is, the *Dirichlet boundary conditions are satisfied exactly* for $\hat{u} \in \widehat{\mathcal{U}}_b^h$. Vice versa

$$\begin{aligned} \text{for } f : \Omega \to \mathbb{R} \text{ the } \check{f} : F^h(\Omega^h) \to \mathbb{R} \text{ is defined as} \\ \check{f}(x) := f((\phi^h)^{-1}(x)) = (f \circ (\phi^h)^{-1})(x), \end{aligned} \tag{2.98}$$

The F^h in (2.93) can be used to define the *isoparametric interpolation operator:*

$$\begin{aligned} I^h : \mathcal{U} := \{u : \Omega \to \mathbb{R}\} \to \mathcal{U}^h := \{u^h :\ \Omega_c^h \to \mathbb{R}\}, \widehat{I^h u} := (I^h u) \circ \emptyset^h \in \hat{\mathcal{U}}^h \\ I^h u|_T \ = \ I^h u|_T \text{ as in (2.31) } \forall T = T_c \in \mathcal{T}_c^h \text{ with } |\overline{T}_c \cap \partial\Omega^h| \le 1, \quad \text{otherwise} \\ I^h u|_{T_c} \ = \ \sum_{i=1}^d N_i(u \circ F^h|_{T_c} \circ F_{T_c}) \cdot (\phi_i \circ (F^h|_{T_c} \circ F_{T_c})^{-1}) \\ = \ \sum_{i=1}^d N_i^{T_c}(u)\phi_i^{T_c} \quad \forall \ T_c \in \mathcal{T}_c^h; \end{aligned} \tag{2.99}$$

hence we have replaced the affine F_T by their components $(F^h|_{T_c} \circ F_{T_c})_i \in \mathcal{P}_{m-1}, i = 1,2$, for those T_c with two boundary points. Obviously, I^h is again a bounded linear operator as the original I^h, if the norms are now defined w.r.t the $\mathcal{T}_c^h$ above. Furthermore, the original $\phi_i \circ (F_T)^{-1}$ and $N_i(u \circ F_T)$ have to be replaced by the more complicated $\phi_i \circ (F^h|_{T_c} \circ F_{T_c})^{-1}$ and the still linear $N_i(u \circ F^h|_{T_c} \circ F_{T_c})$.

Theorem 2.7.1. *For Ω as in (2.86) define Ω^h as in (2.87). Let, for $0 < h \leq 1$, the triangulation $\mathcal{T}^h$ for Ω^h satisfy (2.34) for some l, m, p. Suppose F^h and $\mathcal{T}_e^h$ are defined as in (2.93), with F^h piecewise of degree $m-1$. Let $(K, \mathcal{P}, \mathcal{N})$ be a C^0 reference element, let $\mathcal{U}^h$,$\mathcal{U}_b^h$, I^h and $\widehat{I^h}u$ be defined as in (2.94), (2.97) and (2.99) resp. Then there exists a positive constant C, depending on the reference element, n, m, p and the number in χ in (2.32) such that for $0 \leq s \leq m$,*

$$\begin{aligned}
\|u - \widehat{I^h}u\|_{W_p^m(\Omega)}^h &\leq C\ h^{m-s}|u|_{W_p^m(\Omega)}, \\
\|\widehat{I^h}u\|_{W_p^s(\Omega)}^h &\leq (1 + C\ h^{m-s})\|u\|_{W_p^m(\Omega)}\ \forall u \in W_p^m(\Omega)\ \textit{and} \\
\|u - \widehat{I^h}u\|_{W_p^s(T)} &\leq C\ (\ \text{diam}\ T)^{m-s}|u|_{W_p^m(T)}, \\
\|\widehat{I^h}u\|_{W_p^s(T)} &\leq (1 + \mathcal{O}((\ \text{diam}\ T)^{m-s}))\|u\|_{W_p^m(T)}. \qquad (2.100)
\end{aligned}$$

For $p = \infty$, we have $\max_{T \in \mathcal{T}^h} \|u - \widehat{I^h}u\|_{W_\infty^s(T)} \leq C\ h^{m-s-n/p}|u|_{W_p^m(\Omega)}\ \forall\ u \in W_p^m(\Omega)$.

Theorem 2.7.2. *Under the conditions of Theorem 2.7.1 (, the $\mathcal{P}$ automatically satisfies $\mathcal{P} \subseteq W_p^j(T) \cap W_q^l(T)$, where $1 \leq q \leq p, 0 \leq l \leq j$) and for $T \in \mathcal{T}_c^h$ let $\mathcal{U}^h = \{u^h : u^h$ is measurable and $u^h \in \mathcal{U}^h$ according to (2.94) $\forall T_c \in \mathcal{T}_c^h\}$. Then there exists a constant $C = C(l, p, q, \chi)$ such that*

$$\|u^h\|_{W_p^j(\Omega_c)}^h \leq Ch^{l-j-(n/q-n/p)}\|u^h\|_{W_q^l(\Omega_c)}^h \qquad (2.101)$$

for all $u^h \in \mathcal{U}^h$ and for all $u^h \in \mathcal{V}^h$. For $p < q$ the (2.101) remains correct for a quasi-uniform family $\{\mathcal{T}^h\}$.

Theorem 2.7.3. *Under the conditions of Theorems 2.6.3 and 2.7.1 there exists again an anti-crime operator, E_b^h, defined in full analogy to (2.68) satisfying (2.79) and the exact boundary conditions.*

Proof Away from the boundary we leave the construction of Theorem 2.6.3 unchanged. By (2.97) the $\hat{u}^h$ satisfy the Dirichlet conditions exactly, so we are done.

At the other side, we can handle the boundary directly as well, see Figures 2.22, 2.23, 2.24, 2.25. We sketch this proof and omit some technical details. Only in specific cases it will be possible to determine a polynomial $p \in \mathcal{P}_{m-1}$

s.t. $p \equiv 0$ on $\partial\Omega_r$, the curved part of the boundary of Ω between the vertices P_i, P_e, see below. So we choose another construction by combining the extended u_e^h, see below, with two arbitrarily smooth functions c_e^h and ξ_e^h. The goal is a crime-free u_a^h with $u_a^h|_{\partial\Omega_r} \equiv 0$. We start with the construction of u_e^h.

We chose a triangle $T \in \mathcal{T}^h$, the triangulation for the polygonal $\Omega^h \approx \Omega$, with $|\overline{T} \cap \partial\Omega| > 0$ and vertices $P_c,\ P_i,\ P_e$. As in Section 2.6 we increase the number of points on $\overline{e}$ defining $\mathcal{N}^e,\ m^e$ such that a continuous transition crossing this edge to the neighboring triangle is guaranteed. This includes the edges in T with $|\overline{T} \cap \partial\Omega| > 0$. As a consequence we obtain continuous $u_e^h : \Omega^h \to \mathbb{R}$, however still violating $u_e^h|_{\partial\Omega} = 0$.

Now, we modify this u_e^h to the curved boundary ∂T_c to obtain a u_c^h. We use $m - 1 := m^e$ and the new $\mathcal{P}_{m^e}$ as $\mathcal{P}_{m-1}$ for the rest of this proof. Accordingly, we choose the $F^h \in \mathcal{P}_{m-1}$ in (2.93) to obtain Figure 2.22. Here,

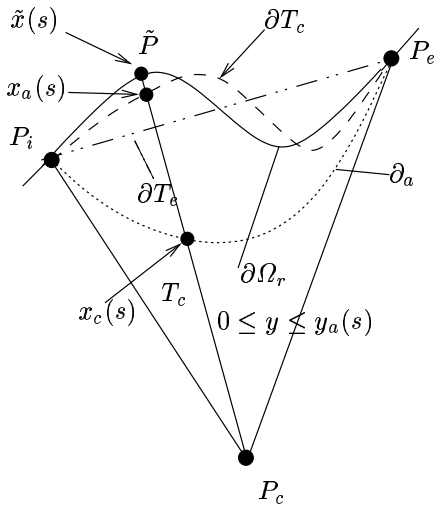

Fig. 2.22. Isoparametry for $\partial\Omega$

the straight $\partial T_e = \overline{P_i\ P_e}$ is mapped by F^h, s.t. $\partial T_c = F^h(\partial T_e)$ satisfies $\text{dist}(\partial T_c,\ \partial\Omega_r) = \mathcal{O}(h^m)$, compare Figure 2.20. We define the curved u_c^h according to Algorithmus 2.1, 2.2 and to the above definition in (2.94) - (2.99). Due to (2.93) *(i)* we keep $u_e^h \equiv u_c^h$ for $|\partial T \cap \partial\Omega| = 1$. For $|\partial T \cap \partial\Omega| = 2$ we give the following construction, see Figure 2.22. With the above $F^h \in \mathcal{P}_{m-1}$, s.t. $\text{dist}(\partial T_c,\ \partial\Omega_r) = \mathcal{O}(h^m)$. let $\partial T_c = F^h(\partial T_e)$ be the approximating curved part of the boundary of a modified T_c. This T_c is obtained by replacing in T the ∂T_e by ∂T_c, hence, $\partial T_c \cap \partial\Omega = \partial T_e \cap \partial\Omega = \{P_i, P_e\}$, see Figure 2.22. Now, we replace in $\mathcal{N}_T^e$ the m original boundary points, P_j on ∂T_e by the $F^h(P_j)$ on ∂T_c to define $\mathcal{N}_{T_c}^e$. Lenoir [41] shows that for small enough h the unisolvence

for T, $\mathcal{N}_T^e$ implies unisolvence for T_c, $\mathcal{N}_{T_c}^e$. We can choose $\psi := u_e^h|_{T_c}, u_e^h \in$

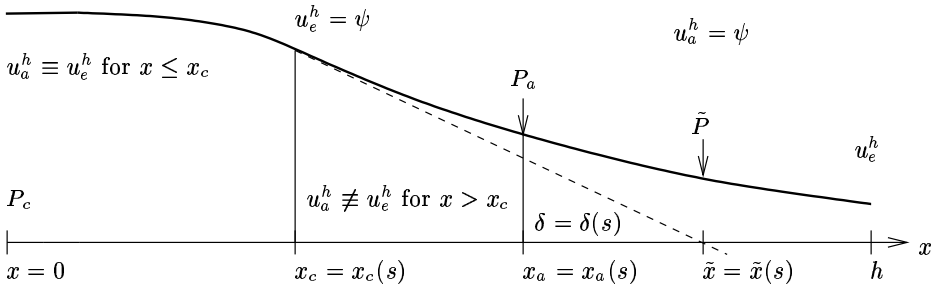

Fig. 2.23. graph for $u_e^h(-)$ and $u_a^h(\dots)$ between P_e and $\tilde{P}$

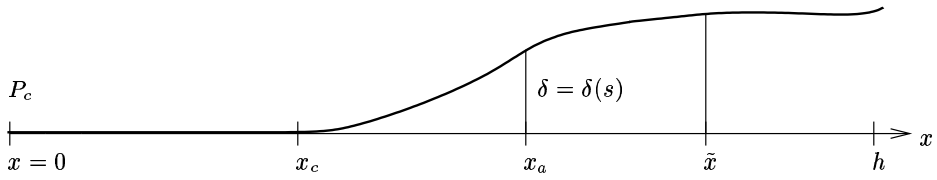

Fig. 2.24. Correction term $c_e = u_c^h - u_e^h$

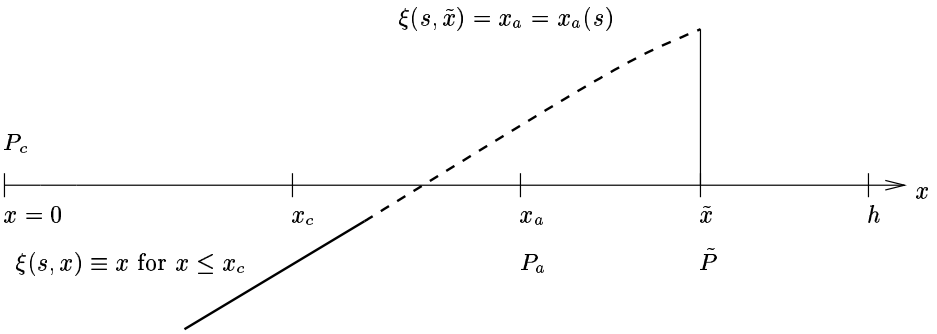

Fig. 2.25. Transformation of x from $\bar{P_e\tilde{P}}$ to $\bar{P_eP_a}$

$\mathcal{U}_b^h$ in (2.94), s.t. $\psi(F^h(P_j)) = 0$ for the m points $F^h(P_j), j = 0, \dots, m-1$ on ∂T_c and $F^h(P_0) = P_0 = P_i, F^h(P_{m-1}) = P_{m-1} = P_e \in \partial T_c \cap \partial\Omega^h$. We assume h small enough s.t. any ray in T_c starting in P_c intersects $\partial\Omega_r$ and

∂T_c in exactly one point. A combination with $\text{dist}(F^h(\partial T_e), \partial\Omega_r) = \mathcal{O}(h^m)$ shows that $|\psi(P)| = \mathcal{O}(h^m)$ for $P \in \partial T_c$.
Next, we define a parameterization for ψ in T_c by introducing the c_e^h, ξ^h below. First, we parametrize ∂T_c by its arclength, s, as $x_a(s), 0 \le s \le h_a$ such that $x_a(s) =$ dist $(P_c, x_a(s))$. Then, we parameterize ψ in T_c as

$$\psi(s,x), 0 \le s \le h_a,\ 0 \le x \le x_a(s) \text{ hence, } \psi(s, x_a(s)) = \mathcal{O}(h^m).$$

Let the ray from P_c to $x_a(s)$ intersect $\partial\Omega_r$ in $\tilde{x}(s)$. We know that

$$x_a(s) - \tilde{x}(s),\ \psi(s,\ x_a(s)),\ \psi(s,\ \tilde{x}(s)) = \mathcal{O}(h^m).$$

Next we define the two smooth functions c_e^h, ξ^h. To this end we introduce a smooth curve $\partial_a \subset T_c$ s.t. ∂_a is not tangential to $\overline{P_i P_c}, \overline{P_e P_c}$ and separates P_c from ∂T_c and $\partial\Omega_r$. We leave ψ unchanged in that part of T_c between P_c and ∂_a. Between ∂_a and ∂T_c we modify it s.t. we obtain analogue local differences as in the proof of Theorem 2.6.3. Assume the ray from P_c to $x_a(s)$ intersects ∂_a in $x_c(s)$. We consider ψ and define the c_e^h, ξ^h along this ray for fixed s. The actual ψ, extended to $\psi : [0,h], h \ge x_a(s), \tilde{x}(s)$, is plotted in Figure 2.23 with a solid line and $\delta = \delta(s) := \psi(s,\ x_a(s))$. We want to change it, indicated by the dotted line, s.t. it vanishes in $\tilde{x}(s)$. Now, define $c_e^h, \xi^h \in C^\infty[0, h_a] \times [0,h]$ as indicated in Figures 2.24 and 2.25.
So let, for fixed $s, 0 \le s \le h_a$,

$$c_e^h(s,x) := \begin{cases} 0 \text{ for } 0 \le x \le x_c := x_c(s) \\ > 0 \text{ for } x_c(s) < x < \tilde{x} := \tilde{x}(s) \\ = \delta := \psi(s,\ x_a(s)) \text{ for } x = \tilde{x}(s). \end{cases}$$

and

$$\xi^h(s,x) := \begin{cases} (s,\ x) \text{ for } 0 \le x \le x_c(s) \\ (s,\ x),\ x > x_c(s) \text{ for } x_c(s) < x < \tilde{x}(s) \\ (s,\ x_a := x_a(s)) \text{ for } x = \tilde{x}(s). \end{cases}$$

For fixed s, the $c_e^h(s,x)$ and $\xi^h(s,x)$ are monotone increasing in $x_c \le x \le \tilde{x}$. Let T_c^Ω be bounded by $\partial\Omega_r$ instead of the above T_c bounded by ∂T_c. Then $\xi^h : T_c^\Omega \to T_c$. Finally

$$u_a^h : T_c^\Omega \to T_c,\quad u_a^h(s,\ x) := (\psi - c_e^h)\xi^h(s,x) \text{ and } u_a^h(x,s) \equiv 0 \text{ on } \partial\Omega_r.$$

Furthermore, we can use the estimates (2.82), (2.84), to prove that this u_a^h satisfies the boundary conditions and the estimates in Theorem 2.6.3. Since $(s, \tilde{x}(s))$ parametrizes $\partial\Omega_r$ we have solved the problem to define a $u_e^h \in \mathcal{U}_b$. ■

3. Conforming Finite Elements

The usual approach in the finite element community considers essentially the weak operator, A, and its corresponding weak bilinear form, $a(\cdot,\cdot)$, see below. Here we generalize this approach such that the generalized form includes the usual finite element approach and the corresponding approach, based on the strong forms of the operator, A_s, and the bilinear form, $a_s(\cdot,\cdot)$. Throughout this Booklet we we will often discuss the weak and the strong form. As we will see in this Chapter, for a smooth enough situation the weak and the strong form of an elliptic problem are analytically equivalent. For *conforming FEMs* this is correct as well, even numerically. Both approaches yield the same (linear) systems with the same coefficients. We will come back to this point in Chapter 6.

There are two good reasons for this unusual approach: Firstly, via the detour to the strong problem, the influence of non conformity can be nicely estimated. This is a generalization of the approach in [18]. (Another possibility are the duality arguments as presented by Rachford/Wheeler, see [39].) Furthermore, we do need this approach anyway for spectral methods which we will present in Subsection 4.2.1.

Secondly, by combining the strong version with quadrature approximations we are able to re-interpret the results as a new class of collocation methods for non-degenerate subdivisions. This seems not to be discussed in the literature until now. We even can define methods of higher order. They are important for path following of parameter dependent solutions of nonlinear problems. The study of turning and bifurcation points requires these methods to avoid spurious solutions.

We present elliptic differential operators of second order, however in this, Chapters 4, 5, and parts of 7 it would require only a simple modification to include operators of order $2m$, however more involved discussions to study the different combinations of natural and prescribed boundary conditions.

3.1 Main Idea and Example for Finite Elements

We start with the simple, but characteristic combination of the Laplace operator and Dirichlet boundary conditions:

Example 3.1 *For* $u \in H^2(\Omega)$ *and* Dirichlet boundary conditions, *hence,* $u \in H_0^1(\Omega)$, $f \in L^2(\Omega)$ *and* $f \in H^{-1}(\Omega)$, *we define two operators, denoted as strong and weak forms* A_s *and* $A_w = A$ *and corresponding bilinear forms* $a_s(\cdot,\cdot)$ *and* $a(\cdot,\cdot)$, *resp., as*

$$\begin{aligned} A_s &: H^2(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega), \\ A = A_w &: H_0^1(\Omega) \to H^{-1}(\Omega) \text{ and} \\ a(\cdot,\cdot) &: H_0^1(\Omega) \times H^1(\Omega) \to \mathbb{R} \text{ by} \\ a(u,v) &:= < Au, v >_{H^{-1}(\Omega)\times H^1(\Omega)} \\ &:= \int_\Omega \nabla u \nabla v + cuvdx \quad \forall v \in H_0^1(\Omega), \\ A_s u &:= -\Delta u + cu, \\ a_s(\cdot,\cdot) &: H^2(\Omega) \cap H_0^1(\Omega) \times L^2(\Omega) \to \mathbb{R}, \\ a_s(u,v) &:= (A_s u, v)_{L^2(\Omega)} = \int_\Omega (-\Delta u + cu)v dx, \text{ resp.} \end{aligned} \tag{3.1}$$

here $\nabla u^h(x) \cdot \nabla v^h(x) = (\nabla u^h(x))^T \nabla v^h(x)$. *The difference is indicated by appropriate indexing, so* A_s *and* $A_w = A$. *We have to distinguish the exact and discrete strong and weak solutions* u_0 *and* $u^h = 0$ *throughout the paper. Usually the context indicates the actual meaning, eg,* $a(u_0, v) = \int_\Omega fv dx \ \ \forall\ v \in H_0^1(\Omega)$ *defines the weak exact solution. Nevertheless we introduce*

Notation 3.1 *We always use the notation* u_0 *and* u_0^h *for the exact and discrete weak or strong solutions, whenever the context indicates the actual meaning. Whenever we have to distinguish the two types, we use*

$$\hat{u}_0,\ \hat{u}^h = 0 \text{ and } \breve{u}_0,\ \breve{u}^h = 0 \text{ for the weak and strong solutions,}$$

resp. [1] *on its head,* $\breve{u}_0$. *The weak solution tries to get rid of everything on its head,* $\hat{u}_0$, *by the sharp* $\hat{}$.

The exact weak and strong solutions $u_0 = \hat{u}_0$, *or* $\overset{u}{0} = \breve{u}_0$ *are then defined, resp., by*

[1] to help the memory, the strong solution can carry something in the basket $\breve{}$

$$
\begin{aligned}
u_0 = \breve{u}_0 \in\ & \mathcal{U}_b := H^2(\Omega) \cap H_0^1(\Omega) : \quad A_s\, u_0 = -\Delta u_0 + c u_0 = f \\
\Leftrightarrow\ & (A_s u_0, v)_{L^2(\Omega)} = \int_\Omega f v dx = (f,v)_{L^2(\Omega)} =: f(v) \qquad (3.2) \\
& = \int_\Omega (-\Delta\, u_0\ + c\ u_0\ )\ v\, dx = a_s(u_0, v) \\
& = a(\breve{u}_0, v) - \int_{\partial\Omega} \frac{\partial u_0}{\partial \nu}\ v ds,\ \frac{\partial \breve{u}_0}{\partial \nu}\ \textit{the outer normal derivative,} \\
& = \int_\Omega f v dx \quad \forall\ v \in L^2(\Omega)\ \textit{and} \\
u_0 \in\ & \mathcal{U}_b := H_0^1(\Omega) : a(u_0, v) = \int_\Omega \nabla\, u_0\, \nabla v + c\, u_0\, v\, dx \\
& =< A u_0, v >_{H^{-1}(\Omega)\times H^1(\Omega)} = f(v) \quad \forall\ v \in H_0^1(\Omega). \qquad (3.3)
\end{aligned}
$$

We obtaion the same solution $u_0 = \hat{u}_0 = \breve{u}_0$, *hence*

$$
a_s(u_0, v)\ = a(u_0, v)\ \textit{since}\ \int_{\partial\Omega} \frac{\partial u_0}{\partial \nu}\ v ds = 0\ \forall\ v \in H_0^1(\Omega). \qquad (3.4)
$$

With the outer normal $\nu = (\nu_1, \ldots, \nu_n)^T$ *for* $\partial\Omega$ *the Green formula yields*

$$
\begin{array}{l}
-\int_\Omega v \partial_j w dx = +\int_\Omega (\partial_j v) w dx - \int_{\partial\Omega} v w \nu_j ds, \quad \textit{hence} \\
\int_\Omega v \partial_i (\partial_i u) dx = -\int_\Omega (\partial_i v) \partial_i u + \int_{\partial\Omega} v \nu_i \partial_i u ds.
\end{array} \qquad (3.5)
$$

This translation from the weak to the strong form requires partial integration and $u_0 \in H^2(\Omega)$. *This is due to the Greens Formula. For this smooth situation and* $v \in \mathcal{U}_b := H_0^1(\Omega)$ *the (3.2), (3.3) are analytically equivalent. Furthermore, we define the restrictions*

$$
\begin{aligned}
a|(\cdot,\cdot) := a(\cdot,\cdot)|_{(H^2(\Omega)\cap H_0^1(\Omega))^2}\ & \textit{and}\ A| := A_{H^2(\Omega)\cap H_0^1(\Omega)},\ \textit{satisfying} \\
< A|u, v >_{H^{-1}(\Omega)\times H^1(\Omega)} & = a|(u,v) = a_s(u,v) = (A_s u, v)_{L^2(\Omega)} \\
& \forall\ \ u \in H^2(\Omega), v \in H_0^1(\Omega). \qquad (3.6)
\end{aligned}
$$

Since $H_0^1(\Omega)$ is dense in $L^2(\Omega)$, we may and do restrict, for Dirichlet conditions, the test functions to $v \in H_0^1(\Omega)$ and to vanishing boundary terms. For these spaces $\subset H_0^1(\Omega)$, the linear operators A_s and A and the bilinear forms $a_s(\cdot,\cdot) : H^2(\Omega) \times L^2(\Omega) \to \mathbb{R}$ and $a(\cdot,\cdot) : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ uniquely define each other as in (3.2), (3.3). They even coincide for $u \in H^2(\Omega)$ and $v \in H_0^1(\Omega)$.

The *basic idea for Finite Elements* (FEs) is the following: We replace the $u_0 \in \mathcal{U}_b,\ v \in \mathcal{V}_b$ in (3.2), (3.3) by elements in finite dimensional spaces $\mathcal{U}_b^h,\ \mathcal{V}_b^h$. They may or may not satisfy the above boundary conditions. For *conforming FEs* they do and we assume $\mathcal{U}_b^h \subset \mathcal{U}_b,\ \mathcal{V}_b^h \subset \mathcal{V}_b,\ \mathcal{U}_b^h, \mathcal{V}_b^h \subset H^1(\Omega)$. Then $a(u^h, v^h) : \mathcal{U}_b^h \times \mathcal{V}_b^h \to \mathbb{R}$ is well defined. In this case, we determine the weak solution $u_0^h \in \mathcal{U}_b^h$ from

$$u_0^h \in \mathcal{U}_b^h : a(u_0^h, v^h) =< Au_0^h, v^h >_{H^{-1}(\Omega)\times H^1(\Omega)} \tag{3.7}$$
$$= \int_\Omega \nabla u_0^h \nabla v^h + c u_0^h v^h \, dx = f(v^h) \quad \forall\, v^h \in \mathcal{V}_b^h \ \subset H_0^1(\Omega).$$

However, compared to (3.2)-(3.3) we have to realize an essential difference: The conforming $\mathcal{U}_b^h, \mathcal{V}_b^h$ which we discuss here are subspaces of $C(\Omega)$, but not of $H^2(\Omega)$. So we are no longer allowed to transform (3.7) directly into an analoge of (3.2), since $a_s(\cdot,\cdot) : \mathcal{U}_b^h \times \mathcal{V}_b^h$ is not defined at all. So we have to define an extension a_s^h and $A_{s,h}$ of a_s and A_s to $\mathcal{U}_b^h \times \mathcal{V}_b^h$ and $\mathcal{U}_b^h \to \mathcal{V}_b^{h'}$ in the form

$$u_0^h \in \mathcal{U}_b^h : (A_{s,h} u_0^h, v^h)_{L^2(\Omega)} := a_s^h(u_0^h, v^h) \tag{3.8}$$
$$:= \sum_{T\in\mathcal{T}^h} \int_T (-\Delta u_0^h \ + c\ u_0^h\ )\ v^h \, dx \ = \int_\Omega f v^h dx = f(v^h) \quad \forall\, v^h \in \mathcal{V}_b^h.$$

To find the relation between the two equations (3.7) and (3.8), we apply partial integration, hence (3.5), for every $T \in \mathcal{T}^h$. Then we obtain for general $u^h \in \mathcal{U}_b^h$

$$a(u^h, v^h) = \sum_{T\in\mathcal{T}^h} \Big(\int_T (-\Delta u^h \ + c\ u^h\ )\ v^h \, dx + \sum_{e\in T} \int_e \frac{\partial u^h}{\partial \nu_e} v^h ds \Big)$$
$$= a_s^h(u^h, v^h) + \sum_{T\in\mathcal{T}^h} \sum_{e\in T} \int_e \frac{\partial u^h}{\partial \nu_e} v^h \tag{3.9}$$

where ν_e is the unit normal vector to the edge e. Now we consider the transition from a triangle T_l to its neighbour T_r , $T_l \in \mathcal{T}^h$. We denote the restriction of v^h, u^h to T_l, T_r as $v_l^h = v^h|_{T_l}$, $u_r^h = u^h|_{T_r}$ a.s.o. Then the above sum is

$$a(u^h, v^h) = a_s^h(u^h, v^h) + \sum_{T\in\mathcal{T}^h} \sum_{e\in T} \int_e \frac{\partial u^h}{\partial \nu_e} v^h ds \tag{3.10}$$
$$= a_s^h(u^h, v^h) + \sum_{e\in T} \int_e \left(v_l^h \frac{\partial u_l^h}{\partial \nu_e} - v_r^h \frac{\partial u_r^h}{\partial \nu_e} \right) ds$$
$$= a_s^h(u^h, v^h) + \sum_{e\in T} \int_e v_l^h \left(\frac{\partial u_l^h}{\partial \nu_e} - \frac{\partial u_r^h}{\partial \nu_e} \right) + (v_l^h - v_r^h) \frac{\partial u_r^h}{\partial \nu_e} ds$$
$$= a_s^h(u^h, v^h) + \sum_{e\in T} \int_e v_l^h \left[\frac{\partial u^h}{\partial \nu_e} \right] ds.$$

Here $[v^h] := v^h|_{T_l}|_e - v^h|_{T_r}|_e$ and $[\partial u^h/\partial \nu_e] := \partial u^h/\partial \nu_e|_{T_l}|_e - \partial u^h/\partial \nu_e|_{T_r}|_e$ denote the corresponding jumps of v^h and $\partial u^h/\partial \nu_e$ across e, resp. Conforming FEs are continuous, hence, $[v^h] := v^h|_{T_l}|_e - v^h|_{T_r}|_e = 0$ and we have omitted this term above. Since we only will estimate the absolute value of the last

messy sum, see Subsection 4.1 and Subsection 6, we do not have to specify the direction of $\partial u^h/\partial \nu_e$ more precisely. We see immediately, that in general $a_s^h(u^h, v^h) \neq a(u^h, v^h)$.

Mind that we have to modify the $\int_e$ terms slightly for $e \subset \delta\Omega$. The Diriclet conditions imply that $v^h, u^h \equiv 0$ in $\mathbb{R}^2 \setminus \Omega$.

Remark 3.1.1. So already at this early stage, we see : A close relation between the $a(\cdot,\cdot)$ and $a_s^h(\cdot,\cdot) : \mathcal{U}_b^h \times \mathcal{V}_b^h \to \mathbb{R}$ is only possible if the FEs are chosen s.t. along the common edges the $[v^h]$ and the $[\partial u^h/\partial \nu_e]$ disappear sufficently often on e, not nesessarily in the same points $\forall v^h \in \mathcal{V}_b^h$ and $\forall u^h \in \mathcal{U}_b^h$.

There is no numerical experience for separate conditions for $\mathcal{U}_b^h$ and $\mathcal{V}_b^h$. However if $[v^h] = 0$, $[\partial v^h/\partial \nu_e] = 0$ $\forall u^h \in \mathcal{U}_b^h, \mathcal{V}_b^h$ in sufficiently many points the very efficient Doedel collocation methods for model problems are available. In some sense they define "super crime" FEs, see below.

So only under the wague conditions of Remark 3.1.1 we can expect similar behaviour of the strong and weak discrete problems. Nevertheless we continue to list both cases.

By introducing bases in $\mathcal{U}_b^h, \mathcal{V}_b^h$, the (3.7), (3.8) yield, even for exact boundary conditions for the $v^h \in \mathcal{V}_b^h$, *different systems of linear equations*, see (3.2).

We may re-interpret the equations (3.7), (3.8) by defining and applying the projectors $Q^{'h}$, $Q_s^{'h}$. We obtain for (3.7)

$$\begin{aligned} &Q^{'h} \in \mathcal{L}(H^{-1}(\Omega), H^1(\Omega) \cap \mathcal{V}_b^{h\prime}) \text{ with} \\ &\quad < Q^{'h} f - f, v^h >_{H^{-1}(\Omega)\times H^1(\Omega)} = 0 \ \forall\ v^h \in \mathcal{V}_b^h \text{ and} \\ &\quad < Q^{'h}(Au_0^h - f), v^h >_{H^{-1}(\Omega), H^1(\Omega)} = 0 \ \forall\ v^h \in \mathcal{V}_b^h. \end{aligned} \tag{3.11}$$

Similarly, (3.8) requires a strong projector $Q_s^{'h}$.

$$\begin{aligned} &Q_s^{'h} \in \mathcal{L}(L_2(\Omega), L_2(\Omega) \cap \mathcal{V}_b^{h\prime}) \text{ with} \\ &\quad (Q_s^{'h} f - f, v^h)_{L_2(\Omega)} = 0 \ \forall\ v^h \in \mathcal{V}_b^h \text{ and} \\ &\quad (Q_s^{'h}(A_{s,h} u_0^h - f), v^h)_{L_2(\Omega)} = 0 \ \forall\ v^h \in \mathcal{V}_b^h. \end{aligned} \tag{3.12}$$

The last two lines in (3.11), (3.12) show that u_0^h in (3.7), (3.8) solves

$$Q^{'h}(A_h u_0^h - f) = 0 \quad \text{and} \tag{3.13}$$

$$Q_s^{'h}(A_{s,h} u_0^h - f) = 0. \tag{3.14}$$

Remark 3.1.2. It has to be pointed out that $Q^{'h}$ and $Q_s^{'h}$ are defined on different spaces, $H^{-1}(\Omega)$ and $L^2(\Omega)$, resp. So an $f \in H^{-1}(\Omega)$ requires information about $v \in H^1(\Omega)$ including ∇v. E.g. we have for $f := Au$ with fixed $u \in H^1(\Omega)$ that

$$< f, v >_{H^{-1}(\Omega)\times H^1(\Omega)} = \int_\Omega \nabla u \nabla v + cuvdx =< Au, v >_{H^{-1}(\Omega)\times H^1(\Omega)};$$

but for $f \in L^2(\Omega)$ no ∇v needs to be considered in $(f,v)_{L^2(\Omega)} = \int_\Omega fvdx$. In $Q_s^{'h}$ we do not refer to ∇v, but have, e.g.,

$$(A_s u, v)_{L^2(\Omega)} = a_s(u,v) = \int_\Omega (-\Delta u + cu)vdx.$$

This difference is important for the approximate projectors $\tilde{Q}^{'h}$ and $\tilde{Q}_s^{'h}$, obtained by quadrature approximations below. The $\tilde{Q}_s^{'h}$ allows the formulation of e.g. *Doedel methods*. Under the wague conditions in Remark (3.1.1) ther are chances for small errors between the weak and the collocation form.

For the case of conforming weak FE-methods we present their convergence theory in this Chapter 3, but for general second order elliptic operators which we are going to introduce now.

3.2 Elliptic Operators and Bilinear Forms

Before we study the generalization of Section 3.1 the general forms of elliptic operators have to be introduced. This Section is in essence a list of definitions for elliptic operators and the related concepts.

To avoid too many technicalities, we consider only [2] *second order elliptic differential operators* in *strong and weak* form. They are related to the corresponding *bilinear forms* and *boundary operators*.

Assume $\Omega \subset \mathbb{R}^n$ to be a bounded domain, $\partial\Omega$ Lipschitz continuous.

Let the operators A_s , A be defined as

$A_s : H^2(\Omega) \to L_2(\Omega)$, $A : H^1(\Omega) \to H^{-1}(\Omega)$, for $u, v : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ and

$$A_s u := -\sum_{i,j=1}^n \partial_j(a_{ij}\partial_i\, u) - \sum_{j=1}^n \partial_j\, (a_{0j}u) + \sum_{i=1}^n a_{i0}\partial_i\, u + a_{00}u \qquad (3.15)$$

for the strong version, and

$$< Au, v >_{H^{-1}(\Omega)\times H^1(\Omega)} := \int_\Omega \Big(\sum_{i,j=1}^n a_{ij}\partial_i u \partial_j v + \sum_j (a_{0j}u)\partial_j v + \sum_i v a_{i0}\partial_i u + a_{00}uv \Big) dx \qquad (3.16)$$

for the weak version.

[2] Using the standard multi-index notation $i = (i_1, \ldots, i_n) \geq 0$ with $|i| = \max\{i_1, \ldots, i_n\}$ and $\partial^i = \partial_{i_1} \ldots \partial_{i_n}$ the general elliptic operator of order m has the form $Au := -\sum_{|i|,|j|\leq m} \partial^j(a_{ij}\partial^i\, u)$ satisfying $\sum_{|i|,|j|=m} a_{ij}\xi^i\xi^j \geq \epsilon'|\xi|_n^{2m}$.

Furthermore we introduce

$$A_m u := -\sum_{i,j=1}^{n} \partial_j(a_{ij}\partial_i\, u) \text{ with } \sum_{i,j=1}^{n} a_{ij}\xi_i\xi_j \geq \epsilon'|\xi|_n^2\ , \epsilon' > 0, \quad (3.17)$$

$|\xi|_n$ the Euclidean norm in $\mathbb{R}^n$ and the $A_m u$, is denoted as the *main part.*

A generalisation to $A : W_p^1(\Omega) \to W_p^{-1}(\Omega), 1 \leq p \leq \infty$ and $A_s : W_p^2(\Omega) \to L_p(\Omega)$ and test spaces $\mathcal{V}_b = W_q^1(\Omega)$ and $L_q(\Omega),\ 1/p + 1/q = 1$ is possible as well, but will not be presented here.

To study the relation between the weak and strong bilinear forms, we start with partial integration. With the outer normal $\nu = (\nu_1, \ldots, \nu_n)^T$ for $\partial\Omega$,

$$\begin{array}{l} -\int_\Omega v\partial_j w dx = +\int_\Omega(\partial_j v)w dx - \int_{\partial\Omega} vw\nu_j ds \quad \text{implies} \\ \int_\Omega v\partial_j(a_{ij}\partial_i u)dx = -\int_\Omega(\partial_j v)a_{ij}\partial_i u + \int_{\partial\Omega} v a_{ij}\partial_j u\nu_j ds. \end{array} \quad (3.18)$$

Or we use the Gauss-Integral Theorem directly to obtain as a consequence of (3.15), (3.20) and for $v \in H^1(\Omega)$

$$(A_s u, v)_{L^2(\Omega)} = \int_\Omega v A_s u =< Au, v >_{H^{-1}(\Omega)\times H^1(\Omega)} - \int_{\partial\Omega} v B_a u ds.$$

The corresponding bilinear and linear forms $a(\cdot,\cdot), a_s(\cdot,\cdot)$ are then defined and related to A by

$$\begin{aligned} a_s(\cdot,\cdot) &: H^2(\Omega)\times L^2(\Omega) \to \mathbb{R}, a_s(u,v) := \int_\Omega A_s uv dx = (A_s u, v)_{L^2(\Omega)}, \\ a(\cdot,\cdot) &: H^1(\Omega)\times H^1(\Omega) \to \mathbb{R} \\ a(u,v) &:= \int_\Omega \Big(\sum_{i,j=1}^{n} a_{ij}\partial_i\ u\ \partial_j\ v + \sum_{j=1}^{n} a_{0j}\ u\ \partial_j\ v \\ &\quad + \sum_{i=1}^{n} a_{i0}(\partial_i\ u)\ v + a_{00} u\ v\Big) dx \text{ and} \\ a(u,v) &= \int_\Omega (A_s\ u)v dx + \int_{\partial\Omega} (B_a\ u)v ds \text{ or} \qquad (3.19) \\ &= a_s(u,v) + \int_{\partial\Omega} (B_a u)v ds \text{ and} \\ B_a\ u &:= \sum_{i,j=1}^{n} \nu_j a_{ij}\partial_i\ u + \sum_{j=1}^{n} \nu_j a_{0j}\ u. \qquad (3.20) \end{aligned}$$

The corresponding $a_m(\cdot,\cdot)$, corresponding to A_m, see (3.17), is denoted [3] as *main part of* $a(\cdot,\cdot)$, see (3.17). For (3.17), we obtain automatically

[3] the m in a_m only indicates the main part of $a(\cdot,\cdot)$, independent of the degree $m-1$ of the FEs.

$$a_m(u,v) := (\sum_{i,j=1}^{n} a_{ij}\partial_i\, u\, \partial_j\, v)dx \text{ satisfies for } 0<\alpha_1<\alpha_2,\ 0<\beta_1<\beta_2$$
$$\alpha_1|u|^2_{H^1(\Omega)} \le \|u\|^2_{a_m} := a_m(u,u) \le \alpha_2|u|^2_{H^1(\Omega)} \text{ and} \tag{3.21}$$
$$\beta_1\|u\|^2_{H^1(\Omega)} \le \|u\|^2_{a_m} \le \beta_2\|u\|^2_{H^1(\Omega)}\ \forall u \in H^1_0(\Omega).$$

We call B_a the "induced" *natural boundary operator*. $Bu = u = 0$ and $B_a u = 0$ on $\partial\Omega$ are denoted as *Dirichlet* and *natural boundary conditions*. For the special case of Example 3.1 we obtain:

$$A_s u = -\Delta u + a_{00}u \text{ induces } B_a\, u = \partial u/\partial\nu.$$

We use the notation B_a since this boundary operator is induced by $a(\cdot,\cdot)$. The relation of B_a to e.g., Dirichlet boundary conditions is visible in (3.20) and is discussed in the textbooks on PDEs, e.g., [38, 18, 17]. Indeed, this $B_a(u)$ is well defined by the trace operator $(B_a\, u)|_{\partial\Omega}$ for $u \in H^2(\Omega)$. By the standard trick $u := u - \overline{u}$ with $\overline{u} = \varphi$ or $B_a\overline{u} = \varphi$ on $\partial\Omega$ we obtain homogeneous boundary conditions exclusively studied in this Booklet. An extension to parts of $\partial\Omega$ with different boundary conditions is presented in most textbooks, e.g., [18], and can therefore be omitted here.

Dirichlet and *natural boundary conditions*, are realized as

$$\mathcal{U}_b = \mathcal{V}_b = H^1_0(\Omega) \text{ and } \mathcal{U}_b = \{u \in H^2(\Omega) :\ B_a u|_{\partial\Omega} = 0\},\ \text{resp.}$$
$$\text{and are combined with } \mathcal{V}_b = \mathcal{V} = H^1(\Omega). \tag{3.22}$$

Sometimes we denote these boundary conditions as Bu and $B_1 v$ for $\mathcal{U}_b$ and $\mathcal{V}_b$, resp. For this general case again (3.6) is valid.

Obviously, we have introduced above *continuous* or *bounded bilinear forms* and *continuous* or *bounded linear operators*. That is:

$$\begin{aligned}
&\text{Positive constants }, C, \text{ exist, s.t.}\\
&|a(u,v)| \ < \ C\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)}\ \forall u,v \in H^1(\Omega) \text{ and}\\
&|a_s(u,v)| \ < \ C\|u\|_{H^2(\Omega)}\|v\|_{L^2(\Omega)}\ \ \forall u \in H^2(\Omega), v \in L^2(\Omega) \text{ and}\\
&\|Au\|_{H^{-1}(\Omega)} \ < \ C\|u\|_{H^1(\Omega)}\ \ \forall u \in H^1(\Omega) \text{ and}\\
&\|A_s u\|_{L^2(\Omega)} \ < \ C\|u\|_{L^2(\Omega)}\ \ \forall u \in H^2(\Omega).
\end{aligned} \tag{3.23}$$

The following concept is only defined for weak bilinear forms or operators. It plays a central role. A *coercive bilinear form* has the following property:

$$\begin{aligned}
&\text{a positive constant }, C, \text{ exists, s.t.}\\
&a(u,u) \ > \ C\|u\|^2_{H^1(\Omega)} \forall u \in H^1(\Omega).
\end{aligned} \tag{3.24}$$

A coercive bounded bilinear form introduces a norm, $\|\cdot\|_a$, which is *equivalent to the norm* $\|\cdot\|_{H^1(\Omega)}$, see (3.21).For the special case $Au = -\Delta u$, or $a_{ij} = \delta_{ij}$ the corresponding $\|u\|_{a_m}$ is denoted as *energy norm*. The estimate of the semi norm or energy norm $|u|_{H^1(\Omega)}$ for $u \in H^1_0(\Omega)$, see (**??**) as well, is based on

Lemma 3.2.1. *For star shaped $\Omega \subset \mathbb{R}^n$ there exists a constant C such that*

$$C\|v\|_{H^1(\Omega)} \leq |v|_{H^1(\Omega)} + |\int_{\partial\Omega} v ds| \quad \forall \quad v \in H^1(\Omega) \tag{3.25}$$

The proof in [18], (8.2.20)Lemma, is valid for our more general conditions as well.

Remark 3.2.2. It is important to realize the impact of the boundary conditions: In fact, the following two equations in (3.26) yield, for $f \in L^2(\Omega)$, the same solution, u_0, if and only if one of the following conditions is satisfied, see (3.19): Either u_0 satisfies the natural boundary condition $B_a(u_0)|_{\partial\Omega} = 0$. Then the boundary term in (3.19) drops out $\forall v \in H^1(\Omega)$. Or u_0 satisfies the Dirichlet boundary condition $(u_0)|_{\partial\Omega} = 0$. Then the boundary term in (3.19) drops out only $\forall v \in H_0^1(\Omega)$.

For $H_0^1(\Omega)$ and $H_0^1(\Omega) \cap H^2(\Omega)$ we determine the *weak* and *strong solution* u_0 by the condition

$$\begin{aligned} &u_0 \in H_0^1(\Omega) : a(u_0, v) = < f, v >_{H^{-1}(\Omega)\times H^1(\Omega)} \quad \forall v \in H^1(\Omega), \text{and} \\ &u_0 \in H_0^1(\Omega) \cap H^2(\Omega) : a_s(u_0, v) = (f, v)_{L^2(\Omega)} \; \forall v \in L^2(\Omega), \\ &\text{hence } Au_0 = f \in H^{-1}(\Omega) \text{ and } A_s u_0 = f \in L^2(\Omega), \text{ resp.} \end{aligned} \tag{3.26}$$

Depending on the boundary conditions, the equality, see (3.19),

$$a(u_0, v) = \int_\Omega (A_s u_0) v dx + \int_{\partial\Omega} (B_a u_0) v ds = \int_\Omega f v dx \; \forall v \in H^1(\Omega) \tag{3.27}$$

can be used in one or two steps: Firstly we insert $v \in H_0^1(\Omega)$ and delete the boundary term. This yields $Au_0 - f = 0 \in L^2(\Omega)$ or $Au_0 - f = 0 \in H^{-1}(\Omega)$. For Dirichlet conditions with u_0, $v \in H_0^1(\Omega) = \mathcal{U}_b = \mathcal{V}_b$ we are done. Natural boundary conditions require a second step: With $\mathcal{V}_b = H^1(\Omega)$, we obtain for arbitrary $v \in H^1(\Omega)$, or $v|_{\partial\Omega} \in H^{1/2}(\partial\Omega)$ that $\int_{\partial\Omega} B_a(u_0) v ds = 0$. This implies, $B_a u_0 = 0$ on $\partial\Omega$. This fact is the reason for the denotation of *natural boundary conditions*. We collect this procedure as

$$\begin{aligned} &(3.27)\; \forall v \in H_0^1(\Omega) \text{ implies } Au_0 - f = 0 \in H^{-1}(\Omega); \text{ this and} \\ &(3.27)\; \forall v \in H^1(\Omega) \text{ implies } u_0 \in H_0^1(\Omega) \text{ or } B_a u_0|_{\partial\Omega} = 0. \end{aligned} \tag{3.28}$$

If, more generally than in (3.19),

$$\hat{f}(v) := \int_\Omega f v dx + \int_{\partial\Omega} \varphi v ds$$

then the above discussion implies $(B_a(u) - \varphi)|_{\partial\Omega} = 0$.

3.3 Convergence for Conforming Finite Element Methods

In this Section we prove the convergence of conforming or crime free weak FEMs for the case of second order coercive elliptic operators. The non coercive case is included as special case in the results of Chapter 7. For the general operators in (3.20) we determine the FE solution, generalizing (3.7).

$$u_0^h \in \mathcal{U}_b^h : f(v^h) = a(u_0^h, v^h) =< Au_0^h, v^h >_{H^{-1}(\Omega)\times H^1(\Omega)} \tag{3.29}$$
$$\int_\Omega \Big(\sum_{i,j=1}^n a_{ij}\partial_i u \partial_j v + \sum_j (a_{0j}u)\partial_j v + \sum_i v a_{i0}\partial_i u + a_{00}uv \Big) dx \ \forall\ v^h \in \mathcal{V}_b^h.$$

For Dirichlet boundary conditions we have then $\mathcal{V}_b^h \subset H_0^1(\Omega)$. Again (3.30) can be interpreted in the form (3.11). The P^h, Q'^h, are either interpolation or truncation or projection operators, see (2.31), (3.11). Here we assume that no variational crimes spoil the situation. Then we have the basic result

Lemma 3.3.1. Cea Lemma: *Let $\mathcal{U}_b^h \subset \mathcal{U}_b$, $\mathcal{V}_b^h \subset \mathcal{V}_b$ and assume $A^h := (Q'^h A|_{\mathcal{U}_b^h})^{-1} \in \mathcal{L}(\mathcal{V}_b^{h'}, \mathcal{U}_b^h)$ and let $u_0 \in \mathcal{U}_b$, a solution of (3.26), exist. Then $u_0^h \in \mathcal{U}_b^h$, the unique solution of (3.30), exists and we obtain the following error estimate:*

$$||u_0 - u_0^h||_{\mathcal{U}} \le \Big(1 + C||A||_{\mathcal{V}'\leftarrow\mathcal{U}} ||(Q'^h A|_{\mathcal{U}_b^h})^{-1}||_{\mathcal{U}_b^h \leftarrow \mathcal{V}_b^{h'}}\Big) \times \|I - P^h\|_{\mathcal{U}\leftarrow\mathcal{U}} \|u_0\|_{\mathcal{U}}. \tag{3.30}$$

Remark 3.3.2. In Theorem 7.2.3 we will show that under rather general conditions the existence and boundedness of $(A^h)^{-1} = (Q'^h A|_{\mathcal{U}_b^h})^{-1} \in \mathcal{L}(\mathcal{V}_b^h, \mathcal{U}_b^h)$ implies the existence of $A^{-1} \in \mathcal{L}(\mathcal{V}_b, \mathcal{U}_b)$, and hence the existence of a unique solution $u_0 \in \mathcal{U}_b$, as well. We use the notations $A^h = Q'^h A|_{\mathcal{U}_b^h}$ and $A_s^h = Q_s'^h A_s|_{\mathcal{U}_b^h}$.

Proof We estimate

$$\begin{aligned} \|u_0 - u_0^h\|_{\mathcal{U}} &\le \|u_0 - P^h u_0\|_{\mathcal{U}} + \|P^h u_0 - u_0^h\|_{\mathcal{U}} \\ &\le \|I - P^h\|_{\mathcal{U}\leftarrow\mathcal{U}}\ \|u_0\|_{\mathcal{U}} + \|P^h u_0 - u_0^h\|_{\mathcal{U}}. \end{aligned}$$

To use $(Q'^h A|_{\mathcal{U}_b^h})^{-1} = (A^h)^{-1}$ we compare

$$\begin{aligned} A^h(P^h u_0 - u_0^h) &= A^h P^h u_0 - A^h u_0^h = Q'^h A P^h u_0 - Q'^h f \\ &= Q'^h A(P^h u_0 - u_0) = A^h(P^h u_0 - u_0) \end{aligned}$$

implying

$$\|P^h u_0 - u_0^h\|_{\mathcal{U}} \le \|(Q'^h A|_{\mathcal{U}_b^h})^{-1}\|_{\mathcal{U}_b^h \leftarrow \mathcal{V}_b^{h'}} \cdot \|Q'^h A\|_{\mathcal{V}'\leftarrow\mathcal{U}} \cdot \|I - P^h\|_{\mathcal{U}\leftarrow\mathcal{U}}\ \cdot \|u_0\|_{\mathcal{U}}$$

and finally

$$\|u_0 - u_0^h\|_{\mathcal{U}} \leq (1 + C\|(Q'^h A|_{\mathcal{U}_b^h})^{-1}\|_{\mathcal{U}_b^h \leftarrow \mathcal{V}_b^{h'}} \|A\|_{\mathcal{V}' \leftarrow \mathcal{U}}) \cdot \|I - P^h\|_{\mathcal{U} \leftarrow \mathcal{U}} \cdot \|u_0\|_{\mathcal{U}}. \blacksquare$$

The discrete coercivity or $\inf - \sup -$ results for the cases of Dirichlet or natural boundary conditions and $a(\cdot,\cdot)$, are presented in the following Theorem. This applies to violated boundary conditions and continuity as discussed in Chapter 4 as well. For the general case of non coercive $a(\cdot,\cdot)$ the stability proof is delayed to Chapter 7.

Theorem 3.3.3. *Let* A, $a(\cdot,\cdot)$, $a^h(\cdot,\cdot)$ *and* Dirichlet *or* natural boundary conditions *be given as in (3.20) and let* $\mathcal{U}_b^h$, $\mathcal{V}_b^h$ *be approximating (conforming or non conforming) subspaces for* $\mathcal{U}_b, \mathcal{V}_b$, *e.g. satisfying Theorem 2.5.1. Then a* $\mathcal{U}_b$*-coercive continuous* $a(\cdot,\cdot)$ *implies, for* $\mathcal{U}_b^h = \mathcal{V}_b^h$, *again the* $\mathcal{U}_b^h$*-coercivity of* $a(\cdot,\cdot)$, *so there exists a constant* $\alpha > 0$ *s.t. e.g.,*

$$a(u^h, u^h) \geq \alpha(\|u^h\|_{H^1(\Omega)}^h)^2 \ \forall \ \mathcal{U}_b^h. \tag{3.31}$$

For $\mathcal{U}_b^h \neq \mathcal{V}_b^h$, *the uniform discrete inf–sup– condition is satisfied. So, there exist* $\epsilon, \epsilon' > 0$ *such that both inequalities*

$$\sup_{0 \neq v^h \in \mathcal{V}_b^h} |a(u^h, v^h)| / \|v^h\|_{\mathcal{V}}^h \geq \epsilon \|u^h\|_{\mathcal{U}}^h \quad \forall u^h \in \mathcal{U}_b^h, \quad \textit{and}$$
$$\sup_{0 \neq u^h \in \mathcal{U}_b^h} |a(u^h, v^h)| / \|u^h\|_{\mathcal{U}}^h \geq \epsilon' \|v^h\|_{\mathcal{V}}^h \quad \forall v^h \in \mathcal{V}_b^h$$

are satisfied. The $\mathcal{U}_b$*-coercivity of* $a(\cdot,\cdot)$ *implies the unique existence of the exact and discrete solutions* u_0 *and* u_0^h.

Proof We present a direct proof for conforming and nonconforming FEs. We start with the conforming case and Dirichlet and natural boundary conditions.. Then $\mathcal{U}_b^h = \mathcal{V}_b^h \subset \mathcal{U}_b$ and thus the $\mathcal{U}_b^h$ coercivity of $a(\cdot,\cdot)$ is an immediate consequence of the $\mathcal{U}_b$ coercivity.

Instead of $\mathcal{U}_b^h, \mathcal{V}_b^h$ often different spaces of discrete ansatz and test functions are chosen, hence $\mathcal{U}_b^h \neq \mathcal{V}_b^h \subset \mathcal{U}$. Again the discrete inf–sup– condition (2.15) is reduced to the $\mathcal{U}_b^h$ coercivity : For u^h we need a close by $v_u^h \in \mathcal{V}_b^h$. This exists since $u^h \in \mathcal{U}_b^h \subset \mathcal{U}_b$ and approximating $\mathcal{V}_b^h \subset \mathcal{U}_b$ approximates $\mathcal{U}_b$. Therefore for small enough h

$$\sup_{0 \neq v^h \in \mathcal{V}_b^h} |a(u^h, v^h)| / \|v^h\|_{\mathcal{V}}^h \geq |a(u^h, v_u^h)| / \|v_u^h\|_{\mathcal{V}}^h \geq \alpha \|u^h\|_{\mathcal{U}}^h / 2.$$

For the $\sup_{\{0 \neq u^h \in \mathcal{U}_b^h\}} |a(u^h, v^h)| / \|u^h\|_{\mathcal{U}}^h$ we start instead with v^h and choose $u_v^h := I^h v_s$. $\blacksquare$

Theorem 3.3.4. *For a* $\mathcal{U}_b$ *coercive* $a(\cdot,\cdot), \mathcal{U}_b \subset H^1(\Omega)$, *and* $\mathcal{U}_b^h, \mathcal{V}_b^h$ *a sequence of conforming FEs as in Chapter 2 choose* P^h *in (3.30) as the global interpolation operator* I^h *as in (2.31). Then the exact weak solution*

$u_0 \in H^m(\Omega)$ and u_0^h the corresponding FE approximate weak solution in (3.30) exist and satisfy

$$\|u_0 - u_0^h\|_{H^1(\Omega)} \leq Ch^{m-1}\|u_0\|_{H^m(\Omega)} \tag{3.32}$$

Proof : We obtain (3.32) as an immediate consequence of Theorems 2.5.1, 3.3.3 and Lemma 3.3.1.

4. Finite Elements with Variational Crimes

The approach which we present here is aimed to cover variational crimes for FE and spectral methods. We start with a list of variational crimes demonstrated for the Laplacian in Section 4.1. Then we give in Section 4.2, the main discretization ideas for FE and spectral methods specified for the case of variational crimes. After discussing linear problems we present the familiar version for the FE community based upon bilinear forms. Nonlinear problems are indicated. In Section 4.3 we develop the appropriate general discretization concepts for FEs and spectral methods.

4.1 Variational Crimes for a Simple Example

We have indicated already in the Introduction the need for variational crimes. Indeed, we want to present, modifying the original FE approach from Chapter 3, five major types and various combinations of variational crimes. We start with the simple Example 4.1, see Example 3.1, and extend it afterwards to general elliptic operators. In this Section we assume

$$\Omega \text{ is a polygonal domain.}$$

We mainly consider the influence on the corresponding linear operators, the bilinear forms and the boundary terms.

Example 4.1 *For* $u \in H_0^1(\Omega)$, *hence* Dirichlet boundary conditions, *and* $f \in H^{-1}(\Omega)$, *the operator,* A, *and the bilinear form,* $a(\cdot,\cdot)$, *are*

$$A : H_0^1(\Omega) \to H^{-1}(\Omega) \text{ and } a(\cdot,\cdot) : H_0^1(\Omega) \times H^1(\Omega) \to \mathbb{R} \text{ with } (4.1)$$
$$< Au, v >_{H^{-1}(\Omega)\times H^1(\Omega)} = a(u,v) = \int_\Omega \nabla u \nabla v + cuvdx \quad \forall v \in H_0^1(\Omega),$$

resp. The exact weak and strong solutions u_0 *are then defined resp., by the following two equations*

$$u_0 \in \mathcal{U}_b := H_0^1(\Omega) : a(u_0, v) =< Au_0, v >_{H^{-1}(\Omega)\times H^1(\Omega)} \qquad (4.2)$$
$$= \int_\Omega \nabla u_0 \nabla v + c\, u_0\, v\, dx =< f, v >_{H^{-1}(\Omega)\times H^1(\Omega)} \quad \forall\ v \in H_0^1(\Omega)$$

and by Green's Theorem, *see (3.5), and with the* outer normal ν *to* $\partial\Omega$, *see (3.2)*

$$\begin{aligned} u_0 \in \mathcal{U}_b := H^2(\Omega) \cap H_0^1(\Omega) :\ & a_s(u_0, v) = (A_s u_0, v)_{L^2(\Omega)} \\ &= \int_\Omega (-\Delta u_0 + c u_0) v \, dx = \int_\Omega f v dx \quad \forall\ v \in L^2(\Omega) \\ &= a(u_0, v) - \int_{\partial\Omega} \frac{\partial u_0}{\partial \nu}\, v ds \\ &= a(u_0, v) \quad \forall\ v \in H_0^1(\Omega). \end{aligned} \tag{4.3}$$

Remark 4.1.1. We point out once more, that it would be enough to test $A_s, a_s(\cdot,\cdot), A, a(\cdot,\cdot)$ by the same $v \in \mathcal{V}_b$. However, $\mathcal{V}_b$ is considered as subspace of different spaces, hence

$$\begin{aligned} (A_s u, v)_{L^2(\Omega)} = a_s(u, v)\ \forall\ \ v \in \mathcal{V}_b = H_0^1(\Omega),\ \text{ dense in } L^2(\Omega) \text{ versus} \\ < Au, v >_{H^{-1}(\Omega)\times H^1(\Omega)} = a(u, v)\ \forall\ \ v \in \mathcal{V}_b = H_0^1(\Omega). \end{aligned} \tag{4.4}$$

For Dirichlet or natural boundary conditions , the test functions are $v \in H_0^1(\Omega)$ or $\partial u_0/\partial\nu|_{\partial\Omega}$. Then the above boundary term $\int_{\partial\Omega} \partial u_0/\partial\nu\ v ds = 0$. As in Chapter 3 we replace the $u_0 \in \mathcal{U}_b,\ v \in \mathcal{V}_b$ in (4.2), (4.3) by elements in finite dimensional spaces $\mathcal{U}_b^h,\ \mathcal{V}_b^h$. However, now these subspaces usually will violate the above boundary conditions (or even the continuity), so $\mathcal{U}_b^h \not\subset \mathcal{U}_b, \mathcal{V}_b^h \not\subset \mathcal{V}_b$ (or even $\mathcal{U}_b^h \not\subset \mathcal{U}, \mathcal{V}_b^h \not\subset \mathcal{V}$), see (4.16)). We determine the approximate weak and strong solution $u_0^h \in \mathcal{U}_b^h$ from

$$a(u_0^h, v^h) = \int_\Omega \nabla u_0^h \nabla v^h + c\, u_0^h\, v^h \, dx = f(v^h) \quad \forall\ v^h \in \mathcal{V}_b^h \text{ and} \tag{4.5}$$

$$a_s^h(u_0^h, v^h) = \sum_{T\in\mathcal{T}^h} \int_T (-\Delta u_0^h + c\, u_0^h) v^h \, dx = f(v^h) \quad \forall\ v^h \in \mathcal{V}_b^h. \tag{4.6}$$

Again as in chapter 3, the relation between weak and strong discrete problem is governed by the errors in (3.9),(3.10). Now $\mathcal{V}_b^h \subset \mathcal{V}_b$ and $\mathcal{U}_b^h \subset \mathcal{U}_b$ are violated. We introduce the modified projectors

$$\begin{aligned} &Q'^h \in \mathcal{L}(\mathcal{V}', \mathcal{V}_b^{h\prime}) \ \text{ by } \ < Q'^h f - f, v^h >_{\mathcal{V}'\times\mathcal{V}_b^{h\prime}} = 0\ \forall\ \ v^h \in \mathcal{V}_b^h \text{ and} \\ &Q_s'^h \in \mathcal{L}(\mathcal{V}', \mathcal{V}_b^{h\prime}) \ \text{ by } (Q_s'^h f - f, v^h)_{L^2(\Omega)} = 0\ \forall\ \ v^h \in \mathcal{V}_b^h. \end{aligned} \tag{4.7}$$

Mind that the $Q'^h, Q_s'^h$ are tested only in $\mathcal{V}_b^h$. Whenever $A, A_{s,h}$ are defined on $\mathcal{U}_b^h$, then (4.5), (4.6) require to determine the weak and strong discrete solutions as

$$u_0^h \in \mathcal{U}_b^h; < Q'^h(Au_0^h - f), v^h >_{H^{-1}(\Omega)\times H^1(\Omega)} = 0\ \forall\ v^h \in \mathcal{V}_b^h, \text{ and} \tag{4.8}$$

$$u_0^h \in \mathcal{U}_b^h : (Q_s'^h(A_{s,h} u_0^h - f), v^h)_{L^2(\Omega)} = 0\ \forall\ v^h \in \mathcal{V}_b^h, \text{ resp.} \tag{4.9}$$

For the different types of non conforming FEs, we have to admit violations of the above assumptions $\mathcal{U}_b^h \subset \mathcal{U}_b$, $\mathcal{V}_b^h \subset \mathcal{V}_b$ and $\mathcal{U}_b^h \not\subset \mathcal{U}, \mathcal{V}_b^h \not\subset \mathcal{V}$. The $a(u^h, v^h)$ has to be defined on $\mathcal{U}_b^h \times \mathcal{V}_b^h$. Then the spaces do not fit anymore. The influence of the non vanishing boundary term $\int_{\partial\Omega} \partial u^h/\partial\nu \; v^h ds$, see (4.27), and its corresponding interior boundary term $\int_e v_l^h \left[\frac{\partial u^h}{\partial \nu_e}\right] + [v^h]\frac{\partial u_r^h}{\partial \nu_e}\, ds$ in (3.9),(3.10), have to be studied carefully. This is caused by the problem: (4.2) characterizes the solution u_0 of $-\Delta u_0^h + c\, u_0^h = f$ only in $\mathcal{U}_b \times \mathcal{V}_b$, but not in a $\mathcal{U}_b \times \tilde{\mathcal{V}}_b$ if the boundary conditions for the $v \in \tilde{\mathcal{V}}_b$ or the continuity or differentiability across the interior boundaries, e, should be violated. In fact, in FEMs the "detour" via the strong form $A_s u_0 = f$, compare (4.3) and Remark 3.2.2, is only used to estimate the violation of $\mathcal{V}_b^h \not\subset \mathcal{U}_b$ or $\mathcal{V}_b^h \not\subset \mathcal{V}$. For finite difference or spectral methods it is employed by approximate evaluations of A_s, see 3) below. A quadrature approximation of the strong forms, see 4), allows a re-interpretation as collocation method, see 5), e.g. the Doedel collocation

For Example 4.1 we find the following five cases. We denote the first two cases as:

Non conforming Finite Elements:

1) *Finite Elements violating the boundary conditions:* Let $\mathcal{U}_b^h, \mathcal{V}_b^h \subset H^1(\Omega)$, with $\mathcal{U}_b^h, \mathcal{V}_b^h \not\subset H_0^1(\Omega)$ $B(v^h) = B_1(v^h) = v^h|_{\partial\Omega}$, or more general operators

$$\begin{aligned} \mathcal{U}_b^h &= \{u^h \in \mathcal{U}^h \subset \mathcal{U} : B(u^h) \approx 0\} \not\subset \mathcal{U}_b \text{ and} \\ \mathcal{V}_b^h &= \{v^h \in \mathcal{V}^h \subset \mathcal{V} : B_1(v^h) \approx 0\} \not\subset \mathcal{V}_b. \end{aligned} \tag{4.10}$$

This case occurs e.g. for finite elements where $u^h : \Omega \to \mathbb{R}^m$, and, e.g., for Dirichlet conditions, $u^h(P_j) = 0$ only in certain points $P_j \in \partial\Omega$ or even only close to $\partial\Omega$ forcurved boundaries, see Section (2.7). Still the original bilinear $a(\cdot,\cdot)$ and linear forms $f(\cdot)$ are defined in $\mathcal{U}^h \times \mathcal{V}^h$ and $\mathcal{V}^h$. The exact and approximate solutions, $u_0 \in \mathcal{U}_b$ and $u_0^h \in \mathcal{U}_b^h$, solve,

$$a(u_0, v) = f(v) \;\forall\; v \in \mathcal{V}_b, \quad a(u_0^h, v^h) = f(v^h) \;\; \forall\; v^h \in \mathcal{V}_b^h. \tag{4.11}$$

Mind that in contrast to FEs $v^h \in \mathcal{V}_b$ for our $v^h \in \mathcal{V}_b^h \not\subset \mathcal{V}_b$ the second condition $a(u_0^h, v^h) = f(v^h) \;\; \forall\; v^h \in \mathcal{V}_b^h$ *is no subset* of the first condition $a(u_0, v) = f(v) \;\; \forall\; v \in \mathcal{V}_b$. So we have to be prepared to pay for this extension from $\mathcal{V}_b^h \subset \mathcal{V}_b$ to $\mathcal{V}_b^h \not\subset \mathcal{V}_b$, see (4.13).

By (4.3), we obtain, mind that $-\Delta u^h + c\, u^h$ is not defined on Ω,

$$\begin{aligned} a_s^h(u^h, v^h) &:= \sum_{T \in \mathcal{T}^h} \int_T (-\Delta u^h + c\, u^h)\, v^h\, dx \;\forall\; v^h \in \mathcal{V}_b^h, \\ a(u^h, v^h) &= a_s^h(u^h, v^h) + \sum_{e \in T} \int_e v_1^h \left[\frac{\partial u^h}{\partial \nu_e}\right] ds + \int_{\partial\Omega} \frac{\partial u^h}{\partial \nu}\, v^h ds, \end{aligned} \tag{4.12}$$

since $[v^h]|_e = 0$. Here ν_e is again the unit normal vector to the edge e and $[v^h]$ and $[\partial u^h/\partial \nu_e]$ the corresponding oriented jumps of v^h and

$\partial u^h/\partial \nu_e$ across e, resp. Since we only will estimate the absolute value of $\sum_{e\in\mathcal{T}^h}\int_e v^h[\partial u^h/\partial\nu_e]ds$ we do not have to specify the direction of $\partial u/\partial\nu_e$ more precisely.

For the weak exact and discrete solutions, here $u_0 \in H^2(\Omega)$ and $u_0^h \in \mathcal{U}_b^h$, resp., we find as a consequence of (4.3) that

$$\begin{aligned} a(u_0 - u_0^h, v^h) &= (-\Delta u_0 + cu_0, v^h) + \int_{\partial\Omega} \frac{\partial u_0}{\partial \nu}\, v^h ds - f(v^h) \\ &= \int_{\partial\Omega} \frac{\partial u_0}{\partial \nu}\, v^h ds \;\forall\; v^h \in \mathcal{V}_b^h. \end{aligned} \tag{4.13}$$

In this case the additional term in (4.27) does not show up. For $u_0 \in H^2(\Omega)$ the $\partial u_0/\partial\nu$ and the v^h are continous along edges $e \in \mathcal{T}^h$. So, for $v \in \mathcal{V}_b^h$ violating the boundary conditions, we have to expect $a(u_0 - u_0^h,\ v^h) \neq 0$ for a general $v^h \in \mathcal{V}_b^h \not\subset H_0^1(\Omega)$.
For the strtong exact and discrete solutions, here $u_0 \in H^2(\Omega)$ and $u_0^h \in \mathcal{U}_b^h$, the situation is simpler: A cobination of (4.3) and (4.6) yields

$$a_s(u_0 - u_0^h, v^h) = (-\Delta u_0 + cu_0, v^h) - f(v^h) = 0 \;\forall\; v^h \in \mathcal{V}_b^h. \tag{4.14}$$

Here, and in the following cases 2) - 5) we compute the $a(u_0 - u_0^h, v^h)$ and their generalizations. In (4.13) a detour via the strong operator $(A_s u_0, v^h)_{L^2(\Omega)}$ (, well defined,) is the basic tool. We will then estimate the differences, $a(u_0 - u_0^h, v^h)$, in Chapters 5 and 6.

2) *Finite Elements violating the continuity conditions:* Let $\mathcal{U}_b^h, \mathcal{V}_b^h \not\subset H^1(\Omega)$, e.g. , we have non continuous FEs: They are defined on a subdivision $\mathcal{T}^h$ for a polygonal Ω, discontinuous across the edges e, more generally,

$$\mathcal{U}_b^h = \{u^h \in \mathcal{U}^h\} \not\subset \mathcal{U},\ \mathcal{V}_b^h = \{v^h \in \mathcal{V}^h\} \not\subset \mathcal{V} \text{ discontinuous across } e \tag{4.15}$$

This happens for non conforming finite elements, e.g. the Crouzeix-Raviart FE, [18, 17].

Although the strong equation is not needed directly, we formulate it here, to allow an appropriate notation for both cases. This will be needed for Section 4.3. For (4.15) the original norms, e.g., $\|\cdot\|_{H^1(\Omega)}$, the bilinear and sometimes the linear form as well, $a(\cdot,\cdot)$, $a_s(\cdot,\cdot)$, $f(\cdot)$, are no longer defined. They have to be replaced by extended $\|\cdot\|_{H^1(\Omega)}^h$, $a^h(\cdot,\cdot)$, $a_s^h(\cdot,\cdot)$, $f^h(\cdot)$, defined for $u^h \notin \mathcal{U}_b$, $v^h \notin \mathcal{V}_b^h$ separately on the $T \in \mathcal{T}^h$.

We start with the $a^h(\cdot,\cdot)$, $a_s^h(\cdot,\cdot) : \mathcal{U}_b^h \times \mathcal{V}_b^h \to \mathbb{R}$: Let

$$a^h(u^h, v^h) := \sum_{T\in\mathcal{T}^h} \int_T \nabla u^h\, \nabla v^h + c\, u^h\, v^h\, dx \tag{4.16}$$

$$a_s^h(u^h, v^h) := \sum_{T\in\mathcal{T}^h} \int_T (-\Delta u^h + c\, u^h) v^h\, dx \tag{4.17}$$

$$\text{both defined } \forall\; u^h \in \mathcal{U}_b^h,\ v^h \in \mathcal{V}_b^h.$$

To avoid over indexing, in the FEM community the same notation is used for this original and the *extended bilinear form* $a^h(\cdot,\cdot)$

$$\begin{aligned} a^h(\cdot,\cdot) : \mathcal{U}_b^h \times \mathcal{V}_b^h \to \mathbb{R} \text{ and } a^h(\cdot,\cdot) : \{\mathcal{U}_b \cup \mathcal{U}_b^h\} \times \{\mathcal{V}_b \cup \mathcal{V}_b^h\} \to \mathbb{R}, \quad (4.18) \\ \text{continuous with } a(\cdot,\cdot) : \mathcal{U}_b \times \mathcal{V}_b \to \mathbb{R} \\ \text{and identical with} a(\cdot,\cdot) \equiv a^h(\cdot,\cdot)|_{\mathcal{U}_b \times \mathcal{V}_b}. \end{aligned}$$

We proceed similarly for the $a_s^h(\cdot,\cdot)$ in (4.17) and the $\tilde{a}_s^h(\cdot,\cdot)$ in (4.42)

$$a_s^h(\cdot,\cdot),\ \tilde{a}^h(\cdot,\cdot) : \{\mathcal{U}_b \cup \mathcal{U}_b^h\} \times \{\mathcal{V}_b \cup \mathcal{V}_b^h\} \to \mathbb{R}, \tag{4.19}$$

where $\tilde{a}^h(\cdot,\cdot)$ is only defined on smooth enough subspaces of $\mathcal{U}_b \times \mathcal{V}_b$. The restrictions to $\mathcal{U}_b \times \mathcal{V}_b$ satisfy

$$a^h(\cdot,\cdot)|_{\mathcal{U}_b \times \mathcal{V}_b} = a(\cdot,\cdot) \text{ and } a_s^h(\cdot,\cdot)|_{\mathcal{U}_b \times \mathcal{V}_b} = a_s(\cdot,\cdot). \tag{4.20}$$

In Section 4.3 we will consider variational and classical consistency errors. The variational consistency error is usually defined via the bilinear forms, e.g., the $a^h(\cdot,\cdot)$, $a_s^h(\cdot,\cdot)$, $\tilde{a}^h(\cdot,\cdot)$, $\tilde{a}_s^h(\cdot,\cdot)$. The classical consistency error relates the operators $A, A_s : \mathcal{U}_b \to \mathcal{V}_b'$ and their discrete counterparts. In our case of piecewise continuous FEs, the A and A_s are not defined for the $u^h \in \mathcal{U}_b^h$, similarly below. At the other side, we need for the classical consistency machinery a relation between the original operator, A, or a close relative, A_h, and its discretization, A^h. So we define, in full analogy to the extended bilinear forms in (4.18), (4.19), extensions, A_h, $A_{s,h}$, of [1] the A, $A_s u$, see (2.17), (2.18):

$$\begin{aligned} \text{For } u^h \in \{\mathcal{U}_b \cup \mathcal{U}_b^h\} \text{ we define} \quad & A_h, A_{s,h} : \{\mathcal{U}_b \cup \mathcal{U}_b^h\} \to \{\mathcal{V}_b' \cup \mathcal{V}_b^{'h}\} \text{ as} \\ < A_h u^h, v^h >_{H_h^{-1}(\Omega) \times H_h^1(\Omega)}^h & := a^h(u^h, v^h)\ \forall v^h \in \{\mathcal{V}_b \cup \mathcal{V}_b^h\},\ \text{hence,} \\ & = \sum_{T \in \mathcal{T}^h} \int_T \nabla u^h \nabla v^h + c\, u^h v^h\, dx, \quad (4.21) \\ (A_{s,h} u^h, v^h)_{L^2(\Omega)} & := a_s^h(u^h, v^h)\ \forall v^h \in \{\mathcal{V}_b \cup \mathcal{V}_b^h\},\ \text{hence,} \\ & = \sum_{T \in \mathcal{T}^h} \int_T (-\Delta u^h + c u^h) v^h\, dx \text{ and } \quad (4.22) \\ A_h, A_{s,h} : \{\mathcal{U}_b \cup \mathcal{U}_b^h\} & \to \{\mathcal{V}_b' \cup \mathcal{V}_b^{'h}\} \text{ are continuous with } A,\ A_s. \end{aligned}$$

Similarly to the notation and the restriction of the above $a^h(\cdot,\cdot)$, $a_s^h(\cdot,\cdot)$ in (4.16), (4.18)-(4.20), we use the same notation for the different restrictions of the $A_h, A_{s,h}$. They satisfy

[1] in contrast to the $a^h(\cdot,\cdot)$, $a_s^h(\cdot,\cdot)$ a.s.o., we use the index $_h$ to denote the extended operators as A_h, $A_{s,h}$. This allows the appropriate notation for all the discrete operators as A^h, $A_s^h : \mathcal{U}_b^h \to \mathcal{V}_b^{'h}$.

$$A_h|_{\mathcal{U}_b \to \mathcal{V}_b'} = A,\ A_h|_{\mathcal{U}_b^h \to \mathcal{V}_b'^h} = A^h, \text{ and} \tag{4.23}$$
$$A_{s,h}|_{\mathcal{U}_b \to \mathcal{V}_b'} = A_s,\ A_{s,h}|_{\mathcal{U}_b^h \to \mathcal{V}_b'^h} = A_s^h$$

To determine the discrete solution, u_0^h, we restrict the $A_h,\ A_{s,h}$ to $A_h,\ A_{s,h} : \mathcal{U}_b^h \to \mathcal{V}_b^h$. In fact, in the standard approach, the bilinear forms, $a^h(\cdot,\cdot),\ a_s^h(\cdot,\cdot)$, directly define the discrete linear operators $A^h, A_s^h : \mathcal{U}_b^h \to \mathcal{V}_b^{h'}$, for fixed $u^h \in \mathcal{U}_b^h$ by

$$u^h \in \mathcal{U}_b^h : < A^h u^h, v^h >_{H_h^{-1}(\Omega) \times H_h^1(\Omega)}^h := a^h(u^h, v^h)\ \forall v^h \in \mathcal{V}_b^h \tag{4.24}$$
$$u^h \in \mathcal{U}_b^h : (A_s^h u^h, v^h)_{L^2(\Omega)} := a_s^h(u^h, v^h)\ \forall v^h \in \mathcal{V}_b^h. \tag{4.25}$$

The $A_h,\ A_{s,h}$ are related to these operators $A^h,\ A_s^h$ in (4.24), (4.25) as

$$A^h = Q'^h A_h|_{\mathcal{U}_b^h} \text{ and } A_s^h := Q_s'^h A_{s,h}|_{\mathcal{U}_b^h}. \tag{4.26}$$

This is needed in the further discussion.

Certainly $Q'^h f$ is defined for any $f \in \mathcal{V}_b'$, e.g., $Q'^h Au$. In this case, Q'^h requires inappropriately testing by $v^h \in \mathcal{V}_b^h \not\subset \mathcal{V}_b$. Sometimes this discrepancy might cause variational crimes, see (**??**). We handle them, not by comparing Au and $A^h u^h$ directly, but by using the "detour" via the strong problem.

Below, see 3.), we will even approximate the integrals $\int_T \nabla u^h \nabla v^h + c u^h v^h\, dx$ by quadrature formulas.

Analogously to (4.3), the $a^h(\cdot,\cdot),\ a_s^h(\cdot,\cdot)$ are related as

$$a_s^h(u^h, v^h) := \sum_{T \in \mathcal{T}^h} \Big(\int_T (-\Delta u^h + c\, u^h)\, v^h\, dx\ \forall\ v^h \in \mathcal{V}_b^h, \tag{4.27}$$
$$a(u^h, v^h) = a_s^h(u^h, v^h) + \sum_{e \in T} \int_e v_1^h \left[\frac{\partial u^h}{\partial \nu_e} \right] ds \Big) + \int_{\partial\Omega} \frac{\partial u^h}{\partial \nu}\, v^h ds,$$

Then $u_0^h \in \mathcal{U}_b^h$ is defined as (weak) solution of

$$a^h(u_0^h, v^h) = f(v^h)\ \forall\ v^h \in \mathcal{V}_b^h, \text{ sometimes } f^h(v^h) = \sum_{T \in \mathcal{T}^h} \int_T f\, v^h dx \tag{4.28}$$

By combining (4.11), (4.16), (**??**), (4.28), we obtain, see (4.13), for $u_0 \in H^2(\Omega)$,

$$a^h(u_0 - u_0^h, v^h) = \sum_{T \in \mathcal{T}^h} \int_{\partial T} \frac{\partial u_0}{\partial \nu_e}\, v^h ds = \sum_{e \in \mathcal{T}^h} \int_e \frac{\partial u_0}{\partial \nu_e} [v^h] ds. \tag{4.29}$$

Similary to (4.13) the additional terms in (**??**) drop out sice $\partial u_0 / \partial \nu_e$ is continous across the edges e. If the boundary conditions are violated in $\mathcal{V}_b^h$ as well, the corresponding term from (4.13) has to be added. If we assume $v^h = 0$

outside of the $T \in \mathcal{T}^h$ this is even included in (4.29).

Similarly to (**??**) we find again

$$a_s^h(u_0 - u_0^h, v^h) = \sum_{T \in \mathcal{T}^h} \int_T (-\Delta u_0 + c u_0) v^h dx - f(v^h) = 0 \ \forall \ v^h \in \mathcal{V}_b^h \quad (4.30)$$

3) *Approximations for the* $A(\cdot), A_s(\cdot), a^h(\cdot,\cdot), a_s^h(\cdot,\cdot), f(\cdot), f^h(\cdot)$: Independent of violated boundary conditions or continuity, we allow approximations to the original or extended linear operators, bilinear, linear forms, pairings and scalar products, the $A\,u^h$, $A_s\,u^h$, $a^h(u^h, v^h)$, $a_s^h(u^h, v^h)$, $f(v^h)$ $f^h(v^h)$, $< f, v^h >, (f, v^h)$. These approximations may be quadrature formulas, see 4) below, divided differences, or Fourier collocation derivatives and de-aliasing techniques for spectral methods. We denote these approximations as $\tilde{A}_h$, $\tilde{A}_{s,h}$, $\tilde{a}^h(\cdot,\cdot), \tilde{a}_s^h(\cdot,\cdot), \tilde{f}_h(\cdot), <\cdot,\cdot\tilde{>}^h, (\cdot,\cdot\tilde{)}^h$. We keep this notation for the *extensions to smooth enough subspaces* of $\mathcal{U}_b, \mathcal{V}_b$. This allows the necessary evaluation of functions and divided differences, analogous to (4.18), (4.19). Otherwise, point evaluations, e.g., as needed in quadrature rules, are not defined.

First, we modify the projectors as

$$\tilde{Q}'^h \in \mathcal{L}(\mathcal{V}', \mathcal{V}_b^{h\prime}) \ \text{ by } < \tilde{Q}'^h f - f, v^h \tilde{>}^h_{H_h^{-1}(\Omega)\times H_h^1(\Omega)} = 0 \ \forall \ \ v^h \in \mathcal{V}_b^h, \text{ and}$$

$$\tilde{Q}_s'^h \in \mathcal{L}(\mathcal{V}', \mathcal{V}_b^{h\prime}) \ \text{ by } (\tilde{Q}_s'^h f - f, v^h\tilde{)}^h_{L^2(\Omega)} = 0 \ \forall \ \ v^h \in \mathcal{V}_b^h, \text{ or} \quad (4.31)$$

$$\tilde{Q}'^h f - f \tilde{\perp}^h \mathcal{V}_b^h \text{ and } \tilde{Q}_s'^h f - f \tilde{\perp}_s^h \mathcal{V}_b^h. \quad (4.32)$$

Here the

$$\langle\cdot,\cdot\tilde{\rangle}^h := \langle\cdot,\cdot\tilde{\rangle}^h_{H_h^{-1}(\Omega)\times H_h^1(\Omega)} \text{ and } (\cdot,\cdot\tilde{)}^h := (\cdot,\cdot\tilde{)}^h_{L^2(\Omega)} \quad (4.33)$$

indicate the piecewise quadrature or other approximations for the $H_h^{-1}(\Omega) \times H_h^1(\Omega)$ pairing as in (4.16) and the $L^2(\Omega)$ scalar product in appropriate smooth subspaces. As in (4.26), see Remark 3.3.2, we denote as

$$\begin{aligned} &\tilde{A}_h, \tilde{A}_{s,h} \ \text{ the } \textit{extended approximations} \text{ for } A, \ A_s, \ \text{ see (4.21), (4.22),} \\ &\tilde{A}^h, \tilde{A}_s^h \text{ the discrete operators } \tilde{A}^h := \tilde{Q}'^h \tilde{A}_h|_{\mathcal{U}_b^h}, \ \tilde{A}_s^h := \tilde{Q}_s'^h \tilde{A}_{s,h}|_{\mathcal{U}_b^h} \end{aligned} \quad (4.34)$$

Similarly to the notation and extension of the above $a^h(\cdot,\cdot)$, $a_s^h(\cdot,\cdot)$ in (4.16), (4.17), (4.20), we use the same notation for the different extension and restrictions of the $A_h, A_{s,h}$. Mind, that again the $\tilde{Q}'^h$, $\tilde{Q}_s'^h$ inappropriately test with $v^h \in \mathcal{V}_b^h$, thus causing crime errors. Instead of the equality in (4.23), we find approximations

$$\tilde{A}_h|_{\mathcal{U}_b \to \mathcal{V}_b'} \ \approx \ A, \text{ and } \tilde{A}_{s,h}|_{\mathcal{U}_b \to \mathcal{V}_b'} \ \approx \ A_s. \quad (4.35)$$

Then the weak and strong discrete solutions are defined by

$$u_0^h \in \mathcal{U}_b^h \text{ by } \tilde{A}^h u_0^h = \tilde{Q}^{'h}\tilde{A}_h u_0^h = \tilde{Q}^{'h} f \text{ or} \tag{4.36}$$
$$< \tilde{Q}^{'h}(\tilde{A}_h u_0^h - f), v^h >^h_{H_h^{-1}(\Omega)\times H_h^1(\Omega)} = 0 \ \forall \ \ v^h \in \mathcal{V}_b^h, \text{ and}$$
$$u_0^h \in \mathcal{U}_b^h \text{ by } \tilde{A}_s^h u_0^h = \tilde{Q}_s^{'h}\tilde{A}_{s,h} u_0^h = \tilde{Q}_s^{'h} f \text{ or} \tag{4.37}$$
$$(\tilde{Q}_s^{'h}(\tilde{A}_{s,h} u_0^h - f), v^h)^h_{L^2(\Omega)} = 0 \ \forall \ v^h \in \mathcal{V}_b^h, \ \text{ resp.}$$

The most important examples for these approximations are

4) *Quadrature approximations and approximate projectors* $\tilde{Q}^{'h}$: The original definition of the approximate weak solution u_0^h requires in (4.5)

$$a^h(u_0^h, v^h) - f^h(v^h) := \sum_{T\in\mathcal{T}^h} \int_T \nabla u_0^h \nabla v^h + c\, u_0^h\, v^h \ - f v^h dx$$
$$= 0 \ \forall \ v^h \in \mathcal{V}_b^h \ , \tag{4.38}$$

similarly for the strong solution. Non conforming FEs will be discussed in more detail only for the following e.g. Doedelcollocation methods. We replace the exact vanishing of the inner products in (4.38) by the vanishing of the quadrature formulas. We start with

$$\tilde{f}(v^h) := (f \cdot v^h)^{\tilde{}h} := (f, v^h)^{\tilde{}h} \tag{4.39}$$
$$:= \sum_{T\in\mathcal{T}^h} \sum_{P_i\in T} w_i f(P_i) v^h(P_i) \approx (f, v^h)_{L^2(\Omega)}$$

for continuous f and v^h. This is extended to the more general case, e.g., of

$$< f, v^h >_{H^{-1}(\Omega)\times H^1(\Omega)} = \int_\Omega f_{-1}^T \ \nabla \ v^h + f_0 v^h dx \in \mathbb{R} \tag{4.40}$$

(with $\int_\Omega f_{-1}^T \ \nabla \ v^h dx = \int_\Omega (f_{-1}, \ \nabla \ v^h) dx$) as

$$< f, v^h >^{\tilde{}h} = (f_{-1}^T \ \nabla \ v^h)^{\tilde{}h} + (f_0, v^h)^{\tilde{}h} \tag{4.41}$$
$$= \sum_{T\in\mathcal{T}^h} \sum_{P_i\in T} w_i (f_{-1}^T(P_i) \ \nabla \ v^h(P_i) + f_0(P_i) v^h(P_i))$$
$$\approx < f, v^h >_{\mathcal{V}'\times\mathcal{V}_b} .$$

Again this approximation requires continuous $f_{-1}^T, \ \nabla \ v^h, f_0, v^h$ s.t. the evaluations in the P_i are possible. Sometimes we replace $f^h(v^h)$ in (4.38) by $< f^h, \ v^h >^h$. Often, we only need the $< Au^h, v^h >^h$ part and have for f, v^h the $(f, v^h)^{\tilde{}h}$. Nevertheless, there are cases where the $< f, v^h >^h$ still is necessary. But usually, we formulate the equations only for $(f, v^h)^{\tilde{}h} = \tilde{f}(v^h)$.

Now we apply (4.41) to (4.38) and have defined a new problem. Determine the approximate weak and strong solution $u_0^h \in \mathcal{U}_b^h$ s.t.

$$\tilde{a}^h(u_0^h, v^h) - \tilde{f}(v^h) := (\nabla u_0^h \nabla v^h + c u_0^h v^h)\tilde{}^h - \tilde{f}(v^h) :=$$
$$\sum_{T\in\mathcal{T}^h} \sum_{P_i\in T} w_i \left((\nabla u_0^h \nabla v^h + c u_0^h v^h)(P_i) - (f v^h)(P_i)\right) = 0 \text{ and} \quad (4.42)$$
$$\tilde{a}_s^h(u_0^h, v^h) - \tilde{f}(v^h) := ((-\Delta u_0^h + c u_0^h)v^h)\tilde{}^h - \tilde{f}(v^h) := 0 \ \forall\ v^h \in \mathcal{V}_b^h,$$

resp. Again, the additional terms in (3.10) drop out, since for conforming $\mathcal{V}_b^h$ and $u_0 \in H^2(\Omega)$ the v^h are continous and $[\partial u/\partial \nu_e] = 0$ on the edges e. The approximate projectors and discrete operators $\tilde{Q}'^h, \tilde{Q}_s'^h$ and $\tilde{A}^h, \tilde{Q}_s^h$ are defined as in (4.7), (4.34). For the weak solutions, the error, corresponding to (4.13), (4.29), (4.42), has then the form

$$\begin{aligned}
\tilde{a}^h(u_0 - u_0^h, v^h) &= \sum_{P_j\in T} \left(w_j(\nabla u_0 \nabla v^h + c u_0 v^h)(P_j)\right) - < f,\ v^h\tilde{>}^h \\
&= \sum_{P_j\in T} \left(w_j(\nabla u_0 \nabla v^h + c u_0 v^h)(P_j)\right) - < f,\ v^h\tilde{>}^h \\
&\quad - \sum_{T\in\mathcal{T}^h} \int_T \left((\nabla u_0 \nabla v^h + c u_0 v^h)\right) dx - < f,\ v^h > \\
&= \left(\tilde{a}^h(u_0, v^h) - a^h(u_0, v^h)\right) \qquad (4.43) \\
&\quad + \left(< f\ ,\ v^h >^h - < f\ ,\ v^h\tilde{>}^h\right) \ \forall\ v^h \in \mathcal{V}_b^h.
\end{aligned}$$

Similarly, we find for the strong solutions, u_0 and u_0^h, if they exist,

$$\begin{aligned}
\tilde{a}_s^h(u_0 - u_0^h, v^h) &= \sum_{P_j\in T} \left(w_j((-\Delta u_0 + c u_0)v^h)(P_j)\right) - (f,\ v^h)\tilde{}^h \\
&= \left(\tilde{a}_s^h(u_0, v^h) - a_s^h(u_0, v^h)\right) \qquad (4.44) \\
&\quad + \left((f\ ,\ v^h)^h - (f\ ,\ v^h)\tilde{}^h\right) \ \forall\ v^h \in \mathcal{V}_b^h.
\end{aligned}$$

Both forms require smooth enough u_0, f to evaluate $\Delta u_0(P_j),\ \nabla u_0(P_j),\ f(P_j)$. The estimates will be based on the quadrature errors in the last two lines.

Remark 4.1.2. In these quadrature approximations (4.43), (4.44), the quadrature points $P_{j,T}$ may be chosen totally independent of the interpolation points for the FEs. This contrasts to the situation for the following collocation methods. Here quadrature and interpolation= collocation points coincide, see (4.49), below.

5) *Collocation methods* We obtain collocation methods in several steps. We start with the strong problem (4.3) and consider the corresponding strong solutions. Then we replace, as in 4) the integrals over Ω by quadrature formulas, see (4.42). Next, we employ an interpolation basis for $\mathcal{V}_b^h$. Finally, we estimate the quadrature errorsfor the exact forms and relate the strong and weak forms. This will allow to use the coercivity.

We explicitly formulate (4.42) as quadrature formula for the strong bilinear forms, if $< f, v^h \tilde{>}^h = (f, v^h)\tilde{}^h$, as

$$
\begin{aligned}
(f, v^h)\tilde{}^h &= \sum_{T\in\mathcal{T}^h} \sum_{P_i\in T} w_i f(P_i) v^h(P_i) = \\
\tilde{a}_s^h(u^h, v^h) &= \sum_{T\in\mathcal{T}^h} \left((-\Delta u^h + c u^h, v^h)\tilde{}^h_{|_T}\right) \\
&= \sum_{T\in\mathcal{T}^h} \Big(\sum_{P_j\in T} w_j (-\Delta\, u^h + c\, u^h)(P_j) v_i^h(P_j)\Big) \;\forall\; v^h \in \mathcal{V}_b^h .
\end{aligned}
\tag{4.45}
$$

To estimate the errors later on, we have to recall, see see (4.3), (4.27), (4.13), (4.16), (**??**), (4.29), (4.42), that

$$
\begin{aligned}
\tilde{a}^h(u^h, v^h) = \;& \tilde{a}_s^h(u^h, v^h) + (\tilde{a}^h(u^h, v^h) - a(u^h, v^h)) \\
&+ \sum_{e\in\mathcal{T}^h} \int_e v_l^h \left[\frac{\partial u^h}{\partial \nu_e}\right] + [v^h] \frac{\partial u_r^h}{\partial \nu_e}\, ds + (a_s^h(u^h, v^h) - \tilde{a}_s^h(u^h, v^h)) \\
&\quad \text{for FEs satisfying the boundary conditions} \\
&+ \text{additionally} + \int_{\partial\Omega} \frac{\partial u^h}{\partial \nu}\, v^h ds \\
&\quad \text{for violated boundary conditions}
\end{aligned}
\tag{4.46}
$$

The relation of the collocation formulation to the above weak quadrature bilinear formulation, see (4.45), is given for the strong solution $u_0^h \in \mathcal{U}_b^h$ and $< f, v^h >= (f, v^h)_{L^2(\Omega)}$, as

$$
\begin{aligned}
\tilde{a}^h(u^h, v^h) - \tilde{a}_s^h(u^h, v^h) = \;& (\tilde{a}^h(u^h, v^h) - a(u^h, v^h)) \\
&+ (a_s^h(u^h, v^h) - \tilde{a}_s^h(u^h, v^h)) + \sum_{e\in\mathcal{T}^h} \int_e v_l^h \left[\frac{\partial u^h}{\partial \nu_e}\right] ds \\
&\quad \text{for conforming FEs} \\
&+ \text{additionally} + \sum_{e\in\mathcal{T}^h} \int_e [v^h] \frac{\partial u_r^h}{\partial \nu_e}\, ds \\
&\quad \text{for discontinuous FEs} \\
&+ \text{additionally} + \int_{\partial\Omega} \frac{\partial u^h}{\partial \nu}\, v^h ds \\
&\quad \text{for violated boundary conditions .}
\end{aligned}
\tag{4.47}
$$

The quadrature and the non conformity error terms $\int_{\partial\Omega} \partial u_0^h/\partial\nu\; v^h ds$ and $\sum_{e\in\mathcal{T}^h} \int_e \partial u_0^h/\partial\nu_e [v^h] ds$ have to be studied separately, see Sections 6.2, 6.3.

This shows that the difference $\tilde{a}^h(u^h, v^h) - \tilde{a}_s^h(u^h, v^h)$ can be small only,if the quadrature errors for $a(\cdot,\cdot)$ and $a_s^h(\cdot,\cdot)$ are small and if

$$[v^h](P_e^i) = 0 \text{ and } [\frac{\partial u^h}{\partial \nu}](Q_e^j) = 0 \tag{4.48}$$
$$\text{for sufficiently many points } P_e^i, Q_e^j \in \overline{e}.$$

To show the equivalence of the strong quadrature and the collocation formulation we choose specific FE spaces $\mathcal{V}_b^h$: We assume the $\mathcal{N}$ in Definition 2.2.2 consists of exactly those d points used in (4.45). Hence, any $v^h \in \mathcal{V}_b^h$ has to be uniquely determined by the $v^h(P_j)$, $\forall\ P_j \in \overline{T}$, $\forall\ T \in \mathcal{T}^h$. Thus, with the *Dirac delta functions* $\delta(P_i)$ $\mathcal{N}_T = \{\delta(P_i) : \forall\ P_i \in \overline{T}\}$, $\forall\ T \in \mathcal{T}^h$. So, the following straight forward definition can be used for conforming and non conforming cases, see Definition 2.2.2 and (2.31). We assume:

$$\begin{aligned} T \in \mathcal{T}^h, T = F_T(K) \text{ is affine equivalent to } K \text{ and} \\ (K,\ \mathcal{P},\ \mathcal{N}) \text{ is a FE} \quad \text{s.t. } \mathcal{N}_T = \{\delta(P_i) : \forall P_j \in \overline{K} \text{ in (4.42)}\} \\ \text{induces a unisolvent basis } \mathcal{N} \text{ for } \mathcal{P}'. \end{aligned} \tag{4.49}$$

Furthermore, we have to guarantee (4.48). If the P_e^i are included in $\mathcal{N}_T$, then automatically $[v^h](P_e^i) = 0$. However the next condition $[\frac{\partial u^h}{\partial \nu}](Q_e^j) = 0$ does not directly fit to collocation. For later reference we formulate these two conditions to define $\mathcal{U}_b^h$, $\mathcal{V}_b^h$ as

$$\mathcal{U}^h := \{u^h \in FEs : [\frac{\partial u^h}{\partial \nu}](Q_e^j) = 0 \forall e \in \mathcal{T}^h \text{ for sufficiently many points } Q_e^j \in \overline{e}\} \tag{4.50}$$
$$\mathcal{V}^h := \{u^h \in FEs : [v^h](P_e^i) = 0 \forall e \in \mathcal{T}^h \text{ for sufficiently many points } P_e^i \in \overline{e}\} \tag{4.51}$$

This property has to be combined with a given set of collocation points in (4.49). We have to guarantee that in fact a $\mathcal{P}, \mathcal{P}_{m-1} \subseteq \mathcal{P} \subseteq P_{m+\tau}$, see (2.34), exists, s.t. $(K, \mathcal{P}, \mathcal{N})$ unisolvantly defines a FE. We certainly have to show that we are not talking about an empty set of methods and come back to the Doedel collocation in the next Subsection.

For the transition from the quadrature formulation to the collocation, we choose an *interpolation basis* $v_i^h \in \mathcal{V}_b^h$, see (4.49) below, s.t. $v_i^h(P_j) = \delta_{i,j}\ \forall\ P_j \in T\ \forall T\ \in \mathcal{T}^h$. Due to the required unisolvense of $(K,\ \mathcal{P},\ \mathcal{N})$, see (2.31), and by (4.49), this interpolation basis is uniquely determined. This fits to our earlier approach if this $(K,\ \mathcal{P},\ \mathcal{N})$ defines, as in (2.31), the global interpolation strictly locally. Then we we obtain the equivalence to collocation and $u_0^h \in \mathcal{U}_b^h$ is the strong discrete solution of

$$(\ 4.45) \iff \tilde{a}_s^h(u_0^h, v_j^h) - < f, v_j^h \tilde{>}^h = 0\ \forall\ v_j^h \in \mathcal{V}_b^h \iff \tag{4.52}$$
$$(-\Delta\, u_0^h + c u_0^h)(P_j) - < f, v_j^h \tilde{>}^h = 0\ \forall\ P_j \in T\ \forall T\ \in \mathcal{T}^h \iff$$
$$(-\Delta\, u_0^h + c u_0^h - f)(P_j) = 0\ \forall\ P_j \in T \forall T\ \in \mathcal{T}^h \tag{4.53}$$
$$\text{if } < f, v^h > = (f, v^h)_{L^2(\Omega)},$$

hence, (4.53) represents the classical collocation method.

Until now, no other realization is known to satisfy the necessary conditions. The question is totally open, whether there exists unisolvent $(K, \mathcal{P}, \mathcal{N})$, defining, as in (2.31), the global interpolation strictly locally. The set of these methods is not empty if we are willing to modify (4.49) according to (4.50). A very efficient example represent the following Doedel collocations.

4.1.1 Doedel collocations

Doedel collocation methods: Until now, only one class of collocation methods is known, which satisfies (4.49) and (4.50). E.Doedel chooses, in our notation,

$$\begin{aligned} \mathcal{U}^h := \mathcal{V}^h := \{u^h \in FEs : [v^h](P_e^i) = 0 \text{ and} [\frac{\partial u^h}{\partial \nu}] \\ (P_e^i) = 0 \forall e \in \mathcal{T}^h \text{ for sufficiently many points } P_e^i \in \overline{e}\} \\ \mathcal{U}_b^h := \mathcal{V}_b^h := \{u^h \in \mathcal{U}^h = \mathcal{V}^h : v^h(P_e^i) = 0 \forall P_e^i \in \partial\Omega\} \end{aligned} \tag{4.54}$$

Then he collocates according to They work astonishingly well. However they have one essential drawback: They are a kind of "super gangster" and violate all kinds of taboos. Nevertheless, in cooperation with E. Doedel and B. Goldluecke we have been able the proof the existence of a well behaved (non local) interpolation basis for an important special case. We think to have the necessary tools to give all the necessary results for the general case. For general elliptic operators the method in in Section has to be considerably modified. This is the goal of B. Goldlueckes dissertation.

4.2 Finite Element and Spectral Methods

We aim for the classical "stability and consistency yield convergence" approach for our context, see Chapter and, e.g., Stetter [57]. As motivation we want to summarize and generalize the above FEMs and give a short introduction to spectral methods. We choose the corresponding general framework in Sections 4.3 - 4.4, below, to include these and many other cases.
The examples in Sections 3.1, 3.3, 4.1 show the following: To analyze finite elements in particular with variational crimes, we have to relate the approximating subspaces $\mathcal{U}^h, \mathcal{V}^h$ for the original $\mathcal{U}$, $\mathcal{V}$ with linear operators, projectors, linear and bilinear forms and appropriate generalizations. We list several combinations for FEMs and introduce a general notation. To include all the cases treated in Section 4.2 we choose the neutral notations of $\mathcal{U}, \mathcal{V}$, $\mathcal{U}_b, \mathcal{V}_b$ and $\mathcal{U}^h, \mathcal{V}^h$, $\mathcal{U}_b^h, \mathcal{V}_b^h$ for the Banach spaces and their discrete approximations. The $\mathcal{U}_b \subset \mathcal{U}$ and $\mathcal{V}_b \subset \mathcal{V}$ are closed subspaces defined by appropriate boundary conditions, usually with $\mathcal{V}_b' = \mathcal{V}'$. $\mathcal{U}_b$ and $\mathcal{V}_b$ and $\mathcal{U}_b^h$ and $\mathcal{V}_b^h$ are needed to guarantee the unique solvability of the problem $Au_0 = f$. The same holds for the discrete counterparts, e.g., $A^h u_0^h = f^h$ below.

In the following two Subsections on finite element and spectral methods we start with the linear problem $Au = f$ and delay the nonlinear problem to the end of the corresponding Subsections. In Sections 3.1 and 4.1 we had studied the model problem $A_s u = -\Delta u + cu = f$. In this Section we want to generalize the results to the case of general elliptic operators and bilinear forms as presented in Section 3.2. We list the weak and strong forms of operators, bilinear and linear forms and projectors by using no index for the weak form and the index $_s$ for the strong forms and finally introduce the general notation. We delay the study of the relation of weak and strong formulation to later Chapters.
The weak and strong linear operators, A and A_s, and the boundary operator B_a, see (3.15), (3.26), resp., are

$$A : H^1(\Omega) \to H^{-1}(\Omega) \text{ and } a(\cdot,\cdot) : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}, \tag{4.55}$$
$$< Au, v >_{H^{-1}(\Omega)\times H^1(\Omega)} = a(u,v) = \int_\Omega \big(\sum_{i,j=1}^n a_{ij}\partial_i\, u\, \partial_j\, v + \sum_{j=1}^n a_{0j}\, u\, \partial_j\, v + \sum_{i=1}^n a_{i0}(\partial_i\, u)\, v + a_{00} u\, v\big) dx \text{ and}$$
$$A_s : H^2(\Omega) \to L_2(\Omega) \text{ and } a_s(\cdot,\cdot) : H^2(\Omega) \times L^2(\Omega) \to \mathbb{R}, \tag{4.56}$$
$$A_s u := -\sum_{i,j=1}^n \partial_j(a_{ij}\partial_i\, u) - \sum_{j=1}^n \partial_j\, (a_{0j} u) + \sum_{i=1}^n a_{i0}\partial_i\, u + a_{00} u$$
$$a_s(u,v) := (A_s u, v)_{L^2(\Omega)}.$$

Natural (nat.) and Dirichlet (Dir.) boundary conditions are realized, with

$$B_a u := \sum_{i,j=1}^n \nu_j a_{ij}\partial_i\, u + \sum_{j=1}^n \nu_j a_{0j}\, u, \text{ as}$$
$$\begin{aligned}
\text{nat. } \mathcal{U}_b &= \{u \in H^1(\Omega) :\ B_a u|_{\partial\Omega} = 0\}, \text{ with } \mathcal{V}_b = \mathcal{V} = H^1(\Omega),\\
\text{Dir. } \mathcal{U}_b &= \mathcal{V}_b = H_0^1(\Omega) \text{ for the weak, and}\\
\text{nat. } \mathcal{U}_b &= \{u \in H^2(\Omega) :\ B_a u|_{\partial\Omega} = 0\}, \text{ with } \mathcal{V}_b = \mathcal{V} = L^2(\Omega),\\
\text{Dir. } \mathcal{U}_b &= \mathcal{V}_b = H^2(\Omega) \cap H_0^1(\Omega) \text{ for the strong problems.}
\end{aligned} \tag{4.57}$$

More generally we use the form:

$$\text{For } A : \mathcal{U}_b \to \mathcal{V}_b', \text{ determine } u_0 \in \mathcal{U}_b : Au_0 = f \in \mathcal{V}'. \tag{4.58}$$

As in (4.18), (4.19), we may have to extend A or A_s to the piecewise Sobolev spaces, A_h or $A_{s,h}$. Again we use the general notation, see (4.23), (4.26),

$$A_h : (\mathcal{U}_b \cup \mathcal{U}_b^h) \to (\mathcal{V}' \cup \mathcal{V}_b^{h'}) \text{ with } A_h|_{\mathcal{U}_b \to \mathcal{V}_b'} = A. \tag{4.59}$$

4.2.1 Finite Element Methods

In the next collections of formulas we present the linear forms, projectors and bilinear forms as observed in the above methods, here generalized to the

operators A and A_s in (4.55) and (4.56).
We start with modifying the linear forms , see (4.2), (4.3), (4.41), (4.28), (4.39),

$$\begin{aligned}
\langle f, v\rangle_{H^{-1}(\Omega)\times H^1(\Omega)} &\in \mathbb{R}, \quad \forall\ f \in H^{-1}(\Omega),\ \forall\ v \in \mathcal{V} = H^1(\Omega)\\
\langle f^h, v^h\rangle^h_{H_h^{-1}(\Omega)\times H_h^1(\Omega)} &\in \mathbb{R}, \quad \forall\ f \in H_h^{-1}(\Omega),\ \forall\ v^h \in \mathcal{V}^h \subset H_h^1(\Omega),\\
\langle \tilde{f}^h, v^h\tilde{\rangle}^h &\in \mathbb{R}, \quad \forall\ \tilde{f}^h \in C_h^{-1}(\Omega), v^h \in \mathcal{V}_b^h \subset C_h^1(\Omega),\\
(f_s, v)_{L^2(\Omega)} &\in \mathbb{R}, \quad \forall\ f_s \in L^2(\Omega), v^h \in \mathcal{V}_b^h \subset L^2(\Omega)\\
(f_s^h, v^h)_{L^2(\Omega)} &\in \mathbb{R}, \quad \forall\ f_s^h \in L^2(\Omega), v^h \in \mathcal{V}_b^h \subset L^2(\Omega)\\
(\tilde{f}_s^h v^h\tilde{)}^h, &\in \mathbb{R}, \quad \forall\ \tilde{f}_s^h, \in C_h(\Omega), v^h \in \mathcal{V}_b^h \subset C_h(\Omega).
\end{aligned} \tag{4.60}$$

In the general form we denote these cases as

$$\langle f, v\rangle_{\mathcal{V}'\times\mathcal{V}} \in \mathbb{R} \text{ and } \langle f^h, v^h\rangle^h \in \mathbb{R}, \tag{4.61}$$

were the $\langle f^h, v^h\rangle^h$ often is defined only in smooth subspaces $\mathcal{V}_s^{'}, \mathcal{V}_s$ of $\mathcal{V}^{'}$ and $\mathcal{V}$. The different $\langle f, v\rangle_{\mathcal{V}'\times\mathcal{V}}$, $\langle f^h, v^h\rangle^h$ a.s.o., give rise to the corresponding concepts of orthogonality

$$\begin{aligned}
\langle f, v\rangle_{H^{-1}(\Omega)\times H^1(\Omega)} &= 0, \text{ indicated as } f \perp v\\
\langle f^h, v^h\rangle^h_{H_h^{-1}(\Omega)\times H_h^1(\Omega)} &= 0, \text{ indicated as } f^h \perp^h v^h\\
\langle \tilde{f}^h, v^h\tilde{\rangle}^h &= 0, \text{ indicated as } \tilde{f}^h \tilde{\perp}^h v^h\\
(f_s, v)_{L^2(\Omega)} &= 0, \text{ indicated as } f_s \perp_s v\\
(f_s^h, v^h)_{L^2(\Omega)} &= 0, \text{ indicated as } f_s^h \perp_s^h v^h\\
(\tilde{f}_s^h v^h\tilde{)}^h &= 0, \text{ indicated as } \tilde{f}_s^h \tilde{\perp}_s^h v^h
\end{aligned} \tag{4.62, 4.63}$$

Again, more generally we use the notations, see (4.61),

$$f \perp v \text{ for the original and } f^h \perp^h v^h \text{ for the discrete spaces .} \tag{4.64}$$

These linear forms and orthogonality give rise to the definition of projectors, on different spaces, see (4.31), (4.32),

$$\begin{aligned}
&Q^{'h} : H_h^{-1}(\Omega) \to \mathcal{V}_b^{h'} \subset H_h^{-1}(\Omega),\\
&\quad \langle Q^{'h} f - f, v^h\rangle_{H_h^{-1}(\Omega)\times H_h^1(\Omega)} = 0 \ \ \forall\ v^h \in \mathcal{V}_b^h \Leftrightarrow Q^{'h} f - f \perp^h \mathcal{V}_b^h,\\
&\tilde{Q}^{'h} : C_h(\Omega) \to \mathcal{V}_b^{h'} \subset H^{-1}(\Omega),\\
&\quad \langle \tilde{Q}^{'h} f - f, v^h\tilde{\rangle}^h = 0 \ \ \forall\ v^h \in \mathcal{V}_b^h \Leftrightarrow \tilde{Q}^{'h} f - f \tilde{\perp}^h \mathcal{V}_b^h,\\
&Q_s^{'h} : L^2(\Omega) \to \mathcal{V}_b^{h'} \subset L^2(\Omega),\\
&\quad (Q_s^{'h} f - f, v^h)_{L^2(\Omega)} = 0 \ \ \forall\ v^h \in \mathcal{V}_b^h \Leftrightarrow Q_s^{'h} f - f \perp_s \mathcal{V}_b^h,\\
&\tilde{Q}_s^{'h} : C_h(\Omega) \to \mathcal{V}_b^{h'} \subset L^2(\Omega),\\
&\quad (\tilde{Q}_s^{'h} f - f, v^h\tilde{)}^h = 0 \ \ \forall\ v^h \in \mathcal{V}_b^h \Leftrightarrow \tilde{Q}_s^{'h} f - f \tilde{\perp}_s^h \mathcal{V}_b^h,
\end{aligned} \tag{4.65}$$

or more generally, for all the above cases, we define

$$Q'^h \in \mathcal{L}(\mathcal{V}', \mathcal{V}_b^{h\prime}) \text{ by } < Q'^h f - f, v^h >^h = 0 \quad \forall\ v^h \in \mathcal{V}_b^h. \tag{4.66}$$

For FEMs and the real- (or, for spectral methods below, complex -) valued bilinear forms and operators are defined by replacing the $\int_\Omega \nabla u^h \nabla v^h + c u^h v^h dx$ and $-\Delta u^h + cu$ by the general forms $a(u^h, v^h)$ and $A_s u$ in (4.55) and (4.56), resp. Similarly to (4.5), (4.6), (4.16), (4.16), (4.42), (4.45), the (usually real valued) bilinear forms and their discrete analogues are thus:

$$\begin{aligned}
a(u,v) &\quad \text{defined for} \quad \forall\ (u,v) \in H^1(\Omega) \times H^1(\Omega),\\
a^h(u^h,v^h) &\quad \text{defined for} \quad \forall\ (u^h,v^h) \in H_h^1(\Omega) \times H_h^1(\Omega),\\
\tilde{a}^h(u^h,v^h) &\quad \text{defined for} \quad \forall\ (u^h,v^h) \in C_h^1(\Omega) \times C_h^1(\Omega),\\
a_s(u,v) &\quad \text{defined for} \quad \forall\ (u,v) \in H^2(\Omega) \times L^2(\Omega),\\
a_s^h(u^h,v^h) &\quad \text{defined for} \quad \forall\ (u^h,v^h) \in H_h^2(\Omega) \times L^2(\Omega),\\
\tilde{a}^h(u^h,v^h) &\quad \text{defined for} \quad \forall\ (u^h,v^h) \in C_h^2(\Omega) \times C_h(\Omega).
\end{aligned} \tag{4.67}$$

In fact, all these $a(\cdot,\cdot), \ldots, \tilde{a}^h(\cdot,\cdot)$ are defined independent of the bounadry conditions. These come in whenever the weak or strong problems have to be solved. As above, the $\tilde{a}^h(\cdot,\cdot)$ and $\tilde{a}_s^h(\cdot,\cdot)$ are obtained from the $a^h(\cdot,\cdot)$ and $a_s^h(\cdot,\cdot)$ by approximation, e.g., by replacing the exact integrals over the T by quadrature formulas. More generally we formulate

$$a(\cdot,\cdot) : \mathcal{U} \times \mathcal{V} \to \mathbb{R} \text{ and } a^h(\cdot,\cdot) : \mathcal{U}^h \times \mathcal{V}^h \to \mathbb{R}, \tag{4.68}$$

where we again assume that tacidely the $a^h(\cdot,\cdot)$ (and $a_s^h(\cdot,\cdot)$) are to be extended to the original $\mathcal{U}, \mathcal{V}$ or possibly smooth subspaces $\mathcal{U}_s \subseteq \mathcal{U}, \mathcal{V}_s \subseteq \mathcal{V}$. The later are necessary for $\tilde{a}^h(\cdot,\cdot)$ and $\tilde{a}_s^h(\cdot,\cdot)$.

For the general systematic discussion of discretization methods we have to choose appropriate combinations, see below, for the original operators and bilinear forms with their discrete counterparts and projectors. We introduce A_h, $A_{s,h}$, $\tilde{A}_h$, $\tilde{A}_{s,h}$ fully analogous to (4.21), (4.22), (4.23), (4.34), by replacing $-\Delta u + cu$ by Au or $A_s u$. Again, we use the general notation to determine the exact and discrete weak and strong solutions u_0 and u_0^h, resp., see Notation 3.1, by

$$\begin{aligned}
u_0 \in \mathcal{U}_b \text{ s.t. } a(u_0,v) = \langle Au_0, v\rangle_{\mathcal{V}'\times\mathcal{V}} = \langle f, v\rangle_{\mathcal{V}'\times\mathcal{V}} \ \forall\ v \in \mathcal{V}_b \text{ and}\\
u_0^h \in \mathcal{U}_b^h \text{ s.t. } a^h(u_0^h,v^h) = \langle A^h u_0^h, v^h\rangle_{\mathcal{V}'\times\mathcal{V}^h} = \langle f^h, v^h\rangle^h_{\mathcal{V}'\times\mathcal{V}^h} \ \forall\ v^h \subset \mathcal{V}_b^h \text{ or}\\
A^h := Q'^h A_h|_{\mathcal{U}_b^h} \text{ or } A^h = \tilde{Q}'^h \tilde{A}_h|_{\mathcal{U}_b^h}.
\end{aligned} \tag{4.69}$$

Here the A_h and $\tilde{A}_h$ indicate the general notation for the above extended and either slightly modified original A, see (4.59), or an approximate evaluation. Examples are the above quadrature approximations, difference and spectral

approximations by divided differences and Fourier collocation derivatives.

For the above cases in (4.55) - (4.60) the solutions u_0 and u_0^h have to be determined in the following combinations, see (4.2), (4.28), (4.42), (4.3), (4.45), (4.53). We start with the exact and approximate weak formulation: Determine the exact and approximate

$$\begin{aligned}
&\text{solution } u_0 \in \mathcal{U}_b \quad \forall v \in \mathcal{V}_b \quad \text{and} \quad u_0^h \in \mathcal{U}_b^h \ \forall v^h \in \mathcal{V}_b^h \quad \text{by:}\\
&a(u_0, v) - \langle f, v\rangle_{H^{-1}(\Omega)\times H^1(\Omega)} = \langle Au_0 - f, v\rangle_{H^{-1}(\Omega)\times H^1(\Omega)} \text{ and } (4.70)\\
&a(u_0^h, v^h) - \langle f, v^h\rangle_{H_h^{-1}(\Omega)\times H^1(\Omega)} = \langle Au_0^h - f, v^h\rangle_{H^{-1}(\Omega)\times H_h^1(\Omega)}\\
&\quad = \langle Q'^h(Au_0^h - f), v^h\rangle_{H^{-1}(\Omega)\times H^1(\Omega)}, \text{ hence } A^h = Q'^h A|_{\mathcal{U}_b^h} \qquad (4.71)\\
&a^h(u_0^h, v^h) - \langle f, v^h\rangle^h_{H_h^{-1}(\Omega)\times H_h^1(\Omega)} = \langle A^h u_0^h - f, v^h\rangle^h_{H^{-1}(\Omega)\times H_h^1(\Omega)}\\
&\quad = \langle Q'^h(A_h u_0^h - f), v^h\rangle^h_{H_h^{-1}(\Omega)\times H_h^1(\Omega)}, \text{ hence } A^h = Q'^h A_h|_{\mathcal{U}_b^h} \quad (4.72)\\
&\tilde{a}^h(u_0^h, v^h) - \langle f, v^h\tilde{\rangle}^h = \langle Au_0^h - f, v^h\tilde{\rangle}^h\\
&\quad = \langle \tilde{Q}'^h(A_h u_0^h - f), v^h\tilde{\rangle}_{C_h(\Omega)}, \text{ hence } \tilde{A}^h = \tilde{Q}'^h A_h|_{\mathcal{U}_b^h}. \qquad (4.73)
\end{aligned}$$

Furthermore, we have the corresponding strong combinations, where $a, A, f, a^h, A^h, f^h, Q'^h, \tilde{Q}'^h, <\cdot,\cdot>$ have to be replaced by their strong counterparts $a_s, \ldots, \tilde{Q}_s'^h$, $(\cdot,\cdot)$. We only present the equations, e.g., corresponding to (4.70) and (4.73).

$$\begin{aligned}
&a_s(u_0, v) - (f, v)^h_{L^2(\Omega)} = (A_s u_0 - f, v)_{L^2(\Omega)} \ \forall\ v \in \mathcal{V}_b \text{ and}\\
&\tilde{a}_s^h(u_0^h, v^h) - (f, v^h\tilde{)} = (\tilde{A}_s^h u_0^h - f, v^h\tilde{)}^h = (\tilde{Q}_s'^h(A_s u_0^h - f), v^h\tilde{)}^h\\
&\qquad\qquad \forall\ v^h \in \mathcal{V}_b^h, \text{ hence } \tilde{A}_s^h = \tilde{Q}_s'^h A_s|_{\mathcal{U}_b^h}. \qquad (4.74)
\end{aligned}$$

In our general notation we determine u_0 and u_0^h by

$$\begin{aligned}
&u_0 \in \mathcal{U}_b : a(u_0, v) = <f>_{\mathcal{V}'\times\mathcal{V}} \quad \forall\ v \in \mathcal{V}_b, \text{ so } Au_0 = f \text{ and}\\
&u_0^h \in \mathcal{U}_b^h : a^h(u_0^h, v^h) = <f^h, v^h>^h_{\mathcal{V}'\times\mathcal{V}} \text{ or } <f^h, v^h>^h \qquad (4.75)\\
&\qquad\qquad \forall\ v^h \in \mathcal{V}_b^h \text{ or } A^h u_0^h = f^h.
\end{aligned}$$

It is important to realize that some of the above modified operators, bilinear and linear forms coincide with or approximate, for smooth enough $u \in \mathcal{U}_s, v \in \mathcal{V}_s$, the original forms. So we have, compare (3.6),

$$\begin{aligned}
&a^h(u, v) = a(u, v),\ a_s^h(u, v) = a_s(u, v) \quad \forall u \in \mathcal{U}, v \in \mathcal{V},\\
&\tilde{a}^h(u, v) \approx a(u, v),\ \tilde{a}_s^h(u, v) \approx a_s(u, v) \quad \forall u \in \mathcal{U}_s, v \in \mathcal{V}_s,\\
&\langle f, v\rangle^h_{\mathcal{V}'\times\mathcal{V}} = \langle f, v\rangle_{\mathcal{V}'\times\mathcal{V}}, \qquad (4.76)\\
&\langle f, v\tilde{\rangle}^h_{\mathcal{V}'\times\mathcal{V}} \approx \langle f, v\rangle_{\mathcal{V}'\times\mathcal{V}} \quad \forall\ f \in \mathcal{V}_s', v \in \mathcal{V}_s.
\end{aligned}$$

Recall, that Q'^h is either defined w.r.t. the original weak or strong scalar product or its piecewise definition, see (2.5), (2.17) for the corresponding

norms. It requires testing with $v^h \in \mathcal{V}_b^h \not\subset \mathcal{V}_b$. This might cause some problems for $Au,\ A_s u \in \mathcal{V}_b'$. Furthermore, if we want to use $\tilde{Q}'^h$ for FEs based upon quadrature approximations, the terms $A^h u^h - f$ and v^h, a.s.o. have to be smooth enough to allow point evaluations. By choosing a strong form of $Au = f$ and an interpolation basis

$$\mathcal{V}_b^h = \text{ span } \{v_{i,T} \in \mathcal{V}_b^h, v_{i,T}(P_j) = \delta_{i,j},\ \forall P_j \in T,\ \forall T \in \mathcal{T}^h\},$$

this method can often be re-interpreted as collocation method, see Sections 4.1, 4.2.2, 6.5, 6.6.

To present a general form for FE and spectral methods, we introduce a unifying notation for linear, bilinear forms, linear operators and projectors and their discrete counterparts.

Notation 4.1 *We collect the above* unifying notations*: We use the* uniform notation *from (4.61), (4.68) for the linear and bilinear forms* $f(\cdot),\ f^h(\cdot)$ *and* $a(\cdot,\cdot),\ a^h(\cdot,\cdot)$ *. For the projectors and the exact, approximate and extended linear operators* $Q'^h : \mathcal{V}_b' \to \mathcal{V}_b^{h'},\ A : \mathcal{U}_b \to \mathcal{V}_b',\ A_h : \{\mathcal{U}_b \cup \mathcal{U}_b^h\} \to \{\mathcal{V}_b \cup \mathcal{V}_b^h\},\ A^h : \mathcal{U}_b^h \to \mathcal{V}_b^{h'}$ *we have, for fixed* $u^h \in \mathcal{U}_b^h$,

$$\begin{aligned}
A &: \mathcal{U}_b \to \mathcal{V}_b',\ \langle Au, v\rangle_{\mathcal{V}_b' \times \mathcal{V}_b} = a(u,v)\ \ \forall\ v \in \mathcal{V}_b,\\
A_h &: \{\mathcal{U}_b \cup \mathcal{U}_b^h\} \to \{\mathcal{V}_b \cup \mathcal{V}_b^h\},\\
&\quad \langle A_h u^h, v^h\rangle_{\{\mathcal{V}_b' \cup \mathcal{V}_b'^h\} \times \{\mathcal{V}_b \cup \mathcal{V}_b^h\}} = a^h(u^h, v^h)\ \ \forall\ v^h \in \{\mathcal{V}_b \cup \mathcal{V}_b^h\},\ \textit{and}\\
A^h &= Q'^h A_h|_{\mathcal{U}_b^h},\ \textit{with}\ \langle A^h u^h, v^h\rangle^h = a^h(u^h, v^h)\ \ \forall\ v^h \in \mathcal{V}_b^h,
\end{aligned} \tag{4.77}$$

for $A^h u := Q'^h Au$ *see the remarks following (4.26). Whenever we want to stress quadrature or the strong form of* A^h *or its corresponding strong or quadrature bilinear counterpart,* $a_s(\cdot,\cdot),\ a_s^h(\cdot,\cdot)$ *or* $\tilde{a}^h(\cdot,\cdot)$, *or* $\tilde{a}_s^h(\cdot,\cdot)$, *see (4.45), we still use the notations*

$$\begin{aligned}
< \tilde{A}^h u^h, v^h \tilde{>}^h &:= \tilde{a}^h(u^h, v^h)\ \textit{and}\\
(A_s^h u^h, v^h)_{L^2(\Omega)} &:= a_s^h(u^h, v^h)\ \textit{and}\\
(\tilde{A}_s^h u^h, v^h\tilde{)}^h &:= \tilde{a}_s^h(u^h, v^h)\ \forall\ \ v^h \in \mathcal{V}_b^h.
\end{aligned} \tag{4.78}$$

Then, see (4.69), we use a unified neutral notation for the exact and approximate (weak or strong) solutions u_0 *and* u_0^h. *They are defined by the unified equations:*

$$\begin{aligned}
a(u_0, v) &= f(v)\ \forall\ v \in \mathcal{V}_b\ \textit{or, for short,}\ \ Au_0 = f,\ \textit{and}\\
a^h(u_0^h, v^h) &= f^h(v^h)\ \forall\ v^h \in \mathcal{V}_b^h\ \textit{or, for short,}\ \ A^h u_0^h = f^h.
\end{aligned} \tag{4.79}$$

The different forms of explicit notations are collected in (4.70) - (4.75).

We extend this approach to the nonlinear case. This is more or less straight forward for the strong versions. For weak formulations, the nonlinear terms usually are discussed by introducing multi linear forms. This is well established for the Navier-Stokes equations, see [64, 16]. We do not want to go into the details here. Rather, we present our standard nonlinear example, $G : \mathcal{U}_b \to \mathcal{V}'$. Again, we have to distinguish the weak and strong formulations, e.g.,

$$\begin{aligned} G_s &: H^2(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega), \\ G_s(u) &:= -\Delta u + f(u) = g = 0 \text{ in } \Omega,\ u|_{\partial\Omega} = 0 \end{aligned} \tag{4.80}$$

or the corresponding weak formulation induced by the usual bilinear and higher nonlinear forms as

$$\begin{aligned} &G : H_0^1(\Omega) \to H^{-1}(\Omega), \quad < Gu, v >_{H^{-1}(\Omega)\times H^1(\Omega)} := \\ &a(u,v) + < f(u) - g, v >_{H^{-1}(\Omega)\times H^1(\Omega)} \quad \forall v \in H_0^1(\Omega) \end{aligned} \tag{4.81}$$

where $a(u,v)$ is defined as in (4.67). E.g., for the cases $f(u) = u^2$ and $f(u) = u^3$ the $< f(u), v >_{H^{-1}(\Omega)\times H^1(\Omega)}$ gives rise to a tri linear and quadri linear form $< u_1 \times u_2, v >_{H^{-1}(\Omega)\times H^1(\Omega)}$ and $< u_1 \times u_2 \times u_3, v >_{H^{-1}(\Omega)\times H^1(\Omega)}$, resp. The main problem is to show that indeed, e.g., $< u_1 \times u_2 \times u_3, v >_{H^{-1}(\Omega)\times H^1(\Omega)}$, is well defined, see [64, 16] for the case of the Navier-Stokes equations.. As in (4.58) we use the notation for these and more general cases

$$G : \mathcal{U}_b \to \mathcal{V}_b',\ \text{determine } u_0 \in \mathcal{U}_b : G(u_0) = g \in \mathcal{V}'. \tag{4.82}$$

Similarly to the above piecewise and quadrature variants in (4.67), (4.69) we have to consider the weak and strong form of G and possible approximations $\tilde{G}_h$, usually again depending upon h. We formulate the discrete operators as

$$\begin{aligned} G^h u^h &:= Q^{'h}(G(u^h)) \text{ or } G^h u^h := Q^{'h}\tilde{G}_h(u^h)) \text{ or} \\ &G^h u^h := \tilde{Q}^{'h}(G_h)(u^h)) \text{ or } G^h u^h := \tilde{Q}^{'h}(\tilde{G}_h(u^h)) \text{ and} \end{aligned} \tag{4.83}$$

$$\begin{aligned} G_s^h u^h &:= Q_s^{'h}(G_s(u^h)) \text{ or } G^h u^h := Q_s^{'h}\tilde{G}_h(u^h)) \text{ or} \\ &G^h u^h := \tilde{Q}_s^{'h}(G_s)(u^h)) \text{ or } G^h u^h := \tilde{Q}_{s,h}^{'h}(\tilde{G}_s(u^h)). \end{aligned} \tag{4.84}$$

As in (4.69), we formulate the general exact and discrete operators as

$$\begin{aligned} G : \mathcal{U}_b \to \mathcal{V}',\ \text{and } G^h : \mathcal{U}_b^h \to \mathcal{V}', G^h := Q^{'h}G \text{ or} \\ G^h := Q^h\tilde{G}_h \text{ or } G^h := \tilde{Q}^h\tilde{G}_h. \end{aligned} \tag{4.85}$$

Finally, we compute, the exact and the discrete solutions u_0 and u_0^h from

$$\begin{aligned} G : \mathcal{U} \to \mathcal{V}';\ \text{determine } u_0 \in \mathcal{U}_b \text{ s.t. } G(u_0) = 0 \Leftrightarrow G(u_0) \perp \mathcal{V}_b, \\ G^h : \mathcal{U}^h \to \mathcal{V}^{h'};\ \text{determine } u_0^h \in \mathcal{U}_b^h \text{ s.t. } G^h(u_0^h) = 0 \Leftrightarrow G^h(u_0^h) \perp^h \mathcal{V}_b^h. \end{aligned} \tag{4.86}$$

For a more detailled formulation of the nonlinearity we refer to Taylor for the analytical background, [62], and Caloz/Rappaz for the finite element version, [22].

4.2.2 Spectral Methods

They, too, determine the approximate solutions via the variational approach in Chapter 3 and Section 4.1. The approximating spaces are defined differently. Usually, the only type fo variational crime to be considered are quadrature approximations. They are equivalent to collocation methods. Spectral methods are well presented in the classical and recent books and surveys [31, 23, 5, ?], for special results see [11, 12, ?]. These methods are particularly appropriate for Γ- equi-variant problems with continuous groups Γ. This equi-variance has to be reproduced in the discretization as well. We restrict the discussion to Hilbert spaces $\mathcal{U} = H^m_w(\Omega), \mathcal{V} = L^2_w(\Omega) = \mathcal{V}', \Omega \subset \mathbb{R}^n$; here w denotes a weight function in a scalar product, $(\cdot,\cdot)_w$, and $\|\cdot\|_{H^k_w(\Omega)}$ the corresponding weighted Sobolev norms, used in the rest of this Section. Hence, we have, e.g.

$$(u,v)_w := \int_\Omega u\ v\ w\ dx \quad \text{and } \|u\|^2_{H^k_w(\Omega)} = \sum_{|\alpha|\le k} (D^\alpha u, D^\alpha u)_w.$$

This Hilbert space setting is appropriate, since it is the standard setting for spectral methods and is needed for the most important collocation version. This requires even $G(u) \in C(\Omega)$, hence more than the usual $G(u) \in \mathcal{V}' = H^{-1}(\Omega)$ or $\mathcal{V}' = L^2(\Omega)$.

We assume, for $\mathbf{Z}^n_0 \subset \mathbf{Z}^n$, a

$$\text{complete orthogonal basis } \{\varphi_k\}_{k\in\mathbf{Z}^n_0} \text{ for } \mathcal{U} \text{ and } \mathcal{V} \text{ w.r.t. } \|\cdot\|_{\mathcal{U}} \text{ and } \|\cdot\|_{\mathcal{V}}, \text{ resp.,} \tag{4.87}$$

with real- or complex valued $\varphi_k(x)$. The finite dimensional approximating spaces are

$$\mathcal{U}^N = \mathcal{U}^h = \text{ span}\{\varphi_k :\ k = (k_1,\dots,k_n) \in \mathbf{K}^N\} \subset \mathcal{U}, \quad \text{finite } \mathbf{K}^N \subset \mathbf{Z}^n_0, \tag{4.88}$$

here $u^h \in \mathcal{U}^h$ and the discretization index h is defined via the range of indices k. The most important examples for $\mathcal{U}^N$ are trigonometric (Fourier) and Legendre or Chebyshev polynomials. With $N = (N_1,\dots,N_n) \in \mathbb{N}^n_0$ this multi-index $k \in \mathbf{K}^N$ satisfies, e.g., $|k_i| \le N_i$ and let $\hat{N} := |\mathbf{K}^N|, \tilde{N} := \min\{N_1,\dots,N_n\}$. Corresponding operators of truncation, interpolation and orthogonal projection, and their approximations, onto the $\mathcal{U}^h_b, \mathcal{V}^h_b$, are denoted by $T^h, I^h, P^h, Q^{'h}$, and $\tilde{P}^h, \tilde{Q}^{'h}$, resp. Every $u \in \mathcal{U}$ (or $\mathcal{V}$) is, alternatively, *approximated by truncation* $T^h u$ as

$$T^h u = T^h(\sum_{k\in\mathbf{Z}^n_0} \hat{a}_k\varphi_k) := \sum_{k\in\mathbf{K}^N} \hat{a}_k\varphi_k \in \mathcal{U}^h,$$

or by (unique) *interpolation* in distinct points $y_j \in \Omega, j$ multi-indices, as

$$I^h : \mathcal{U} \to \mathcal{U}^h \text{ is uniquely defined by}$$

$$(I^h u - u)|_{y_j} = 0, j \in \mathbf{J}^N \subset \mathbf{Z}^n, |\mathbf{J}^N| = \hat{N} = |\mathbf{K}^N|.$$

Now, $(\cdot,\cdot)_w$ is approximated for trigonometric and Legendre or Chebyshev polynomials by the iterated trapezoidal (with equidistant points) and the Gaussian quadrature formulas, resp., defined as

$$(u,v)_w \approx (u,v)_w^{\tilde{h}} := \sum_{j \in \mathbf{J}^N} u(y_j)\bar{v}(y_j)w_j, \quad (\|u\|^h_{L^2_w(\Omega)})^2 := (u,u)_w^{\tilde{h}}. \tag{4.89}$$

With $\rho = -1, 0, 1, 2$ for equidistant, Gauss, Gauss-Radau and Gauss-Lobatto quadrature points y_j we have $=$ instead of $\approx$ in (4.89) for $\rho = -1, 0, 1$ and $u, v \in \mathcal{U}^h = \mathcal{U}^N$. For $\rho = 2$ this is only valid if one of the u or $v \in \mathcal{U}^{N-1} \subset \mathcal{U}^N = \mathcal{U}^h$. The corresponding truncation and interpolation errors satisfy

$$T^h u - u \perp_w \mathcal{U}^h \text{ and } I^h u - u \perp^h_w \mathcal{U}^h \; \forall \; u \; \in \mathcal{U}, \mathcal{V}. \tag{4.90}$$

As mentioned already, the $\mathcal{U}^h$, $\mathcal{V}^h$, φ_k, y_j often are chosen according to a group Γ, e.g., $\Gamma = SO(1)$ and $\varphi_k(x) = \exp(ikx)$. If we approximate periodic problems with periodic boundary conditions by the corresponding periodic functions, see [13], then we have $\mathcal{U}^h_b = \mathcal{U}^h$, $\mathcal{V}^h_b = \mathcal{V}^h$. Since (4.90) corresponds to the $\perp_s$ in Subsection 4.2.1 (, we did not use a $\perp_w$ to indicate a weak $\perp$!) this reads as

$$\mathcal{U}^h_b = \mathcal{U}^h, \mathcal{V}^h_b = \mathcal{V}^h \quad T^h = Q_s^{'h}, I^h = \tilde{Q}_s^{'h}. \tag{4.91}$$

These $\mathcal{U}_b, \mathcal{V}_b$, $\mathcal{U}^h_b, \mathcal{V}^h_b$ and the $(u,v)_w, (u,v)_w^{\tilde{h}}$ are Γ-invariant. The $T^h, I^h : \mathcal{U} \to \mathcal{U}^h$, coincide with Γ-equi-variant and orthogonal projectors w.r.t. $(\cdot,\cdot)_w$ and $(\cdot,\cdot)_w^{\tilde{h}}$, resp. Hence, e.g.,

$$\begin{aligned}
&(\gamma u, \gamma v)_w = (u,v)_w, \; (\gamma u, \gamma v)^h_w = (u,v)_w^{\tilde{h}} \; \forall \; u,v \; \in \; \mathcal{U}, \; \mathcal{V}, \; \gamma \; \in \Gamma, \\
&(Q_s^{'h} f - f, v^h)_w = 0 \text{ and } \quad (\tilde{Q}_s^{'h} f - f, v^h)_w^{\tilde{h}} = 0 \quad \forall \quad v^h \in \mathcal{V}^h_b \text{ with} \\
&T^h \; = \; Q_s^{'h}, I^h \; = \; \tilde{Q}_s^{'h}, \text{ s.t. } Q_s^{'h} f - f \perp_w \mathcal{V}^h_b \text{ and } \tilde{Q}_s^{'h} f - f \perp^h_w \mathcal{V}^h_b \text{ and} \\
&Q_s^{'h} \gamma f = \gamma Q_s^{'h} f \text{ and } \tilde{Q}_s^{'h} \gamma f = \gamma \tilde{Q}_s^{'h} f \; \forall \; \gamma \; \in \Gamma.
\end{aligned} \tag{4.92}$$

The situation changes for non periodic problems and in case of domain decomposition techniques. Then the boundary conditions defining $\mathcal{U}^h_b, \mathcal{V}^h_b$ are realized via collocation in appropriate boundary points. For the standard spectral and their standard domain decomposition methods enough boundary points are chosen, s.t., e.g., for Dirichlet boundary conditions, $u^h(y_j) = 0 \; \forall \; j \; \in \; \mathbf{J}^N, y_j \in \partial\Omega \; \Leftrightarrow u^h|_{\partial\Omega} \; \equiv \; 0$. So, as mentioned above already, variational crimes for the standard spectral methods are solely due to quadrature approximations. For the one-dimensional and higher-dimensional errors we obtain, see [31, 23, 5, 65, 11, 12, ?] for more detailed errors

Theorem 4.2.1. *Let $0 \le \ell \le m$ and $\Omega \subset \mathbb{R}$, hence $n = 1$. For a periodic function $u \in H^m_w(\Omega)$ choose $\Omega = (0, 2\pi)$, $w(x) \equiv 1$ and Fourier polynomials $(K = F)$. For non-periodic functions choose $\Omega = (-1, 1)$, $w(x) = 1/\sqrt{1-x^2}$ and $w(x) \equiv 1$ for Chebyshev and Legendre polynomials $(K = C$ and $L)$, respectively. Furthermore, let $m > \iota_K(\ell)$ with*

$$\iota_F(\ell) := \ell, \ \iota_C(\ell) := 2\ell, \ \iota_L(\ell) := 2\ell + n/2. \tag{4.93}$$

Then the errors converge for $N \to \infty$ as

$$\|T^h_K u - u\|_{H^\ell_w(\Omega)} \ \le \ \| I^h_K u - u \|_{H^\ell_w(\Omega)} \ \le C N^{-m+\iota_K(\ell)} \| u \|_{H^m_w(\Omega)} \tag{4.94}$$

The quadrature errors are estimated by $\| I^h_K u - u \|_{L^2_w(\Omega)}$, however see (4.89) and below. For functions defined on $\Omega \subset \mathbb{R}^n, n > 1$, the N in (4.94) has to be replaced by $\tilde{N}$. If different approximations, e.g., a combination of Fourier and Legendre, are used for different variables, the minimal $\iota^K(\ell)$ has to be chosen.

For the study of variational crimes we need, [23], for $u^h \in U^h, 1 \le p \le q \le \infty$ and $r \ge 1$ the following inverse estimates

$$\|u^h\|_{w^r_q(\Omega)} \le C N^{\mu_K(r)+\nu_K(\frac{1}{p}-\frac{1}{q})} \|u^h\|_{L^p(\Omega)} \quad \forall u \in \mathcal{U}^h, k = F, C, L \text{ and} \tag{4.95}$$
$$\mu_F(r) = r, \mu_C(r) = \nu_L(r) = 2r \text{ and } \nu_F = \nu_C = 1, \nu_L = 2$$

To avoid too many technicalities for the following spectral collocation methods, see Subsecion 4.2.2, we assume (4.112) as

$$\begin{aligned} G(u) &= Au + \lambda R(u) = Au + \lambda R_e(u, \nabla u, \int_{\Omega_0} u) \\ &= Au + \lambda(u^2 + \nabla u(u + \int_{\Omega_0} u) + g). \end{aligned} \tag{4.96}$$

with $G(u_0) = 0$, $A = G_u(u_0)$ a bounded linear operator, often $Au = -\Delta u$, and R a nonlinear operator. This (4.96) already shows the handling of the essential difficulties. For the general case, see [13]. With the above projectors Q'^h_s, $\tilde{Q}'^h_s$ the different types of spectral methods can be formulated as, see (4.90),

$$\text{determine } u^h_0 \in \mathcal{U}^h_b \text{ such that } \ Q'^h_s G(u^h_0) = 0 \text{ or } G(u^h_0) \perp_w \mathcal{V}^h_b, \tag{4.97}$$
$$\text{determine } u^h_0 \in \mathcal{U}^h_b \text{ such that } \ \tilde{Q}'^h_s G(u^h_0) = 0 \text{ or } G(u^h) \perp^h_w \mathcal{V}^h_b. \tag{4.98}$$

Sometimes $G(u^h)$ is replaced by an approximate operator $\tilde{G}^h(u^h)$, see [6], **Klaus Klaus im book wieder aendern Booklet ??** Subsubsecion **??**. The difference to FEMs is due to the fact, that mostly the strong formulations with $Q'^h_s, \tilde{Q}'^h_s$ are used as in (4.97), (4.98), w.r.t. to $L^2_w(\Omega) = \mathcal{V} = \mathcal{V}'$. The weak form with, e.g., $\mathcal{V}' = H^{-1}(\Omega)$ and the $Q'^h, \tilde{Q}'^h$ play a minor role. This

is mainly caused by the dominance of collocation spectral methods. To apply it we need continuous $G(u)$ or $G(u^h)$. For our above example (4.96) we have to evaluate

$$\begin{aligned} G^h(u^h)(y_j) &:= (A^h u^h)(y_j) + \lambda R^h(u^h)(y_j),\ A^h u^h := T^h\ (Au^h) \qquad (4.99)\\ &\text{with } j \in \mathbf{J}^N \text{ and usually } T^h(Au^h) = Au^h, \text{ and}\\ R^h(u^h)(y_j) &:= \lambda((u^h)^2(y_j) + (\nabla u^h)(y_j)(u^h(y_j) + \sum_{i\in\mathbf{J}^N} u^h(y_i)w_i) + g(y_j)). \end{aligned}$$

In spectral methods usually the $(A(u^h))(y_j), (\nabla u^h)(y_j)$ are not evaluated directly, but via some linear approximation operators, e.g., the Fourier collocation derivative, see [23]. We denote these kinds of *linear* operators (again) as

$$\begin{array}{ccccccc} A^h u^h & \approx & (A)u^h, & & L_1^h u^h & \approx & \nabla u^h \quad \text{s.t.}\\ (A^h u^h)(y_j) & \approx & (Au^h)(y_j), & & (L_1^h u^h)(y_j) & \approx & (\nabla u^h)(y_j)\\ \ell u^h & = & \int_{\Omega_0} u^h & \approx & \sum_{i\in\mathbf{J}^N} u^h(y_i)w_i & =: & \ell^h u^h. \end{array} \qquad (4.100)$$

We introduce the restriction operator

$$\rho^h : C(\Omega) \to \mathbb{R}^{\hat{N}}, (\rho^h(u))(y_j) := u(y_j),\quad j \in \mathbf{J}^N. \qquad (4.101)$$

The $G^h(u^h)$ in (4.99) can now be re-interpreted as

$$\rho^h G^h(u^h) = \rho^h(A^h u^h) + \lambda \rho^h R^h(u^h). \qquad (4.102)$$

For spectral methods in the Hilbert space $L^2_w(\Omega)$ the ρ^h can, for smooth situations, equivalently be defined by the $\tilde{Q}_s^{'h}$ in (4.92). So we can re-write (4.102) as

$$\begin{aligned} &\tilde{Q}_s^{'h} G^h(u^h) = \tilde{Q}_s^{'h}(A^h u^h) + \lambda \tilde{Q}_s^{'h} R^h(u^h),\\ &\quad \text{and determine } u_0^h \text{ s.t. } \tilde{Q}_s^{'h} G^h(u_0^h) = 0. \end{aligned} \qquad (4.103)$$

So we are back at (4.98). We distinguish two cases: Either the restriction operator ρ^h is applied to the exactly evaluated $R(u^h)$ to obtain $\rho^h R(u^h) = \rho^h R^h(u^h)$. Or, e.g., Fourier collocation differentiation and de-aliasing techniques are employed to introduce an appropriate R_e. Then, we observe that

$$\begin{aligned} \rho^h R^h(u^h) &:= R_e(\rho^h u^h, \rho^h(L_1 u^h), \ell^h u^h) \qquad (4.104)\\ &= \rho^h R_e(u^h, \nabla u^h, \ell^h u^h) + \mathcal{O}(\|I^h R(u^h) - R(u^h)\|_{L^2_w(\Omega)})\\ &= \rho^h R(u^h) + \mathcal{O}(\|I^h R(u^h) - R(u^h)\|_{L^2_w(\Omega)}),\\ &\quad \text{and corresponding relations for the partials of } R, R^h, \end{aligned}$$

All these operators ρ^h, A^h, L_1^h, and the functional ℓ^h are bounded and linear, w.r.t. appropriate norms. So the corresponding conditions for the derivatives

are automatically satisfied. So, we have a situation very similar to FEMs. The main difference is the other form of approximation subspaces and the concentration on the strong formulations. Instead of (4.83) the following strong version defines u_0^h :

$$G_s^h u_0^h := Q_s^{'h}(G_s(u_0^h)) \ = 0 \text{ or } G_s^h u_0^h := Q_s^{'h}\tilde{G}(u_0^h)) \ = 0 \ \text{ or}$$
$$G_s^h u_0^h := \tilde{Q}_s^{'h}(G_s)(u_0^h)) \ = 0 \text{ or } G_s^h u_0^h := \tilde{Q}_s^{'h}(\tilde{G}_s(u_0^h)) \ = 0. \quad (4.105)$$

This corresponds either to the (4.97), (4.98) or to the collocation version (4.99), however generalized to allow the above approximations.

4.3 General Concepts for Convergence of Finite Elements

In Sections 3.1, 4.1, 4.2, we have collected many examples of FE and spectral methods with and without variational crimes. This shows the following: To analyze these methods, in particular with variational crimes, we have to relate the approximating subspaces $\mathcal{U}^h, \mathcal{V}^h$ with linear operators, projectors, linear and bilinear forms and appropriate generalizations. To simultaneously discuss all the above cases, we choose the neutral notations of $\mathcal{U}, \mathcal{V}$ and $\mathcal{U}^h, \mathcal{V}^h$ for the Banach spaces and their discrete approximations, resp. The $\mathcal{U}_b \subset \mathcal{U}$, $\mathcal{V}_b \subset \mathcal{V}$ and $\mathcal{U}_b^h \subset \mathcal{U}^h$, $\mathcal{V}_b^h \subset \mathcal{V}^h$ are closed subspaces defined by appropriate boundary conditions, usually with $\mathcal{V}_b' = \mathcal{V}'$. Only sometimes we require $\mathcal{U}_b^h \subset \mathcal{U}_b, \mathcal{V}_b^h \subset \mathcal{V}_b$ or$\mathcal{U}_b^h \subset \mathcal{U}, \mathcal{V}_b^h \subset \mathcal{V}$. $\mathcal{U}_b$ and $\mathcal{V}_b$ and $\mathcal{U}_b^h$ and $\mathcal{V}_b^h$ are needed to guarantee the unique solvability of the problem $Au_0 = f$ and its discrete counterpart $A^h u_0^h = f^h$ below. We use norms and bilinear forms either with or without indices, e.g. $\|u\| = \|u\|_{\mathcal{U}}$ for $u \in \mathcal{U}$ or $u \in \mathcal{U}_b$. We give several definitions:

Definition 4.3.1. Conforming and non conforming (Petrov-Galerkin) approximating spaces: *Let $\mathcal{U}, \mathcal{V}$ be Banach spaces, $\mathcal{U}_b \subset \mathcal{U}$, $\mathcal{V}_b \subset \mathcal{V}$ closed subspaces and $\mathcal{U}^h, \mathcal{V}^h$ be finite-dimensional spaces, assigned to every $h \in H$ with* [2] *accumulation point* 0. *We call $\mathcal{U}^h$* approximating spaces for $\mathcal{U}$ *if we*

$$dist(u, \mathcal{U}^h) := \inf_{u^h \in \mathcal{U}^h} ||u - u^h||_{\mathcal{U}}^h \to 0 \quad \text{for } h \to 0 \ \forall \quad u \in \mathcal{U}, \quad (4.106)$$

sometimes only required for $u \in \mathcal{U}_b$. In the sequel we assume admissible approximating spaces $\mathcal{U}_b^h \subset \mathcal{U}^h$, $\mathcal{V}_b^h \subset \mathcal{V}^h$ *for* $\mathcal{U}_b$ **and** $\mathcal{V}_b$ *defined by*

$$dim\, \mathcal{U}_b^h = dim\, \mathcal{V}_b^h, \ dist\,(u, \mathcal{U}_b^h) \ \to 0 \ , dist\,(v, \mathcal{V}_b^h) \to 0 \text{ for}$$
$$h \to 0 \quad \forall\, u \in \mathcal{U}_b, \ \forall\, v \in \mathcal{V}_b. \quad (4.107)$$

[2] we do not want to over-formalize the notation and have chosen $h \in H$

The terms conforming *and* Galerkin *and* Petrov-Galerkin approximating spaces *are used if* $\mathcal{U}_b^h \subset \mathcal{U}_b$, $\mathcal{V}_b^h \subset \mathcal{V}_b$ *and* $\mathcal{U}_b^h = \mathcal{V}_b^h$ *and* $\mathcal{U}_b^h \neq \mathcal{V}_b^h$, *resp.*
We denote a combination $\mathcal{U}_b^h \not\subset \mathcal{U}_b$, $\mathcal{V}_b^h \not\subset \mathcal{V}_b$ *or* $\mathcal{U}_b^h \not\subset \mathcal{U}$, $\mathcal{V}_b^h \not\subset \mathcal{V}$ *as* non conforming (Petrov-Galerkin) approximating spaces. *Sometimes, all these cases are denoted as* generalized Petrov-Galerkin approximating spaces.
If, in addition to (4.106) for $\mathcal{U}_b$, *the* $\mathcal{V}^h$ *satisfy (4.106)* $\forall \quad v \in \mathcal{V}_b$ **and** $\forall \quad v \in \mathcal{V}''$, *the bi-dual space of* $\mathcal{V}$, *then* $\mathcal{U}^h$, $\mathcal{V}^h$ *is called* a bi-dual pair of non conforming (Petrov-Galerkin) approximating spaces. *The bi-duality condition is usually satisfied, in particular, it is correct for our FE and spectral approximations.*

For the following considerations we need operators $P^h \in \mathcal{L}(\mathcal{U}, \mathcal{U}_b^h)$, and $Q^{'h} \in \mathcal{L}(\mathcal{V}', \mathcal{V}_b^{h'})$ sometimes even $Q^h \in \mathcal{L}(\mathcal{V}, \mathcal{V}_b^h)$. The Q^h is needed for the dual problem and is then defined analogously to $Q^{'h}$. We may choose the P^h relatively arbitrarily, however, the $Q^{'h}$ has to be chosen to fit to the variational methods, see (4.109), Remark 4.3.2. So let

$$P^h \in \mathcal{L}(\mathcal{U}, \mathcal{U}_b^h), \ e.g., \ P^h u := I^h u \quad \forall u \in \mathcal{U}, \tag{4.108}$$

with the interpolation operator for finite element or spectral methods or some kind of quasi-interpolation or best approximation operator.

We have seen, in (4.60), that for the non conforming Petrov-Galerkin spaces, the original $< f, v^h >_{H^{-1}(\Omega)\times H^1(\Omega)}$ may not be defined, and have to be extended as $< f, v^h >^h_{H_h^{-1}(\Omega)\times H_h^1(\Omega)} = < f, v^h >^h$ in the generalized form. Or we even have used $a < \tilde{f}^h v^h \tilde{>}^h$. They still coincide with or approximate the original pairings on $\mathcal{V}' \times \mathcal{V}_b$. We use for all cases the general notations ,see (4.61), $< f^h, v >_{\mathcal{V}'\times\mathcal{V}_b}$ and $< f^h, v^h >^h$. On this basis and including all the cases in (4.65), we are able to define interpolation and the operators $Q^{'h}$ as

$$\begin{aligned} Q^{'h} \in \mathcal{L}(\mathcal{V}', \mathcal{V}_b^{h'}) \text{ s.t. } Q^{'h} f := f|_{\mathcal{V}_b^h} \ \Leftrightarrow \ < Q^{'h} f - f, v^h >^h = 0 \text{ or} \\ \tilde{Q}^{'h} f := \tilde{f}^h|_{\mathcal{V}_b^h} \ \Leftrightarrow \ < \tilde{Q}^{'h} f - f, v^h \tilde{>}^h = 0 \quad \forall \, v^h \in \mathcal{V}_b^h \end{aligned} \tag{4.109}$$

According to (4.65) we use the same notation, $Q^{'h}$ for all cases, including $\tilde{Q}^{'h}$.

Our (stationary) nonlinear problems can be formulated again in a strong or a weak form. We study the operator $G : H_0^1(\Omega) \to H^{-1}(\Omega)$ and determine $u_0 \in H_0^1(\Omega)$ s.t. $G(u_0) = 0$. Similarly to Section 4.1,

$$\begin{aligned} &G(u_0) = 0 \text{ is tested with } \mathcal{V}_b \text{ w.r.t. } < \cdot, \cdot >_{H^{-1}(\Omega)\times H^1(\Omega)} . \text{ Then} \\ &G(u_0) = 0 \ \Leftrightarrow \ < G(u_0), v >_{H^{-1}(\Omega)\times H^1(\Omega)} = 0 \ \forall \quad v \in \mathcal{V}_b \\ &\qquad \Leftrightarrow G(u_0) \perp \mathcal{V}_b. \end{aligned} \tag{4.110}$$

If we want to emphasize the strong form, we use the following notations: $\perp_s$ indicates orthogonality w.r.t. $(\cdot,\cdot)_{L^2(\Omega)}$ and $G_s(u)$ the strong form corresponding to the weak form $G(u)$, such that $G_s(u) = 0$ can be appropriately tested by $(\cdot,\cdot)_{L^2(\Omega)}$ on $\mathcal{V}_b$. Then (4.110) is modified into:

$$\begin{aligned} G_s(u_0) = 0 &\Leftrightarrow (G_s(u_0), v)_{L^2(\Omega)} = 0 \ \forall \ \ v \in \mathcal{V}_b \\ &\Leftrightarrow G_s(u_0) \perp_s \mathcal{V}_b. \end{aligned} \tag{4.111}$$

Most of the time we consider for FEMs the weak form (4.110). Mainly in connection with collocation we consider the strong form (4.111) as well. Again we use the neutral denotation for both of (4.110) and (4.111) (,with $\perp$ indicating orthogonality w.r.t.$\langle\cdot,\cdot\rangle_{\mathcal{V}'\times\mathcal{V}}$,) and determine $u_0 \in \mathcal{U}_b$ s.t.

$$G(u_0) = 0 \Leftrightarrow < G(u_0), v >_{\mathcal{V}'\times\mathcal{V}} = 0 \ \forall \ v \in \mathcal{V}^h \ \Leftrightarrow \ G(u_0) \perp \mathcal{V}_b. \tag{4.112}$$

To reduce the technical difficulties we confine the presentation for the beginning to the case of a bounded linear operator, e.g., $A := G_u(u_1)$, in either weak or strong form. The evaluation and consistent differentiability of the nonlinear operators will be discussed separately, see Section §4.4. For (4.112) and $A \in \mathcal{L}(\mathcal{U}, \mathcal{V}')$, we determine

$$u_0 \in \mathcal{U}_b : Au_0 = f, \quad f \in \mathcal{V}' \Leftrightarrow a(u_0, v) =< f, v >_{\mathcal{V}'\times\mathcal{V}} \ \forall \ \ v \in \mathcal{V}_b, \tag{4.113}$$

with $a(u,v) =< Au, v >_{\mathcal{V}'\times\mathcal{V}} \ \forall \ v \in \mathcal{V}_b, \ \forall u \in \mathcal{U}_b$ fixed. Again we obtain, for Dirichlet boundary conditions and for the two choices $\mathcal{U}_b, \mathcal{V}_b = H_0^1(\Omega), H^1(\Omega)$ and $= H^2(\Omega) \cap H_0^1(\Omega), L^2(\Omega)$ the weak and strong bilinear forms $a(\cdot,\cdot)$ and $a_s(\cdot,\cdot)$, resp., similarly for natural boundary conditions.

For the different linear and bilinear forms, projectors and operators we use Notation 4.1. This includes the functionals f, f^h in (4.60), (4.61), or the corresponding pairings $< \cdot, \cdot >_{\mathcal{V}'\times\mathcal{V}_b^h}$, $< \cdot, \cdot >^h$, the bilinear forms $a^h(\cdot,\cdot)$ e.g. $a_s(\cdot,\cdot), a_s^h(\cdot,\cdot), \tilde{a}_s^h(\cdot,\cdot)$, see (4.68), and the projectors and operators $Q^{'h}$, e.g. $Q^{'h}, Q_s^{'h}, \tilde{Q}^{'h}, \tilde{Q}_s^{'h}$, in (4.65), the A_h, e.g. A, A_s, A_h in (4.55), (4.58), (4.59).

We want to repeat the discrete equations. The so-called *(exact or conforming) Petrov-Galerkin methods* for (4.113) determine, with the original A, $a(\cdot,\cdot)$, the $u_0^h \in \mathcal{U}_b^h \subset \mathcal{U}_b$ s.t., see Definition 4.3.3,

$$\begin{aligned} &A^h u_0^h - Q^{'h} f \ = Q^{'h}(Au_0^h) - Q^{'h} f \ = 0 \ \Leftrightarrow \\ &< Q^{'h}(Au_0^h - f), v^h >_{\mathcal{V}'\times\mathcal{V}_b} = a(u_0^h, v^h) - < f, v^h >_{\mathcal{V}'\times\mathcal{V}_b^h} = 0 \ \forall \ \ v^h \in \mathcal{V}_b^h. \end{aligned} \tag{4.114}$$

The so-called *non conforming Petrov-Galerkin methods* for (4.113) allow variational crimes, e.g., $\mathcal{U}_b^h \not\subset \mathcal{U}_b$ or $\mathcal{U}_b^h \not\subset \mathcal{U}$, and similarly for $\mathcal{V}_b^h$. With the above general notations, see (4.65), (4.66), they determine, with extended A_h, $a^h(\cdot,\cdot)$, the $u_0^h \in \mathcal{U}_b^h$ s.t.

$$\begin{aligned} &A^h u_0^h - Q^{'h} f \ = Q^{'h}(A_h u_0^h) - Q^{'h} f \ = 0 \text{ with } A^h = Q^{'h} A_h|_{\mathcal{U}_b^h} \ \Leftrightarrow \\ &< A^h u_0^h - Q^{'h} f, v^h >^h = a^h(u_0^h, v^h) - < f, v^h >^h \ = \ 0, \ \forall \ \ v^h \in \mathcal{V}_b^h. \end{aligned} \tag{4.115}$$

Finally, we combine $Au,\ f \in \mathcal{V} \cap C(\overline{\Omega})$ with the $\tilde{Q}'^h$, e.g., defined by quadrature, to obtain

$$\tilde{A}^h u_0^h - \tilde{Q}'^h f = \tilde{Q}'^h (A_h u_0^h) - \tilde{Q}'^h f = 0 \text{ with } \tilde{A}^h = \tilde{Q}'^h A_h|_{\mathcal{U}_b^h} \Leftrightarrow (4.116)$$
$$< \tilde{A}^h u_0^h - \tilde{Q}'^h f, v^h >^h = \tilde{a}^h(u_0^h, v^h) - < f, v^h >^h = 0 \ \forall \ v^h \in \mathcal{V}_b^h.$$

If we choose a special interpolation basis $v_{i,T} \in \mathcal{V}_b^h, v_i(y_j) = \delta_{i,j}\ \forall i,j\ \forall T \in \mathcal{T}^h$, for the different quadrature points y_j, this method can often be re-interpreted as collocation method, see Sections 4.1, (4.47) and below.

Remark 4.3.2. The P^h, Q'^h should be chosen s.t. their errors have similar orders of magnitude as those of the consistency errors below. Otherwise the results are no longer optimal.

The main examples for this theory are finite element- and spectral methods. Difference methods are treated in this context in [16].

Definition 4.3.3. Generalized Petrov- Galerkin (P-G) methods: *Let* $\mathcal{U}_b^h$, $\mathcal{V}_b^h$ *be* conforming and generalized P-G admissible approximating spaces, *see Definition 4.3.1 and* A, A_h, $a^h(\cdot,\cdot)$, $<\cdot,\cdot>^h$, Q'^h, $f^h(\cdot)$, *be the exact and extended or approximate operators, bilinear forms, pairings, projectors and functionals, see (4.55), (4.58), (4.68), (4.61), (4.65), (4.64): Then we call (4.114), (4.115) and their approximate variants, (4.116), the* induced discretization methods, *in particular.*

(a) if $\mathcal{U}^h = \mathcal{V}^h \subset \mathcal{U}_b$, *we call (4.114) a* Galerkin method,
(b) for $\mathcal{U}^h \neq \mathcal{V}^h$ *we call (4.114) and (4.115) a* Petrov-Galerkin (P-G) method
(c) for $U_b^h \subset \mathcal{U}_b, \mathcal{V}_b^h \subset \mathcal{V}_b$ *we call (4.114) a* conforming FE or spectral for short a P- G method
(d) for an approximate $\tilde{A}^h \approx A^h$ *and/or* $\tilde{Q}'^h$, $\tilde{f}^h(\cdot)$
we call (4.115) a generalized Petrov-Galerkin (P-G) method,
(e) if $\mathcal{U}_b^h \not\subset \mathcal{U}_b$ *or* $\mathcal{U}_b^h \not\subset \mathcal{U}$, *similarly for* $\mathcal{V}_b^h$, *we call (4.115)* a Petrov-Galerkin (P-G) method, with variational crimes.
(f) the last two cases (d), (e) we call non conforming FE or spectral, for short Petrov-Galerkin (P-G) methods
(g) if we use the strong versions $A_s, A_{s,h}, \tilde{A}_s^h, Q_s'^h, \tilde{Q}_s'^h$ *we call these procedures* strong (conforming or non conforming) methods *e.g. a* strong Galerkin method
(h) for short we call the cases (d) - (g) generalized P-G- methods.

For FE and spectral methods each of the above choices $A^h = Q'^h A|_{\mathcal{U}_b^h}$, $A^h = Q'^h A_h|_{\mathcal{U}_b^h}$, $\tilde{A}^h = \tilde{Q}'^h A_h|_{\mathcal{U}_b^h}$ implicitly defines a linear operator Φ^h. This $\Phi^h : \mathcal{L}(\mathcal{U}_b, \mathcal{V}_b') \to \mathcal{L}(\mathcal{U}_b^h, \mathcal{V}_b^{h'})$ indicates a general procedure, defining the discrete operator A^h for the original A, according to the chosen method. Again, we use the general notation Φ^h. It yields the "discretization" $A^h = \Phi^h(A)$. A similar

approach is possible, as we have seen in Section 4.2, for nonlinear operators as well. So we have $A^h = \Phi^h(A)$ or $G^h = \Phi^h(G)$ for linear and nonlinear operators A and G, resp. It is important that Φ^h acts on the different u dependent terms in A or G by applying linear operators, e.g. the Fourier collocation derivative, see [23, 11, 13, 6, 7, 65] for details. This operator

$$\begin{aligned} &\Phi^h : (\mathcal{U}_b \rightarrow \mathcal{V}_b') \rightarrow (\mathcal{U}_b^h \rightarrow \mathcal{V}_b^{h'}) \text{ for nonlinear operators and} \\ &\text{its restriction } \Phi^h : \mathcal{L}(\mathcal{U}_b, \mathcal{V}_b') \rightarrow \mathcal{L}(\mathcal{U}_b^h, \mathcal{V}_b^{h'}) \text{ with } \Phi^h(A) = A^h \\ &\text{is indeed linear, since } \Phi^h(A + C) = \Phi^h(A) + \Phi^h(C). \end{aligned} \tag{4.117}$$

This last condition excludes , e.g., *Runge-Kutta methods.* These methods are not important in our context.

So, we have the following *diagram,* with the general notation A^h, Q'^h, Φ^h, e.g., $\tilde{A}^h, \tilde{Q}'^h, \tilde{\Phi}^h$ or $A_s^h, Q_s'^h, \Phi_s^h$

$$\begin{array}{ccccccc} \mathcal{U} \supset & \mathcal{U}_b & & \xrightarrow{A, A_h} & \mathcal{V}_b' & \overset{\text{exactly tested by}}{\longleftrightarrow} & \mathcal{V}_b \subset \mathcal{V} \\ & P^h \downarrow \uparrow E^h & & \Phi^h \downarrow & Q'^h \downarrow & & \\ & \mathcal{U}_b^h & & \xrightarrow{A^h} & \mathcal{V}_b^{h'} & \overset{\text{(approximately) tested by}}{\longleftrightarrow} & \mathcal{V}_b^h \end{array} \tag{4.118}$$

Here P^h, E^h, Q'^h are uniformly bounded; w.r.t. the appropriate norms. E^h is the identical or a more general embedding, sometimes even with an additional embedding $E_v^h : \mathcal{V}^h \rightarrow \mathcal{V}$. We will come back to the necessary comparison of A and A^h in Chapter 6.

We have to check, how well the discrete solution, u_0^h, satisfies the original equation. We will end up with two different, however, closely related concepts the variational and the classical consistency errors.

Similarly to the classical definitions, we start defining a measure of how strongly a *nonconforming Petrov-Galerkin method* (or *with variational crimes*) violates the conformity conditions. For $\mathcal{U}_b^h \subset \mathcal{U}_b$, $\mathcal{V}_b^h \subset \mathcal{V}_b$ satisfied or violated, we have defined the exact and discrete solutions u_0 and u_0^h by

$$\begin{aligned} &u_0 \in \mathcal{U}_b : a(u_0, v) - < f, v >_{\mathcal{V}' \times \mathcal{V}_b} = 0 \ \forall \ v \in \mathcal{V}_b \text{ and, e.g.,} \\ &u_0^h \in \mathcal{U}_b^h : a^h(u_0^h, v^h) - < f, v^h >_{\mathcal{V}' \times \mathcal{V}_b}^h \ = \ 0 \ \forall \ v^h \in \mathcal{V}_b^h \text{ resp.} \end{aligned} \tag{4.119}$$

Hence, the second equation does and does not represent a subset of conditions for conforming and non conforming Petrov-Galerkin methods methods, resp. Subtraction yields

$$a(u_0 - u_0^h, v^h) \ = \ 0 \ \forall \ \ v^h \in \mathcal{V}_b^h \text{ (for conforming methods).} \tag{4.120}$$

If we replace $a(u_0 - u_0^h, v^h)$ by the general $a^h(u_0 - u_0^h, v^h)$, and consider non-conforming Petrov-Galerkin methods this term is nontrivial, see (4.13), (4.29). Then

$$a^h(u_0 - u_0^h, v^h) \text{ or } \sup_{0 \neq v^h \in \mathcal{V}_b^h} \{|a^h(u_0 - u_0^h, v^h)|/\|v^h\|\} \tag{4.121}$$

is denoted as *variational consistency error* in the FE community. However, it slightly differs from the classical notation, e.g. [57]. With respect to the induced operator A^h and if u_0 is smooth enough to define $A_h u_0$, we transform the relation (4.121) into

$$Q^{'h} A_h u_0 - A^h u_0^h \; = A^h(P^h u_0 - u_0^h) + Q^{'h}(A_h u_0 - A_h P^h u_0).$$

If additionally u_0 is smooth enough the

$$\|Q^{'h} A_h (u_0 - P^h u_0)\|_{\mathcal{V}'} = \mathcal{O}(\|u_0 - P^h u_0\|_{\mathcal{U}}) \to 0.$$

This can be generalized to arbitrary smooth enough u , if we introduce $f_u := Au$ and u^h is the discrete solution of $A^h u^h = Q^{'h} f_u$. In general notation and with $A_h u = Au$, this reads as

$$\begin{aligned} Q^{'h} A_h u - A^h u^h &= A^h(P^h u - u^h) + Q^{'h} A_h (u - P^h u) \\ &= A^h P^h u - Q^{'h} f_u + Q^{'h} A_h (u - P^h u) \\ &= A^h P^h u - Q^{'h} A_h u + Q^{'h} A_h u - A^h P^h u \\ &= A^h P^h u - Q^{'h} A u + \mathcal{O}(\|u - P^h u\|_{\mathcal{U}}). \end{aligned} \tag{4.122}$$

For the following definition, we use the notation as in (4.26), (4.34), (4.77),

$$A^h u := Q^{'h} A_h u, \text{ or } \tilde{A}^h u := \tilde{Q}^{'h} A_h u \; \forall u \in \mathcal{U}_b \text{ smooth enough.} \tag{4.123}$$

We start with the linear case here and motivate and study the nonlinear case in Section 4.4.

Definition 4.3.4. Consistency errors: *Under the conditions of Definition 4.3.3 we choose $a^h(\cdot,\cdot)$ as in (4.67), (4.68) and $A^h u^h = Q^{'h} Au$. Then*

$$a^h(u - u^h, v^h) \text{ or } \sup_{0 \neq v^h \in \mathcal{V}_b^h} \{|a^h(u - u^h, v^h)|/\|v^h\|\} \tag{4.124}$$

is called (variational) consistency error *in u. (For conforming and non conforming methods it is trivial and nontrivial, resp.) This notation in (4.124) is used similarly for the isoparametric variants. Define a discrete operator A^h by applying a Petrov-Galerkin method to a linear operator A. Then*

$$\|A^h P^h u - Q^{'h} A_h u\|_{\mathcal{V}'} \tag{4.125}$$

is called (classical) consistency error *in u. A method is called* variationally and classically consistent *for A in u, if*

$$\begin{aligned} &\sup_{0 \neq v^h \in \mathcal{V}_b^h} \{|a^h(u_0 - u_0^h, v^h)|/\|v^h\|\} \to 0 \text{ and} \\ &\|A^h P^h u - Q^{'h} A_h u\|_{\mathcal{V}'} \to 0 \text{ for } h \to 0, \end{aligned} \tag{4.126}$$

resp. If even

$$\|A^h P^h u - Q'^h A_h u\|_{\mathcal{V}'} \text{ or } \sup_{0 \neq v^h \in \mathcal{V}_b^h} \{|a^h(P^h u - u, v^h)|/\|v^h\|\} = \mathcal{O}(h^p),$$

we call it consistent of order p. *These* $\mathcal{O}(h^p)$ *are usually estimated by* $Ch^p \|u\|_{W_q^k(\Omega)}$ *with appropriate* p, k, q.

Remark 4.3.5. The variational consistency error usually is only considered and estimated for the exact solution, u_0, or if necessary for an approximant $P^h u_0$, see Chapter 6 for estimates. A generalization to arbitrary u is possible, see (4.122). The classical consistency errors are often studied for general u. There is the simple relation between the two terms in (4.122). Hence, *through the rest of the Booklet we assume, and, in fact, do have*

$$\|u - P^h u\|_{\mathcal{U}} \leq \|I - P^h\|_{\mathcal{U}_b^h \leftarrow \mathcal{U}_b} \|u\|_{\mathcal{U}} \ \to 0 \ \forall h \to 0, \ \forall u \in \mathcal{U}_b. \quad (4.127)$$

Both consistency errors even have *the same size*, if both dominate the interpolation error $\|u_0 - P^h u_0\|_{\mathcal{U}}$.

We need the important concept of stability for linear and nonlinear operators in Section 4.4. We start her with the linear stability and extend it to the nonlinear case in Chapter 7.

Definition 4.3.6. Stability: *Let* $\mathcal{U}^h$, $\mathcal{V}^h$ *be* conforming or general admissible approximating spaces, *see Definitions 4.3.1,4.3.3. Assume a bounded linear operator* $B^h \in \mathcal{L}(\mathcal{U}_b^h, \mathcal{V}_b^h)$ *is assigned to every* $h \in H$. *The sequence* $(B^h)_{h \in H}$, *or for short, only* B^h, *is called* stable, *if there are positive constants* h_0, C, *independent of* h, *such that for* $h \leq h_0$ *the* $(B^h)^{-1} \in \mathcal{L}(\mathcal{V}_b^h, \mathcal{U}_b^h)$ *exists and that* $||(B^h)^{-1}||_{\mathcal{U}_b^h \leftarrow \mathcal{V}_b^h} \leq C$.

4.4 Stability and Consistency yield Convergence

We want to unfold this well-known fact for FE and spectral methods with variational crimes. We do that [3] in four steps:

We start with a direct comparison between A and A^h or G and G^h as in [57] in this Section. We show that "Stability and Consistency yield Convergence". Then we have to prove stability and consistency in the following Chapters. We present in Chapter 5 generalized Strang Lemmas for the familiar (weak) bilinear form approach for FEMs. In Chapter 6 we estimate the errors necessary for the generalized Strang Lemmas in Chapter 5 and prove stability for coercive bilinear forms. Finally, in Chapter 7 we prove the general stability and convergence.

[3] We use throughout the notations, see 4.1, and Definitions 4.3.1,4.3.3,4.3.4, 4.3.6 in Sections 4.2 and 4.3.

For the $a^h(\cdot,\cdot), f^h, A^h$, and for $Q^{'h}, P^h$, in the above general notation, we want to use the results in Sections 4.2 and 4.3 as motivation for the general nonlinear case. For consistency, the Gu and its discrete counterpart have to be compared, hence,

$$G^h P^h u - Q^{'h} G_h u = (\Phi^h G)(P^h u) - Q^{'h} G_h u, \tag{4.128}$$

in particular, for the solutions u_0, u_0^h of $G(u_0) = 0$ and $(\Phi^h G)(u_0^h) = 0$. For the linear case the results in Lemmas 5.1.2 - 5.1.4 will allow estimates of the form

$$\|u_0^h - u_0\|_{\mathcal{U}}^h \leq C(\text{dist }(u_0, \mathcal{U}_b^h) + \sup_{0 \neq v^h \in \mathcal{V}_b^h} \frac{|a^h(u_0 - u_0^h, v^h)|}{\|v^h\|_{\mathcal{V}}^h}). \tag{4.129}$$

for the extended $a^h(u_0 - u_0^h, v^h)$ defined $\forall v^k \in \mathcal{V}_b^h$. Usually, $u_0 \notin \mathcal{U}_b^h$, so $a^h(u_0, v^h)$ might not be defined and stability is not applicable and available for $a^h(u_0, v^h)$, but only for

$$a^h(\cdot,\cdot) : \mathcal{U}_b^h \times \mathcal{V}_b^h \to \mathbb{R}, \text{ e.g. } a^h(P^h u_0, v^h).$$

So, we better study,

$$\begin{aligned} a^h(P^h u_0, v^h) - f^h(v^h) &= a^h(P^h u_0 - u_0^h, v^h) \\ &= a^h(u_0 - u_0^h, v^h) + a^h(P^h u_0 - u_0, v^h) \ \forall \ \ v^h \in \mathcal{V}_b^h, \end{aligned}$$

here $\tilde{a}^h(u_0, v^h)$ is only defined if u_0 is smooth enough. For the induced operator A^h this reads, with well defined $A_h u_0 = A u_0$, see (4.122), as

$$\begin{aligned} A^h P^h u_0 - Q^{'h} A u_0 &= A^h P^h u_0 - f^h = A^h(P^h u_0 - u_0^h) \\ &= Q^{'h} A_h u_0 - A^h u_0^h + A^h P^h u_0 - Q^{'h} A_h u_0 \\ &= Q^{'h} A u_0 - A^h u_0^h + Q^{'h} A_h (P^h u_0 - u_0). \end{aligned} \tag{4.130}$$

The first term $A^h P^h u_0 - Q^{'h} A u_0$ is called *local discretization error* in, e.g., [57], see (4.126), for $G u_0 = A u_0 - f = 0$. The $Q^{'h} A_h u_0 - A^h u_0^h$ corresponds to the above *variational consistency error in FEs*, see Definition 4.3.4, and see Chapter 6 for estimates. The last term, $Q^{'h} A_h (P^h u_0 - u_0)$, is, for smooth u_0 and with $\|P^h u_0 - u_0\|_{\mathcal{U}}^h \to 0$ automatically small. For a stable A^h applied to $A^h(P^h u_0 - u_0^h)$ we obtain

$$\begin{aligned} \|P^h u_0 - u_0^h\|_{\mathcal{U}}^h &\leq C \|A^h P^h u_0 - Q^{'h} A u_0\|_{\mathcal{V}'}^h \text{ or} \\ &\leq C \ \ (\|P^h u_0 - u_0\|_{\mathcal{U}}^h + \|Q^{'h} A u_0 - A^h u_0^h\|_{\mathcal{V}'}^h), \end{aligned}$$

implying (4.129) in a modified form.

Now, we use these insights to motivate the definitions for the general nonlinear case: We started with a weak or strong operator, e.g., $G : H_0^1(\Omega) \cap H_0^{-1}(\Omega)$ or $G : H^2(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega)$, in the general notation

$$G : \mathcal{U}_b \to \mathcal{V}_b',\ \mathcal{U}_b,\ \mathcal{V}_b' \text{ Banach spaces, solve } G(u_0) = 0. \tag{4.131}$$

To obtain good convergence or to apply quadrature approximations, we need a smooth solution $u_0 \in H^{m+1}(\Omega) \cap H_0^1(\Omega)$ or even, e.g.,

$$G : \mathcal{U}_{b,s} = H^{m+1}(\Omega) \cap H_0^1(\Omega) \to \mathcal{V}_{b,s}' = H^{m-1}(\Omega), m > 0. \tag{4.132}$$

Obviously, modifications for differential equations of higher order are necessary. We choose the interpolation or truncation operators $P^h = I^h, T^h$ or projectors Q'^h :

$$P^h : \mathcal{U}_b \to \mathcal{U}_b^h, P^h; \text{ and } Q'^h : \mathcal{V}_b' \to \mathcal{V}_b^{h'} \text{ as linear bounded operators} \tag{4.133}$$

Sometimes we even study extensions $P^h : \mathcal{U} \to \mathcal{U}^h$ and $Q'^h : \mathcal{V}' \to \mathcal{V}^{h'}$. They have, for smooth $u \in \mathcal{U}_b, f \in \mathcal{V}'$, the usual approximation properties

$$\begin{aligned} &\|P^h u - u\|_{\mathcal{U}}^h \to 0 \text{ or } \|Q'^h f - f\|_{\mathcal{V}'}^h \to 0 \text{ for } u \in \mathcal{U}_b,\ \ f \in \mathcal{V}_b' \text{ and} \\ &\|P^h u - u\|_{\mathcal{U}}^h = \mathcal{O}(h^p) \text{ or } \|Q'^h f - f\|_{\mathcal{V}'}^h = \mathcal{O}(h^p) \text{ for } u \in \mathcal{U}_{bs} f \in \mathcal{V}_{b,s}' \end{aligned} \tag{4.134}$$

for $h \to 0$. This implies

$$\begin{aligned} \lim_{h\to 0} \|P^h u\|_{\mathcal{U}}^h = \|u\|_{\mathcal{U}} \text{ and } \lim_{h\to 0} \|Q'^h f\|_{\mathcal{V}'}^h = \|f\|_{\mathcal{V}'} \\ \text{for fixed } u \in \mathcal{U},\ f \in \mathcal{V}'. \end{aligned} \tag{4.135}$$

Furthermore, see Section 4.3, (4.117), the operator Φ^h is applicable to $A, A-f$ or G in (4.112), and

$$\Phi^h(A) = A^h = Q'^h A_h|_{\mathcal{U}_b^h} \text{ or } \Phi^h(G) = G^h = Q'^h G_h|_{\mathcal{U}_b^h}. \tag{4.136}$$

Specific examples are

$$A^h = Q'^h A_h|_{\mathcal{U}_b^h} \text{ or } Q'^h \tilde{A}_h|_{\mathcal{U}_b^h} \text{ or } \tilde{Q}'^h A_h|_{\mathcal{U}_b^h} \text{ or } \tilde{Q}'^h \tilde{A}_h|_{\mathcal{U}_b^h}$$

and similarly

$$G^h = Q'^h G_h|_{\mathcal{U}_b^h} \text{ or } Q'^h \tilde{G}_h|_{\mathcal{U}_b^h} \text{ or } \tilde{Q}'^h G_h|_{\mathcal{U}_b^h} \text{ or } \tilde{Q}'^h \tilde{G}_h|_{\mathcal{U}_b^h}.$$

Here $\tilde{A}$, $\tilde{G}$ and $\tilde{Q}'^h$ have been obtained, see Section 4.1, e.g., by applying linear approximation operators to the $u-$dependent terms in A, G, e.g., Fourier collocation derivatives for u'. Then we obtain, as special case of [57, 58, 66, 68]

Definition 4.4.1. *Let the sequence of spaces* $\{\mathcal{U}_b^h, \mathcal{V}_b^h\}_{h\in H}$ *and of bounded linear operators* $\{P^h, Q'^h, \Phi^h\}_{h\in H}$ *satisfy (4.135) and* $G \in \mathcal{D}(\Phi^h)$. *Then we call* $\mathfrak{M} := \{\mathcal{U}_b^h, \mathcal{V}_b^{h'}, P^h, Q'^h, \Phi^h\}_{h\in H}$ *a* discretization method applicable to G *in (4.131), see (4.112). This defines the* discretion, $\Phi^h G = G^h : \mathcal{U}_b^h \to \mathcal{V}_b^{h'}$

and the[4] discrete equation, $G^h(u_0^h) = 0$. *Let furthermore* $u \in \mathcal{U}$ *(or* $\mathcal{U}_b$*) satisfy* $u \in \mathcal{D}(G_h) \cap \mathcal{D}((\Phi^h G)P^h)$ *s.t.*

$$\|(\Phi^h G)P^h u - Q'^h G_h u\|_{\mathcal{V}'}^h \to 0 \text{ and } \mathcal{O}(h^p) \text{ for } h \to 0, h \in H. \quad (4.137)$$

Then the discretization method is called classically consistent *with* G *at* u, *and* of order p, *resp. Evaluated for the exact solution* u_0 *of* $G(u_0) = 0$, $G^h(P^h u_0) = (\Phi^h G)(P^h u_0), h \in H$ *is called* local discretization error *of* $(\Phi^h G) = G^h$ *or of the discretization method* $\mathfrak{M}$. *Sometimes one might have to choose a subset* $\overline{H} \subset H$ *and admit only* $h \in \overline{H}$.

The choice of P^h, Q'^h for the given $\mathcal{U}_b^h, \mathcal{V}_b^{h'}$ and problem (4.112) is certainly not unique, e.g., the above interpolation or truncation operators. However the combination $\mathcal{U}_b^h, \mathcal{V}_b^{h'}, P^h, Q'^h, \Phi^h$ should be chosen appropriately to yield consistency or even (4.137) with the highest possible p.

We have already introduced in Definition 4.3.6 the stability for linear operators B^h. We generalize it in

Definition 4.4.2. *Let* $G^h :\to \mathcal{V}_b^{h'}$, $G^h \in \mathcal{D}(G^h) \subseteq \mathcal{U}_b^h$, *be defined by the discretization method* $\mathfrak{M}$ *and let* $u^h \in \mathcal{D}(G^h) \subseteq \mathcal{U}_b^h \; \forall h \in H$. *Furthermore, let* $r, S \in \mathbb{R}_+$ *be fixed constants, s.t.,* uniformly in $h \in H$*:*

$$\begin{aligned} u_i^h \in B_r(u^h) &:= \{v^h \in \mathcal{U}^h : \|v^h - u^h\|_{\mathcal{U}}^h \le r\}, i = 1, 2, \\ &\Rightarrow \|u_1^h - u_2^h\|_{\mathcal{U}}^h \le S\|G^h(u_1^h) - G^h(u_2^h)\|_{\mathcal{V}'}^h. \end{aligned} \quad (4.138)$$

Then G^h *is called* stable in u^h *and* S *and* r *are called* stability bound and stability threshold, *resp.*

We see immediately, that for $G^h u^h = A^h u^h + f^h$ with linear A^h, the above definitions for consistency and stability are equivalent to the Definition 4.3.4 and 4.3.6. This result can even be extended in Theorem 7.1.3 .

We will discuss in Theorem 7.1.3 the conditions, which guarantee stability. E.g., it allows to reduce stability for the nonlinear problem to that of the linear problem. In Stetter [57] the first inequality in (4.138) is replaced by the weaker assumption.

$$\|G^h(u_i^h) - G^h(u^h)\|_{\mathcal{V}'}^h \le R, \; i = 1, 2. \quad (4.139)$$

Since for nonlinear G (and hence G^h) this (4.139) may be satisfied for large $\|u^h - u_i^h\|_{\mathcal{U}}^h$ we have chosen the stronger assumption in (4.138). In fact we have, see [57], Corollary 1.2.2, the following

Corollary 4.4.3. *Let in the discretization* $G^h : \mathcal{U}_b^h \to \mathcal{V}_b^{h'}$ *the* G^h *be continuous in* $B_r(u^h)$ *and satisfy (4.138). Then these* $\{u_i^h\}_{h \in H}$ *satisfy (4.139) with* $R = r/S$.

[4] although usually $\mathcal{D}(G_h) \subseteq \mathcal{U}_b, \mathcal{D}(G_h) \neq \mathcal{U}_b$, we keep this simplifying notation.

Consistency plus stability imply convergence. This is the essence of the following theorem, see Stetter [57]:

Theorem 4.4.4. *Let the original problem (4.112) have the exact solution* u_0 *and let* $G^h : \mathcal{U}_b^h \to \mathcal{V}_b^{h'}$ *be its discretization and satisfy the following conditions*

1. $G^h = \Phi^h(G) : \mathcal{U}_b^h \to \mathcal{V}_b^{h'}$ *is defined and continuous in* $B_r(P^h u_0)$ *with* $r > 0$ *independent of* h,
2. G^h *is (classically) consistent with* G *and consistent of order* p *in* $P^h u_0$,
3. G^h *is stable for* $P^h u_0$.

Then the discrete problem $G^h(u^h) = 0$ *possesses the unique solution* $u_0^h \in \mathcal{U}_b^h$ *for all sufficiently small* $h \in H$ *and* u_0^h *converges and converges of order* p *to* u_0, *respectively.*

This Theorem shows that we have to prove consistency and stability for G^h. The proof of (classical) consistency for the conforming FEMs is straight forward, see (4.120):

Theorem 4.4.5. *Let* $G \in C^1(\mathcal{U}_b)$, *and* $u, P^h u \in \mathcal{D}(G)$. *Then for the conforming FEMs the* variational and classical consistency error *vanishes and is estimated by*

$$(\Phi^h G)(P^h u) - Q^{'h}(Gu) = \mathcal{O}(\|P^h u - u\|_{\mathcal{U}}) \text{ resp.}$$

Proof : We have for $G : \mathcal{U}_b \to \mathcal{V}_b'$ and $\mathcal{U}_b^h \subset \mathcal{U}_b$, $\mathcal{V}_b^h \subset \mathcal{V}_b$

$$(\Phi^h G)u^h := G^h u^h := Q^{'h}(Gu^h). \tag{4.140}$$

Then,with $G \in C^1(\mathcal{U}_b)$, and $u, P^h u \in \mathcal{D}(G)$, we find by the Mean Value Theorem

$$(\Phi^h G)(P^h u) - Q^{'h}(Gu) = Q^{'h}(G(P^h u) - G(u))$$

$$= Q^{'h} \int_0^1 G'(u + t(P^h u - u))dt(P^h u - u) = \mathcal{O}(\|P^h u - u\|_{\mathcal{U}}). \qquad \blacksquare$$

This argument breaks down for variational crimes. In fact, the violated continuity and boundary conditions imply complicated errors.

Now, we come back to the general structure of the *FE and spectral methods.* By (4.136) the $\Phi^h A = A^h$ or $\Phi^h G = G^h$ are obtained by applying bounded linear operators, e.g., $Q^{'h}, P^h, I^h, T^h$ to $A_h u^h$ or $G_h u^h$. Otherwise approximations as, e.g., the Fourier collocation derivatives and de-aliasing operators for spectral methods may be applied to the argument u or some of its derivatives, constituting A or G. So we employ and need *inner and outer bounded linear operators*, P^h and $Q^{'h}$, to obtain

$$\|G^h(P^h u) - Q^{'h} G_h(u)\|_{\mathcal{V}'} = \mathcal{O}(h^p)\ \|u\|_{\mathcal{U}_{b,s}} \text{ for smooth } u \in \mathcal{U}_{b,s}. \tag{4.141}$$

These claims for spectral methods have been studied in detail in [11, 12, 13, 6, 7, 15]. The linearity of this Φ^h in (4.117) obviously allows to exchange differentiation with these operators. As indicated above already, this property contrasts to and excludes methods of the Runge-Kutta type, see above. Here implicit and repeated function evaluations destroy the linearity.

The proof for the following Theorem for FE and spectral methods is very similar to that in [13]: For non conforming FEMs we have to come back to this problem at the end of Chapter 6.

Theorem 4.4.6. *Let the nonlinear operator G satisfy $G : \mathcal{U}_b \to \mathcal{V}_b, G \in C^r(\mathcal{D}(G))$, $G(u_0) = 0$, $\|u_0 - u\|_{\mathcal{U}}$ be small, let $\mathcal{U}_b^h$, $\mathcal{V}_b^h$ Φ^h define conforming Petrov-Galerkin methods and let G^h be evaluated corresponding to (4.136), employing inner and outer bounded linear operators. Then the operator G^h is* consistent *and* r-times consistently differentiable *with G, that is, for $j = 1, \ldots, r$,*

$$\begin{aligned} &\|G^h(P^h u) - Q'^h G u\|_{\mathcal{V}'}^h = \mathcal{O}(\|I^h u - u\|_{\mathcal{U}}^h), \\ &\|(G^h)^{(j)}(P^h u)P^h u_1 \cdot \ldots \cdot P^h u_j - Q'^h G^{(j)}(u)u_1 \cdot \ldots \cdot u_j\|_{\mathcal{V}'}^h = \\ &\quad = \mathcal{O}(\|I^h u_1 - u_1\|_{\mathcal{U}}^h \cdot \ldots \|I^h u_j - u_j\|_{\mathcal{U}}^h)(1 + \|I^h u - u\|_{\mathcal{U}}^h) \end{aligned} \tag{4.142}$$

for $u, u_1 \ldots, u_j \in \mathcal{U}$ with $\|u - u_0\|_{\mathcal{U}}^h$ sufficiently small. Analogous results for $\tilde{P}^h, \tilde{Q}'^h$-combinations are valid as well.

Proof: The higher derivatives with fixed, e.g., $u, u_1 \ldots, u_{j-1}$ can be interpreted as corresponding bilinear forms. All the operators defining Φ^h are bounded and linear and $\mathcal{U}_b^h$, $\mathcal{V}_b^h$ define conforming Petrov-Galerkin method. Hence, linear A satisfy (4.141) and the A^h are consistent. This implies the consistency of the $(G^h)^{(j)}(P^h u)P^h u_1 \cdot \ldots \cdot P^h u_j$ as well. That in fact they are related by (4.142) follows as in [12, 13]. ■

5. Generalized Strang Lemmas

This Chapter presents the influence of violated boundary conditions, continuity and approximate evaluation generalizing the well known Cea Lemma. In contrast to Chapters 4 and 7 we have to distinguish here and in Chapter 6 the different bilinear forms.

Following Theorem 4.4.4 we have to show stability and consistency for a general class of operators and generalized Petrov-Galerkin methods. As indicated above, we do that in four steps. In this Chapter we discuss the variational consistency errors for weak and strong formulations of the problem.

5.1 Generalized Strang Lemmas

We consider bilinear forms $a^h(\cdot,\cdot)$, satisfying a discrete inf- sup-condition. This condition has been verified in Theorem 3.3.3 for coercive bilinear weak forms. With the estimates for the $|a^h(u_0 - u_0^h, v^h)|/\|v^h\|_{\mathcal{V}}^h$ and $|\tilde{a}^h(u_0 - u_0^h, v^h)|/\|v^h\|_{\mathcal{V}}^h$ in Chapter 6, we will be able to to prove these inf–sup–conditions for general elliptic bilinear forms in Chapter 7. This implies the case of boundedly invertible A^h and hence the general stability for linear and nonlinear operators, see Chapter 7. So, the following generalized Strang-Lemmas play the role of a certain guideline for the further studies. We estimate

$$\|u_0^h - u_0\|_{\mathcal{U}}^h \leq C(\text{dist } (u_0, \mathcal{U}_b^h) + \sup_{0 \neq v^h \in \mathcal{V}_b^h} \frac{|a^h(u_0 - u_0^h, v^h)|}{\|v^h\|_{\mathcal{V}}^h}), \text{ if}$$
$$a^h(u_0 - u_0^h, v^h) \text{ is defined } \forall v^k \in \mathcal{V}_b^h. \quad (5.1)$$

The last condition is necessary for the case of $\tilde{a}^h(u_0 - u_0^h, v^h)$ which requires smooth enough u_0.

The missing estimates for the variational discretization error $\sup_{0 \neq v^h \in \mathcal{V}_b^h} |a^h(u_0 - u_0^h, v^h)|/\|v^h\|_{\mathcal{V}}^h$ for the different cases are given in this Chapter 6. We formulate the Lemmas for the different cases of variational crimes introduced above. Again, we use the general notations for a^h, A^h, f^h, Q'^h, P^h as introduced in (4.68), (4.69), (4.61), (4.66), (4.108), see Notation 4.1. We impose the following conditions.

Condition 5.1 *Let $a(\cdot,\cdot)$ and $f(\cdot)$ be continuous and admit a solution*

$$u_0 \in \mathcal{U}_b \text{ for } a(u_0, v) = f(v) \ \forall\ v \in \mathcal{V}_b; \tag{5.2}$$

choose the $\mathcal{U}_b, \mathcal{V}_b$ and the corresponding boundary conditions $Bu, B_1 v$ as in (3.19), (3.20). Let $a(\cdot,\cdot)$ and $f(\cdot)$ or $a^h(\cdot,\cdot)$ and $f^h(\cdot)$, the bilinear and linear forms in (4.68), be uniformly continuous on $\mathcal{U}_b^h \times \mathcal{V}_b^h$ and $\mathcal{V}_b^h$. Let

$$a^h(u,v) = a(u,v), f^h(v) = f(v) \quad \forall u \in \mathcal{U}_b \text{ or } v \in \mathcal{V}_b. \tag{5.3}$$

Let $a^h(\cdot,\cdot)$ satisfy a uniform inf–sup– condition, see Theorem 2.1.7 and Chapter 6, with

$$\epsilon^h \geq \epsilon > 0 \text{ on } \mathcal{U}_b^h \times \mathcal{V}_b^h, \tag{5.4}$$

This implies a unique discrete solution

$$u_0^h \in \mathcal{U}_b^h \text{ for } a^h(u_0^h, v^h) = f(v^h) \ \forall\ v^h \in \mathcal{V}_b^h \tag{5.5}$$

and it implies the stability of A^h.

Remark 5.1.1. According to this Condition and the general notations, the above bilinear forms, operators and norms can be used in the weak and strong forms. This is one reason for calling the results in this Chapter the *generalized Strang Lemmas*. This is particularly important for the last Lemma 5.1.4. Its strong version includes the results necessary for collocation methods. Mind that for the weak and strong forms we have the following combination of bilinear forms, Banach spaces and norms. We explicitly only formulate the case of Dirichlet boundary conditions.

$$\begin{aligned} a(\cdot,\cdot) : (\mathcal{U}_b \ \times \mathcal{V}_b) &= (H_0^1(\Omega) \ \times H_0^1(\Omega)) \to \mathbb{R}, \\ \|u\|_{\mathcal{U}_b} = \|u\|_{\mathcal{V}_b} &= \|u\|_{H^1(\Omega)} \text{for the weak, and} \\ a_s(\cdot,\cdot) : (\mathcal{U}_b \ \times \mathcal{V}_b) &= (H^2(\Omega) \cup H_0^1(\Omega) \ \times L^2(\Omega)) \to \mathbb{R}, \\ \|u\|_{\mathcal{U}_b} = \|u\|_{H^2(\Omega)}, \|v\|_{\mathcal{V}_b} &= \|v\|_{L^2(\Omega)} \text{for the strong form.} \end{aligned} \tag{5.6}$$

The following Lemmas generalize the Cea Lemma 3.3.1, see [18, 17].

Lemma 5.1.2. Violated boundary conditions: *Let $\mathcal{U}_b^h \not\subset \mathcal{U}_b, \mathcal{V}_b^h \not\subset \mathcal{V}_b$, however $\mathcal{U}_b^h \subset \mathcal{U}, \mathcal{V}_b^h \subset \mathcal{V}$, hence $u^h \in \mathcal{U}_b^h, v^h \in \mathcal{V}_b^h$ violate the boundary conditions $Bu^h = 0$ and $B_1 v^h = 0$, see (3.22), but are continuous. Let $a(\cdot,\cdot)$ satisfy (5.2), (5.3), (5.4), and $u_0 \in \mathcal{U}_b$ and $u_0^h \in \mathcal{U}_b^h$ be an exact and the approximate solution, defined by*

$$a(u_0, v) = f(v) \ \forall\ v \in \mathcal{V}_b, \quad a(u_0^h, v^h) = f(v^h) \ \forall\ v^h \in \mathcal{V}_b^h. \tag{5.7}$$

Then, with C independent of h,

$$\|u_0 - u_0^h\|_{\mathcal{U}}^h \leq C \left(\inf_{u^h \in \mathcal{U}_b^h} \|u_0 - u^h\|_{\mathcal{U}} + \sup_{0 \neq v^h \in \mathcal{V}_b^h} \frac{|a(u_0 - u_0^h, v^h)|}{\|v^h\|_{\mathcal{V}}} \right). \tag{5.8}$$

As a consequence of (5.2), (5.5), the second term on the right-hand side of (5.8) would be zero if $\mathcal{U}_b^h \subseteq \mathcal{U}_b, \mathcal{V}_b^h \subseteq \mathcal{V}_b$. Therefore, it measures the effect of $\mathcal{U}_b^h \not\subseteq \mathcal{U}_b, \ \mathcal{V}_b^h \not\subseteq \mathcal{V}_b$.

Proof. For any $u^h \in \mathcal{U}_b^h$,

$$\begin{aligned}
\|u_0 - u_0^h\|_{\mathcal{U}} &\le \|u_0 - u^h\|_{\mathcal{U}} + \|u^h - u_0^h\|_{\mathcal{U}} \text{ (triangle inequality)} \\
&\le \|u_0 - u^h\|_{\mathcal{U}} + \frac{1}{\epsilon} \sup_{v^h \in \mathcal{V}_b^h \setminus \{0\}} \frac{|a(u^h - u_0^h, v^h)|}{\|v^h\|_{\mathcal{V}}} \text{ in (2.12)} \\
&= \|u_0 - u^h\|_{\mathcal{U}} + \frac{1}{\epsilon} \sup_{v^h \in \mathcal{V}_b^h \setminus \{0\}} \frac{|a(u^h - u_0, v^h) + a(u_0 - u_0^h, v^h)|}{\|v^h\|_{\mathcal{V}}} \\
&\le \|u_0 - u^h\|_{\mathcal{U}} + \frac{1}{\epsilon} \sup_{v^h \in \mathcal{V}_b^h \setminus \{0\}} \frac{|a(u^h - u_0, v^h)|}{\|v^h\|_{\mathcal{V}}} \qquad (5.9) \\
&\quad + \frac{1}{\epsilon} \sup_{v^h \in \mathcal{V}_b^h \setminus \{0\}} \frac{|a(u_0 - u_0^h, v^h)|}{\|v^h\|_{\mathcal{V}}} \text{ (triangle inequality)}
\end{aligned}$$

$$\begin{aligned}
&\le \|u_0 - u^h\|_{\mathcal{U}} + \frac{C}{\epsilon}\|u^h - u_0\|_{\mathcal{U}} \text{ (continuity in } \mathcal{U}_b^h \times \mathcal{V}_b^h) \\
&\quad + \frac{1}{\epsilon} \sup_{v^h \in \mathcal{V}_b^h \setminus \{0\}} \frac{|a(u_0 - u_0^h, v^h)|}{\|v^h\|_{\mathcal{V}}} \qquad (5.10) \\
&= \left(1 + \frac{C}{\epsilon}\right) \|u_0 - u^h\|_{\mathcal{U}} + \frac{1}{\epsilon} \sup_{v^h \in \mathcal{V}_b^h \setminus \{0\}} \frac{|a(u_0 - u_0^h, v^h)|}{\|v^h\|_{\mathcal{V}}} \quad \blacksquare
\end{aligned}$$

Note that, by continuity,

$$\frac{|a(u_0 - u_0^h, v^h)|}{\|v^h\|_{\mathcal{V}}} \le C\|u_0 - u_0^h\|_{\mathcal{U}} \qquad (5.11)$$

so that

$$\|u_0 - u_0^h\|_{\mathcal{U}} \ge \frac{1}{C} \sup_{v^h \in \mathcal{V}_b^h \setminus \{0\}} \frac{|a(u_0 - u_0^h, v^h)|}{\|v^h\|_{\mathcal{V}}}. \qquad (5.12)$$

Combining (5.12) and (5.9)

$$\begin{aligned}
\max \quad & \left\{ \frac{1}{C} \sup_{v^h \in \mathcal{V}_b^h \setminus \{0\}} \frac{|a(u_0 - u_0^h, v^h)|}{\|v^h\|_{\mathcal{V}}}, \inf_{u^h \in \mathcal{U}_b^h} \|u_0 - u^h\|_{\mathcal{U}} \right\} \\
&\le \|u_0 - u_0^h\|_{\mathcal{U}} \\
&\le \left(1 + \frac{C}{\epsilon}\right) \inf_{u^h \in \mathcal{U}_b^h} \|u_0 - u^h\|_{\mathcal{U}} \qquad (5.13) \\
&\quad + \frac{1}{\epsilon} \sup_{v^h \in \mathcal{V}_b^h \setminus \{0\}} \frac{|a(u_0 - u_0^h, v^h)|}{\|v^h\|_{\mathcal{V}}}.
\end{aligned}$$

Inequality (5.13) indicates that the term $\inf_{u^h \in \mathcal{U}_b^h} \|u_0 - u^h\|_{\mathcal{U}}$, together with $\sup_{v^h \in \mathcal{V}_b^h \setminus \{0\}} |a(u_0 - u_0^h, v^h)| / \|v^h\|_{\mathcal{V}}$ truly reflect the size of the discretization error $\|u_0 - u_0^h\|_{\mathcal{U}}$. Similar estimates are possible for the following Lemmas.

The next Lemma uses the extensions $a^h(\cdot,\cdot), f^h(\cdot)$ of $a(\cdot,\cdot) : \mathcal{U} \times \mathcal{V} \to \mathbb{R}, f(\cdot) : \mathcal{V} \to \mathbb{R}$.

Lemma 5.1.3. Violated continuity: *Let (5.2), (5.3), (5.4) be satisfied, $\mathcal{U}_b^h \not\subset \mathcal{U}, \mathcal{V}_b^h \not\subset \mathcal{V}$ violate the continuity conditions, (4.15), $u_0 \in \mathcal{U}_b$ and $u_0^h \in \mathcal{U}_b^h$ be an exact and the approximate solutions, defined by*

$$a(u_0, v) = f(v) \ \forall\ v \in \mathcal{V}_b \text{ and } a^h(u_0^h, v^h) = f(v^h) \quad \forall\ v^h \in \mathcal{V}_b^h,$$

resp. (5.3). Then, with C, independent of h,

$$\|u_0 - u_0^h\|_{\mathcal{U}}^h \le C \left(\inf_{u^h \in \mathcal{U}_b^h} \|u_0 - u^h\|_{\mathcal{U}}^h + \sup_{0 \ne v^h \in \mathcal{V}_b^h} \frac{|a^h(u_0 - u_0^h, v^h)|}{\|v^h\|_{\mathcal{V}}^h} \right) \quad (5.14)$$

In Lemmas 5.1.2 and 5.1.3 it might be, e.g., for isoparametric FEs, that in (5.7) and (5.14) the $a(u_0, v^h)$ and $a^h(u_0, v^h)$ have to be replaced by a modified $a^h(u_0, v^h)$, which might again not be defined $\forall\ v^h \in \mathcal{V}_b^h$. Then we may have to replace $|a(u_0 - u_0^h, v^h)|$ and $|a^h(u_0 - u_0^h, v^h)|$ by the modified $|a^h(z^h - u_0^h, v^h)|$ with small $\|z^h - u_0\|_{\mathcal{U}}^h$, see Section 6.4. Without loosing accuracy, for z^h the $P^h u_0$ sometimes may be chosen, whenever it satisfies $\|P^h u_0 - u_0\| \le C \|u_0 - u_0^h\|_{\mathcal{U}}^h \to 0$, e.g. realized as the interpolation of u_0, see the following proof. Again, as a consequence of (5.2), (5.5), (5.4), the second term on the right-hand side of (5.14) would be zero if $\mathcal{U}_b^h \subseteq \mathcal{U}, \mathcal{V}_b^h \subseteq \mathcal{V}$. Therefore, it measures the effect of $\mathcal{U}_b^h \not\subseteq \mathcal{U}$, $\mathcal{V}_b^h \not\subseteq \mathcal{V}$.

Proof We present the proof based on the good approximation $z^h \in \mathcal{U}_b^h$, e.g. with $\|z^h - u_0\|_{\mathcal{U}}^h \le 2 \inf_{u^h \in \mathcal{U}_b^h} \|u_0 - u^h\|_{\mathcal{U}}^h$. This z^h replaces u_0 which can be used in the standard situation. As a consequence of the inf − sup − condition for a^h we find for any $u^h \in \mathcal{U}_b^h, v^h \in \mathcal{V}_b^h$, e.g. for $u^h = z^h$,

$$\|u_0^h - z^h\|_{\mathcal{U}}^h \le \epsilon^{-1} \sup_{0 \ne v^h \in \mathcal{V}_b^h} \frac{|a^h(u_0^h - z^h, v^h)|}{\|v^h\|_{\mathcal{V}}^h}. \quad (5.15)$$

The triangle inequality $\|u_0 - u_0^h\|_{\mathcal{U}}^h \le \|z^h - u_0\|_{\mathcal{U}}^h + \|u_0 - u_0^h\|_{\mathcal{U}}^h$ and arguments analogous to the proof of Lemma 5.1.2 yield the desired result. ∎

For more complicated situations the exact scalar products $a(u^h, v^h), f(v^h)$ or $a^h(u^h, v^h), f^h(v^h)$ are not computable. So we introduce, e.g. quadrature approximations $\tilde{a}^h(u^h, v^h)$ and $\tilde{f}^h(v^h)$ defined on $\mathcal{U}^h \times \mathcal{V}^h$ and $\mathcal{V}^h$, but only for smooth enough $(u, v) \in \mathcal{U} \times \mathcal{V}$ and $v \in \mathcal{V}$.

Lemma 5.1.4. Quadrature Approximations and Collocation: *Let (5.2), (5.3), (5.4) be satisfied for $\tilde{a}^h(\cdot,\cdot)$ and $\tilde{f}^h(\cdot)$. This will be a consequence of the estimates for quadrature errors in Chapter 6, see Theorem 6.5.3. Let u_0 and u_0^h be an exact and the approximate solution defined by*

$$a(u_0, v) = f(v) \ \forall\ v \in \mathcal{V}_b \text{ and } \tilde{a}^h(u_0^h, v^h) = \tilde{f}^h(v^h) \ \ \forall\ v^h \in \mathcal{V}_b^h.$$

Finally, let $\tilde{a}^h(u_0, v^h)$ *be defined* $\forall\ v^h \in \mathcal{V}_b^h$, *or replace it , as above, by* $\tilde{a}^h(z^h, v^h)$ *with small* $\|z^h - u_0\|_{\mathcal{U}}^h$. *Then we estimate, with* C *independent of* h,

$$\|u_0 - u_0^h\|_{\mathcal{U}}^h \le C \left(\inf_{u^h \in \mathcal{U}_b^h} \|u_0 - u^h\|_{\mathcal{U}}^h + \sup_{0 \ne v^h \in \mathcal{V}_b^h} \frac{|\tilde{a}^h(u_0^h - u_0, v^h)|}{\|v^h\|_{\mathcal{V}}^h} \right) \text{ or}$$

$$\|u_0 - u_0^h\|_{\mathcal{U}}^h \le C \left(\inf_{u^h \in \mathcal{U}_b^h} \|u_0 - u^h\|_{\mathcal{U}}^h + \sup_{0 \ne v^h \in \mathcal{V}_b^h} \frac{|a^h(u_0, v^h) - \tilde{a}^h(u_0, v^h)|}{\|v^h\|_{\mathcal{V}}^h} + \sup_{0 \ne v^h \in \mathcal{V}_b^h} \frac{|f^h(v^h) - \tilde{f}^h(v^h)|}{\|v^h\|_{\mathcal{V}}^h} \right). \tag{5.16}$$

Remark 5.1.1 shows that these results for the strong formulation yield the estimates for the collocation case.

The first inequality emphazises the *variational consistency error,* the second the *quadrature errors* for $a^h(u_0, v^h)$ compared to $\tilde{a}^h(u_0, v^h)$ and $f^h(v^h)$ and $\tilde{f}^h(v^h)$.

Proof Here we argue with a smooth enough u_0 directly, assuming that $\tilde{a}^h(u_0, v^h)$ is defined and $a^h(u_0, v^h) = f^h(v^h)$ for all $v^h \in \mathcal{V}_b^h$, see (4.20). Then

$$\begin{aligned}
\epsilon \|u_0^h - u^h\|_{\mathcal{U}}^h &\le \sup_{0 \ne v^h \in \mathcal{V}_b^h} |\tilde{a}^h(u_0^h - u^h, v^h)| / \|v^h\|_{\mathcal{V}}^h \\
&= \sup_{0 \ne v^h \in \mathcal{V}_b^h} |\tilde{a}^h(u_0^h - u_0 + u_0 - u^h, v^h)| / \|v^h\|_{\mathcal{V}}^h \\
&= \sup_{0 \ne v^h \in \mathcal{V}_b^h} |\tilde{a}^h(u_0^h - u_0, v^h) + \tilde{a}^h(u_0 - u^h, v^h)| / \|v^h\|_{\mathcal{V}}^h \ \text{ by (5.5)} \\
&\le \sup_{0 \ne v^h \in \mathcal{V}_b^h} |a^h(u_0, v^h) - \tilde{a}^h(u_0, v^h) + \tilde{f}^h(v^h) - f^h(v^h)| / \|v^h\|_{\mathcal{V}}^h \\
&\quad + \sup_{0 \ne v^h \in \mathcal{V}_b^h} |\tilde{a}^h(u_0 - u^h, v^h)| / \|v^h\|_{\mathcal{V}}^h. \qquad (5.17)
\end{aligned}$$

With the uniformly bounded $\tilde{a}^h(\cdot, \cdot)$, we can draw two different conclusions from the fourth or the third line above

$$\begin{aligned}
\epsilon \|u_0^h - u^h\|_{\mathcal{U}}^h &\le \sup_{0 \ne v^h \in \mathcal{V}_b^h} |a^h(u_0, v^h) - \tilde{a}^h(u_0, v^h)| / \|v^h\|_{\mathcal{V}}^h \\
&\quad + \sup_{0 \ne v^h \in \mathcal{V}_b^h} |f(v^h) - \tilde{f}^h(v^h)| / \|v^h\| + C \|u_0 - u^h\|_{\mathcal{U}}^h \qquad (5.18)
\end{aligned}$$

or

$$\begin{aligned}
\epsilon \|u_0^h - u^h\|_{\mathcal{U}}^h &\le \sup_{0 \ne v^h \in \mathcal{V}_b^h} |\tilde{a}^h(u_0^h - u_0, v^h)| / \|v^h\|_{\mathcal{V}}^h \qquad (5.19) \\
&\quad + \sup_{0 \ne v^h \in \mathcal{V}_b^h} |f(v^h) - \tilde{f}^h(v^h)| / \|v^h\| + C \|u_0 - u^h\|_{\mathcal{U}}^h.
\end{aligned}$$

Again the triangle inequality and the same additional terms as in the proof of Lemma 5.1.2, see (5.9), yield (5.16). The z^h modification proceeds analogously to the last proof. ■

6. Consistency and Coercivity for Variational Crimes

Klaus Klaus Ausnahmen fuer starke resultate nachtragen
We want to emphasize that this Chapter mainly discusses estimates for the *weak bilinear forms*. Exceptions are formulated in the coercivity and convergence results of each Section, Theorem 6.1.1 and Sections 6.5 and 6.6. The strong relation between these approximate bilinear forms and collocation enforces the study of the strong forms as well. Instead of the neutral Notation 4.1, we use in this Chapter the $a(\cdot,\cdot)$, $a^h(\cdot,\cdot)$, $\tilde{a}^h(\cdot,\cdot)$, a.s.o. always for the weak bilinear forms. The strong forms are denoted as $a_s(\cdot,\cdot)$, $a_s^h(\cdot,\cdot)$, $\tilde{a}_s^h(\cdot,\cdot)$, similarly for the operators, $A, A^h, \tilde{A}^h$ versus $A_s, A_s^h, \tilde{A}_s^h$.

The Lemmas in the last Chapter have been proved under the Condition 5.1. They show that we need the (uniform) continuity of $f(\cdot)$, $f^h(\cdot)$, $\tilde{f}^h(\cdot)$ and $a(\cdot,\cdot)$, $a^h(\cdot,\cdot)$, $\tilde{a}^h(\cdot,\cdot)$ on $\mathcal{V}_b$, $\mathcal{V}_b^h$ and $\mathcal{U}_b \times \mathcal{V}_b$, $\mathcal{U}_b^h \times \mathcal{V}_b^h$ and the $\inf - \sup -$ conditions for $a(\cdot,\cdot)$, $a^h(\cdot,\cdot)$, $\tilde{a}^h(\cdot,\cdot)$ on $\mathcal{U}_b \times \mathcal{V}_b$, $\mathcal{U}_b^h \times \mathcal{V}_b^h$. Again, we have to distinguish the different bilinear forms. Additionally, we need estimates for $a(u_0 - u_0^h, v^h)$, $a^h(u_0 - u_0^h, v^h)$ and the quadrature or approximation errors $a^h(u_0, v^h) - \tilde{a}^h(u_0, v^h)$, $f^h(v^h) - \tilde{f}^h(v^h)$. To this end, we have to combine approximation errors, inverse estimates for the norms of $u^h \in \mathcal{U}_b^h$ or $\in \mathcal{U}^h$ and $v^h \in \mathcal{V}_b^h$ or $\in \mathcal{V}^h$ and results concerning the extension operators E^h. To apply the preceding results to nonconforming FE and spectral methods, we estimate in this Chapter the consistency errors. Furthermore, we prove the stability for coercive bilinear forms here and for the general case in Chapter 7. It is important to realize that the proofs for the stability for coercive bilinear forms sometimes needs the consistency results, but not vice versa.

The long Chapter is organized as follows: We will prove, for coercive bilinear forms, the discrete coercivity and the inf–sup– condition in Section 6.1. We extend this to the strong bilinear forms in Section **??**. We estimate classical and variational consistency for the different non conforming FEs for linear operators, as introduced in Chapter 5. This will be done in Section 6.2 - 6.5. Starting in Section 6.2 we assume $n = 2$, from Section 6.5 ff. we again allow $n \geq 2$, unless the preceding results for $n = 2$ are required. This will be indicated. In Section 4.4 we had compared variational with classical consistency errors introduced, e.g., in [58, 57]. For the standard approximations, both definitions yield consistency simultaneously. We study here both types

of consistency errors. Variational consistency represents the more familiar approach in the FE community. We present the results for linear second order elliptic equations for FE and spectral methods. In spectral methods without domain decomposition techniques the smoothness of the approximating spaces is guaranteed. The boundary conditions in the finitely (but sufficiently) many points imply exact boundary conditions. This is correct even for some of the domain decompositions. For some recently suggested techniques, still strongly developing, this is partially violated. In this case the techniques in the following Sections will have to be appropriately modified for spectral methods. Therefor, we discuss for spectral methods as variational crimes only quadrature approximations.

6.1 Discrete Coercivity and Gauss Quadrature

Here, we prove the $\mathcal{U}_b^h$ coercivity for $\mathcal{U}_b$ coercive bilinear forms $a(\cdot,\cdot)$ and show the desired discrete inf–sup– condition for the case $\mathcal{U}_b^h \neq \mathcal{V}_b^h$, and thus stability for its discrete counterpart. Furthermore, we include the necessary Gauss quadrature results.

6.1.1 Uniform Continuity, Discrete Coercivity and Consistency

We start by repeating the relation for the exact and the approximate bilinear forms. In fact, both are defined independently of the imposed boundary conditions (3.22). However, they live on subspaces defined by the boundary conditions. They become relevant, when the weak operator, e.g., $A : H_0^1(\Omega) \to H^{-1}(\Omega)$, see (3.19), and the strong operator $A_s : H^2(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega)$ have to be compared. In (3.15) - (3.19) we had found

$$
\begin{aligned}
a(u,v) = < Au, v >_{H^{-1}(\Omega)\times H_0^1(\Omega)} &= \int_\Omega \Big(\sum_{i,j=1}^n a_{ij}\partial_i u\ \partial_j v + \sum_{j=1}^n a_{0j}\ u\ \partial_j v \\
&\quad + \sum_{i=1}^n a_{i0}(\partial_i\ u)\ v + a_{00} u\ v\Big) dx \qquad (6.1)\\
&= \int_\Omega (A_s\ u) v dx + \int_{\partial\Omega} (B_a\ u) v ds \ \forall\ u \in \mathcal{U}_b,\ v \in \mathcal{V}_b \text{ with}\\
B_a\ u &= \sum_{i,j=1}^n \nu_j a_{ij}\partial_i\ u + \sum_{j=1}^n \nu_j a_{0j}\ u \text{ and}\\
B_a\ u &= \partial u/\partial\nu \text{ for the special case}\\
&\quad a_{ij} = c_i\delta_{i,j}\ \forall i,j = 0,\ldots,n, \text{e.g., } A_s u = -\Delta u.
\end{aligned}
$$

It is obvious, that $a(u^h, v^h)$ is defined $\forall\ u^h \in \mathcal{U}_b^h \subset \mathcal{U},\ v^h \in \mathcal{V}_b^h \subset \mathcal{V}$ if only the boundary conditions are violated. If additionally, the continuity

conditions are violated, we had to modify the $a(\cdot,\cdot), \|\cdot\|_{\mathcal{U}}$, sometimes even $f(\cdot)$. To obtain the relation between $a^h(\cdot,\cdot)$ and $a_s^h(\cdot,\cdot)$, we essentially only have to replace $\partial u^h/\partial \nu_e$ in (4.27), (??) by $B_a u^h$. So we find

$$\begin{aligned}
a^h(u^h,v^h) \;=\; &< A^h u^h, v^h >_{\mathcal{V}_b'^h \times \mathcal{V}_b^h} \;=\; \sum_{T\in\mathcal{T}^h} \Big(\int_T \sum_{i,j=1}^n a_{ij}\partial_i\, u^h \partial_j\, v^h \\
&+ \sum_{i=1}^n a_{i0}\,(\partial_i\, u^h) v^h + \sum_{j=1}^n a_{0j}\, u^h\,(\partial_j\, v^h) + a_{00} u^h v^h dx \Big) \\
=:\; &\sum_{T\in\mathcal{T}^h} a_T(u^h,v^h) = \sum_{T\in\mathcal{T}^h} \Big(\int_T (A_s\, u^h) v^h dx + \\
&+ \sum_{e\in T} \int_e \big(v_l^h \left[B_a u^h\right] + [v^h] \frac{\partial u_r^h}{\partial \nu_e} \big) ds + \int_{\partial\Omega} v_l^h B_a u^h\, ds \\
=\; &a_s^h(u^h,v^h) + \sum_{e\in T} \int_e \big(v_l^h \left[B_a u^h\right] + [v^h] \frac{\partial u_r^h}{\partial \nu_e} \big) ds \\
&+ \int_{\partial\Omega} v_l^h B_a u^h\, ds \;\forall\; u^h \in \mathcal{U}_b^h,\; v^h \in \mathcal{V}_b^h \\
&\text{with } a^h(u,v) = a(u,v) \;\forall\; u \in \mathcal{U}_b,\; v \in \mathcal{V}_b.
\end{aligned} \tag{6.2}$$

Now a few comments are necessary to distinguish the different cases. Depending upon the violated boundary or continuity conditions several of the above terms disappear:

$$\begin{aligned}
&a^h(u^h,v^h) - a_s^h(u^h,v^h) = \sum_{e\in T} \int_e v_l^h \left[B_a u^h\right] ds \\
&\qquad \text{for conforming FEs} \\
&a^h(u^h,v^h) - a_s^h(u^h,v^h) = \sum_{e\in T} \int_e \big(v_l^h \left[B_a u^h\right] + [v^h] \frac{\partial u_r^h}{\partial \nu_e} \big) ds \\
&\qquad \text{for violated continuity} \\
&a^h(u^h,v^h) - a_s^h(u^h,v^h) = \int_{\partial\Omega} v_l^h B_a u^h\, ds \\
&\qquad \text{for violated boundary conditions} \\
&\qquad\qquad \forall\; u^h \in \mathcal{U}_b^h,\; v^h \in \mathcal{V}_b^h.
\end{aligned} \tag{6.3}$$

For different violations we have to add the corresponding terms. If the weak and strong solutions $u_0^h \in \mathcal{U}_b^h$ exist, they are determined by

$$a^h(u_0^h,v^h) = < f, v^h >_{\mathcal{V}_b'^h \times \mathcal{V}_b^h} \text{ and } a_s^h(u_0^h,v^h) = (f,v^h)_{\mathcal{V}_b'^h \times \mathcal{V}_b^h} \;\forall v^h \in \mathcal{V}_b^h \tag{6.4}$$

Correspondingly we introduce as in (4.42), (4.45) the

$$\text{quadrature formulas } \tilde{a}^h(u^h,v^h), \tilde{a}_s^h(u^h,v^h) \;\forall\; u^h \in \mathcal{U}_b^h,\; v^h \in \mathcal{V}_b^h, \tag{6.5}$$

see (6.75), (6.76) below.

To prove the *uniform continuity* for the linear and bilinear forms is left a straight forward exercise. So, we assume that there exists a constant, C, independent of h, such that

$$\begin{aligned} &\text{for } \|u\|_{\mathcal{U}} = \|u\|_{H^1(\Omega)},\ \|v\|_{\mathcal{V}} = \|v\|_{H^1(\Omega)} && (6.6)\\ &|a(u,v)| \le C\|u\|_{\mathcal{U}} \cdot \|v\|_{\mathcal{V}} \text{ and } |f(v)| \le C\|v\|_{\mathcal{V}} \ \forall\, u \in \mathcal{U}_b, v \in \mathcal{V}_b \text{ and}\\ &|a(u^h,v^h)|, \quad |a^h(u^h,v^h)| \le C\|u^h\|^h_{\mathcal{U}} \cdot \|v^h\|^h_{\mathcal{V}} \ \forall\, u^h \in \mathcal{U}^h_b, v^h \in \mathcal{V}^h_b\\ &|\tilde{a}^h(u^h,v^h)| \le C\|u^h\|^h_{\mathcal{U}_s} \cdot \|v^h\|^h_{\mathcal{V}_s} \ \forall\, u^h \in \mathcal{U}^h_b \subset \mathcal{U}^h_s, v^h \in \mathcal{V}^h_b \subset \mathcal{V}^h_s, && (6.7)\\ &|f(v^h)|, \quad |f^h(v^h)|,\ |\tilde{f}(v^h)| \le C\|v^h\|^h_{\mathcal{V}} \ \forall\, u^h \in \mathcal{U}^h_b, v^h \in \mathcal{V}^h_b. \end{aligned}$$

This condition is satisfied for $u \in \mathcal{U}, v \in \mathcal{V}, u^h \in \mathcal{U}^h, v^h \in \mathcal{V}^h$ as well, for $|\tilde{a}^h(u,v)|$ only if u, v are smooth enough to allow point evaluations. This is indicated by $\mathcal{U}^h_b \subset \mathcal{U}^h_s$, $\mathcal{V}^h_b \subset \mathcal{V}^h_s$, $\|u^h\|^h_{\mathcal{U}_s} \cdot \|v^h\|^h_{\mathcal{V}_s}$ in (6.7), see (6.75), (6.76).

The coercivity or $\inf - \sup -$ results for the cases of Dirichlet or natural boundary conditions and $a(\cdot,\cdot)$, $a^h(\cdot,\cdot)$ are presented in the following Theorem. This applies to violated boundary conditions and continuity as discussed in Sections 6.2, 6.3 and with the necessary machinery in Sections 6.4 and6.5 as well. For the general case of non-coercive $a(\cdot,\cdot)$ the stability proof is delayed to Chapter 7.

Theorem 6.1.1. *Let* A, $a(\cdot,\cdot)$, $a^h(\cdot,\cdot)$ *and* Dirichlet *or* natural boundary conditions *be given as in (6.2), (6.2) and let* h *be small enough. Then a* $\mathcal{U}_b$*-coercive* $a(\cdot,\cdot)$ *implies, for* $\mathcal{U}^h_b = \mathcal{V}^h_b$, *again* $\mathcal{U}^h_b$*-coercivity for* $a(\cdot,\cdot)$ *and* $a^h(\cdot,\cdot)$. *So there exists a constant* $\alpha > 0$ *s.t., e.g.,*

$$a(u^h,u^h) \ge \alpha(\|u^h\|^h_{H^1(\Omega)})^2 \ \forall\ \mathcal{U}^h_b. \tag{6.8}$$

For $\mathcal{U}^h_b \ne \mathcal{V}^h_b$, *the uniform inf–sup– condition is satisfied. So, there exist* $\epsilon, \epsilon' > 0$ *such that both inequalities*

$$\begin{aligned} &\sup_{0 \ne v^h \in \mathcal{V}^h_b} |a(u^h,v^h)|/\|v^h\|^h_{\mathcal{V}} \ge \epsilon \|u^h\|^h_{\mathcal{U}} \quad \forall u^h \in \mathcal{U}^h_b, \quad \textit{and}\\ &\sup_{0 \ne u^h \in \mathcal{U}^h_b} |a(u^h,v^h)|/\|u^h\|^h_{\mathcal{U}} \ge \epsilon' \|v^h\|^h_{\mathcal{V}} \quad \forall v^h \in \mathcal{V}^h_b \end{aligned} \tag{6.9}$$

are satisfied. Thus, the $\mathcal{U}_b$*-coercivity of* $a(\cdot,\cdot)$ *implies the unique existence of the exact and discrete solutions* u_0 *and* u^h_0.

Except the extensions to $\tilde{a}^h(\cdot,\cdot)$, $a^h_s(\cdot,\cdot)$, $\tilde{a}^h_s(\cdot,\cdot)$, instead of $a(\cdot,\cdot)$ and to non-conforming instead of conforming FEs this Theorem is identical with Theorem 3.3.3 with changes in the Proof. Furthermore

Theorem 6.1.2. *These results in Theorem 6.1.1, and similarly in Theorems 6.2.5, 6.3.1, 6.5.2, remain correct for the approximations* $\tilde{a}^h(\cdot,\cdot)$, *and*

for the strong bilinear forms, $a_s(\cdot,\cdot)$, $a_s^h(\cdot,\cdot)$, $\tilde{a}_s^h(\cdot,\cdot)$, introduced in Chapter 4. This statement requires the condition that the quadrature approximations and enough points on every edge, e, have been chosen to guarantee quadrature errors and differences in (6.3) vanishing with h. These conditions are discussed in the following Sections.

Proof: For non conforming FEs we start with Dirichlet conditions and need the anti-crime operator, E^h, see Theorem 2.6.3. This $E^h : \mathcal{U}^h \to \mathcal{U}$ satisfies, for the case of Dirichlet boundary conditions,

$$\|E^h I^h u - u\|_{W_q^1(\Omega)} \le Ch^{(n-1)/q}\|u\|_{W_q^1(\Omega)} \;\forall\; u \in W_q^2(\Omega). \tag{6.10}$$

By the triangle inequality and the transition from u^h to $E^h u^h$ we immediately obtain, for $\mathcal{U}_b^h = \mathcal{V}_b^h \not\subset \mathcal{U}_b$ the $\mathcal{U}_b^h-$ coercivity of $a(\cdot,\cdot)$.

To prove the inf–sup– conditions let $\mathcal{U}_b^h \not\subset \mathcal{U}_b$, $\mathcal{V}_b^h \not\subset \mathcal{V}_b$, $\mathcal{U}_b^h \neq \mathcal{V}_b^h$, $u := E^h u^h$ and choose $u_s \in H^2(\Omega)$ with $\|u_s - u\|_{H^1(\Omega)} < Ch^{(n-1)/2}\|u_s\|_{H^2(\Omega)}$. With the interpolation operator $I_{\mathcal{V}}^h : \mathcal{V}_b \to \mathcal{V}_b^h$ let $v_u^h := I_{\mathcal{V}}^h u_s \in \mathcal{V}_b^h$. Then Theorems 2.5.1 and 2.6.3, see (6.10), imply $\|v_u^h - u^h\|_{\mathcal{U}}^h \le Ch^{(n-1)/q}\cdot\|u^h\|_{H^1(\Omega)}^h$. Therefore for small enough h

$$\sup_{0\neq v^h\in\mathcal{V}_b^h} |a(u^h,v^h)|/\|v^h\|_{\mathcal{V}}^h \ge |a(u^h,v_u^h)|/\|v_u^h\|_{\mathcal{V}}^h \ge \alpha\|u^h\|_{\mathcal{U}}^h/2.$$

For the $\sup_{\{0\neq u^h\in\mathcal{U}_b^h\}} |a(u^h,v^h)|/\|u^h\|_{\mathcal{U}}^h$ we start instead with v^h and $v := E^h v^h, v_S \approx v, u_v^h := I^h v_s$.

To prove the $\mathcal{U}_b^h-$ coercivity of $a(\cdot,\cdot)$ for natural boundary conditions, we use a modification of the extension operator in Theorem 2.6.3. Instead of aiming for $E^h u^h = 0$ on $\partial\Omega$ our goal is here $B_a E^h u^h = 0$ on $\partial\Omega$, obtainable with a slight modification of the above proof. Again this E^h satisfies Theorem 2.6.3 and allows the same proof as above.

To prove the $\mathcal{U}_b^h-$coercivity of $a^h(\cdot,\cdot)$ we combine $a(u,v) = a^h(u,v)\;\; \forall u \in \mathcal{U}_b,\; v \in \mathcal{V}_b$ again with E^h and choose $u = E_s u^h \in H^2(\Omega)$ with $\|u - u_s\|_{H^1(\Omega)} \le Ch^{1/2}\|u_s\|_{H^2(\Omega)}$. This implies

$$\begin{aligned}
&|a^h(u^h,v^h) - a(E_s^h u^h, E_s^h, v^h)| \\
&\quad\le |a^h(u^h - E_s^h u^h, v^h)| + |a^h(E_s^h u^h, v^h - E_s^h v^h)| \\
&\qquad \text{by} \quad a^h(\cdot,\cdot) = a(\cdot,\cdot) \text{ on } \mathcal{U}_b \times \mathcal{V}_b \\
&\quad\le M\big(\|u^h - E_s^h u^h\|_{H^1(\Omega)}^h\|v^h\|_{H^1(\Omega)}^h + \|E_s^h u^h\|_{H^1(\Omega)}^h\|v^h - E_s^h v^h\|_{H^1(\Omega)}^h\big) \\
&\qquad \text{by } a^h(\cdot,\cdot) \text{ continuous} \\
&\quad\le Ch^{(n-1)/2}\|u^h\|_{H^1(\Omega)}^h\cdot\|v^h\|_{H^1(\Omega)}^h.
\end{aligned}$$

With $|a(E_s^h u^h, E_s^h u^h)| \ge \alpha\|E_s^h u^h\|_{H^1(\Omega)}^2$ the $\mathcal{U}_b^h-$coercivity, and, similarly, the inf–sup– condition is proved.

For curved boundaries and isoparametric FEs, see Section 6.4, these techniques are modified to extend the above result to these cases. For the approximate $\tilde{a}^h(\cdot,\cdot)$ we combine it with the quadrature errors in Section 6.5.

Finally, we combine the results in the following Sections with (4.27), (??). They show, how the difference between the weak and the corresponding strong bilinear form in (6.3), and similarly the exact and the quadrature approximation, can be estimated. The differences are bounded by the errors for quadrature approximations, violated boundary conditions and continuity. Under the conditions for convergence in the Theorems of this Chapter, we have thus guaranteed the inf–sup– conditions for the respective bilinear forms.

This implies the stability for A^h for all these cases. Finally, the unique existence of the u_0, u_0^h is an immediate consequence of Theorems 2.1.6 and 2.1.7 and Remark 2.1.8. ■

Remark 6.1.3. In the preceding proof the density argument is essential: Instead $u \in H^1(\Omega)$ we changed to $u_s \in H^2(\Omega)$, thus allowing to apply the convergence results, e.g., in Theorems 2.5.1 and 2.6.3. This will be important for the consistency results in the following subsections and the stability results in Chapter 7.

We have to estimate the *variational consistency errors* as in Sections 3.2 and 4.1, see (4.13), and the generalizations in Definition 4.3.4. In contrast to the *classical consistency errors*, we always have to assume, for the case of variational consistency errors, that the discrete solution u_0^h exists. So, we assume the existence of the exact and approximate solutions u_0 and u_0^h of

$$a(u_0, v) = f(v) \ \forall v \in \mathcal{V}_b \text{ or } a^h(u_0^h, v^h) = f^h(v^h) \ \forall v^h \in \mathcal{V}_b^h$$

with the general notations in Notation 4.1. If the boundary conditions are violated we find for the *variational consistency error*, see (4.13),

$$a(u_0 - u_0^h, v^h) = \int_{\partial\Omega} v^h B_a u_0 ds$$
$$\text{for violated boundary conditions for } u_0 \in H^2(\Omega), \ \forall \ v^h \in \mathcal{V}_b^h. \quad (6.11)$$

This is obtained, totally analogous to (4.13) by replacing $-\Delta u + cu$ by the general operators A, A_s and B_a in (3.20), (3.26), (3.20), see (6.2), (6.3) as well. Mind that here and below the use of u_0 instead of u^h in (6.3) eliminates some terms due to $u_0 \in H^2(\Omega)$. *For violated continuity we introduce the jump along* $e = \overline{T_1} \cap \overline{T_2}$ as $[v^h] := v^h|_{T_1} - v^h|_{T_2}$, for an appropriate extension of v^h outside of Ω, see Section 6.3. Combining (6.2) with $a^h(u_0^h, v^h) = f(v^h) = a^h(u_0, v^h) \ \forall v^h \in \mathcal{V}_b^h$ by (4.43), we obtain, for the *variational consistency error* for violated continuity

$$\begin{aligned} a^h(u_0 - u_0^h, v^h) &= \sum_{T \in \mathcal{T}^h} \left[\int_T v^h (A_s u_0 - f) dx + \int_{\partial T} v^h B_a u_0 ds \right] \\ &= \sum_{e \in \mathcal{T}^h} \int_e [v^h] B_a u_0 ds \ \forall \ v^h \in \mathcal{V}_b^h \text{for violated continuity.} \end{aligned} \quad (6.12)$$

Again this is obtained from (4.29) by the same generalizations A, A_s, B_a.

Now we turn to the estimates for the consistency errors. We refer to [18, 17] and generalize some of their results to more general approximation spaces, boundary conditions and second order elliptic partial differential equations instead of their $-\Delta u = f$. As in [18, 17], we need estimates, e.g., for $|a^h(u_0 - u_0^h, v^h)|/\|v^h\|^h$, see (5.8), (5.14) and (5.16) for the cases of violated boundary conditions $\mathcal{U}_b^h \not\subset \mathcal{U}_b, \mathcal{V}_b^h \not\subset \mathcal{V}_b$, discontinuous approximation spaces $\mathcal{U}_b^h \not\subset \mathcal{U}, \mathcal{V}_b^h \not\subset \mathcal{V}$, and approximate evaluation (quadrature formulas) of $a^h(\cdot,\cdot)$ and $f(\cdot)$ or $f^h(\cdot)$, resp.

The estimates for, e.g., $|a^h(u_0 - u_0^h, v^h)|/\|v^h\|_{\mathcal{V}}^h$ and the $|a^h(u_0, v^h) - \tilde{a}^h(u_0^h,$
$v^h)|/\|v^h\|_{\mathcal{V}}^h$ have to be determined for each case of variational crimes, see Chapter 4 and 5, in detail. We consider the general case, (3.26), (3.19), starting with the classical consistency errors first. It is independent of the existence of a (unique) discrete solution, in contrast to the variational consistency errors. As in Chapter 2, see (2.34), we require

Condition 6.1 Conditions for FEMs: *Choose FEM with piecewise polynomials $\mathcal{P}$ and a maximal $\tau \geq -1$ (,with usually $\tau = -1$,) as*

$$\mathcal{P}_{m-1} \subseteq \mathcal{P} \subseteq \mathcal{P}_{m+\tau}, \text{ with } \tau \geq -1, \text{ and } \mathcal{P} \neq \mathcal{P}_{m+\tau}, \text{ for } \tau \geq 0.$$

Let the subdivision be non-degenerate, see Definition 2.4.5, s.t. a uniform χ exists with the property: For every $T \in \mathcal{T}^h$ concentric inner and outer circles D_1 and D_2 exist, with $D_1 \subset T \subset D_2$ resp., and $\operatorname{diam} D_2 / \operatorname{diam} D_1 \leq \chi$. *Every $T \in \mathcal{T}^h$ at the boundary has at most one curved side, see Figures 2.18-2.20 above. Let $\partial\Omega$ be piecewise smooth. Non smooth points of $\partial\Omega$ are used as vertices of sub triangles, compare Condition 6.2.*

Mind that for the interpolation results, which we need below, we alway need FEs satisfying (2.34). For the Doedel collocation we have to relax the strictly local definition of the interpolation operator.

6.1.2 Univariate Gauss, Gauss-Radau, and Gauss-Lobatto Quadrature Formulas

We have to distinguish in the rest of the Booklet interpolation and quadrature errors. The latter often have to be estimated for $u^h,\ v^h \in \mathcal{U}^h, \mathcal{V}^h$. So we use three types of Gauss formulas, see Lemma 6.1.4 below. They are distinguished by the requirement that either none, one or two boundary points are included in the corresponding collocation grid points. They are defined for Legendre and Chebyshev polynomials w.r.t the weight functions $w \equiv 1$ and $w = (1 - x^2)^{\frac{1}{2}}$ on the interval $[-1, 1]$, resp. We introduce three types of univariate polynomials, $P_{m'}^{\rho} \in \mathcal{P}^{m'} := \mathcal{P}_{m'}^1,\ \rho = 0, 1, 2$, based on the Legendre and Chebyshev polynomials $T_{m'} \in \mathcal{P}^{m'}$ and $P_{m'} \in \mathcal{P}^{m'}$, resp., to define quadrature formulas of the type of

Gauss: $\rho = 0$: $P^0_{m'} := P_{m'}$ or $P^0_{m'} := T_{m'}$ and for both cases
Gauss-Radau: $\rho = 1$: $P^1_{m'} := P^0_{m'} + aP^0_{m'-1}$, with a such that $P^1_{m'}(-1) = 0$,
Gauss-Lobatto: $\rho = 2$: $P^2_{m'} := P^0_{m'} + aP^0_{m'-1} + bP^0_{m'-2}$, with a, b such that $P^2_{m'}(\pm 1) = 0$.

The m' collocation points $y^\rho_j := y_j$, $j \in \{0, \ldots, m'-1\}$ for the quadrature formulas are defined as the roots of the polynomials

$$P^\rho_{m'}(y^\rho_j) = 0 \text{ for } j \in \{0, \ldots, m'-1\}, \quad \rho = 0, 1, 2. \tag{6.14}$$

The corresponding weights are then computed via the Lagrange elementary polynomials, p^ρ_j, as

$$p^\rho_j \in \mathcal{P}^{m'} : p^\rho_j(y^\rho_l) = \delta_{j,l} \ \forall\ 0 \ \leq \ j, l \ \leq m'-1 \text{ as}$$
$$w_j := w^\rho_j := \int_{-1}^{1} p^\rho_j(x) w(x) dx, \ \rho = 0, 1, 2,$$

with the above weight function, w. The following Proposition is proved in many textbooks, e.g., in [23]:

Proposition 6.1.4. *Let $-1 \leq y_0 < \cdots < y_{m'-1} \leq +1$ be the above roots of the $P^\rho_{m'}$, for $m' \in \mathbb{N}$, $\rho = 0, 1, 2$, and $w_j := w^\rho_j$ the corresponding weights. Then the quadrature approximation $q^{m'}_w(u)$ is defined as*

$$q^{m'}_w(u) := \sum_{j=0}^{m'-1} u(y^\rho_j) w^\rho_j \approx \int_{-1}^{1} u(x) w(x) dx, \tag{6.15}$$
$$q^{m'}(u) := q^{m'}_{w \equiv 1}(u), \quad \textit{for } u \in C[-1, 1] \textit{ and satisfies}$$
$$q^{m'}(\rho) := \sum_{j=0}^{m'-1} p(y^\rho_j) w^\rho_j = \int_{-1}^{1} p(x) w(x) dx, \ \textit{for all } p \in \mathcal{P}^{2m'-1-\rho}.$$

This implies for scalar products, $(u, v)_w$, w.r.t the weight function w that

$$(u, v)_w = (u, v)^{m'}_w := q^{m'}_w(u \cdot v) = \sum_{j=0}^{m'-1} (u \cdot v)(y^\rho_j) w^\rho_j$$
$$\textit{for } u \in \mathcal{P}^{m'-\rho}, v \in \mathcal{P}^{m'-1} \textit{ and } \rho = 0, 1, 2. \tag{6.16}$$

So $(u, v)^{m'}_w$ exactly reproduces scalar products for $u, v \in \mathcal{P}^{m'-1}$ for $\rho = 0, 1$, and for $\rho = 2$ only for one of the u (or v) $\in \mathcal{P}^{m'-2}$.

6.2 Violated Boundary Conditions

In this and in Sections 6.3, 6.4, we consider no spectral methods, but FEs, since only they do violate boundary conditions and continuity. Furthermore

we assume $n = 2$ in this and Sections 6.3 and 6.4. The discrete solution u_0^h is obtained from the original $a(\cdot,\cdot)$ as

$$\begin{aligned} a(u^h, v^h) \text{ and } f(v^h) \text{ in (6.2) are defined } \forall u^h \in \mathcal{U}_b^h,\ v^h \in \mathcal{V}_b^h, \\ \text{determine } u_0^h \in \mathcal{U}_b^h \text{ from } a(u_0^h, v^h) = f(v^h) \ \forall v^h \in \mathcal{V}_b^h. \end{aligned} \tag{6.17}$$

6.2.1 Consistency Estimates for Violated Boundary Conditions

As in Section 4.1 we have to estimate for (6.11), see (4.13), (3.20), (3.22), (4.10), the variational consistency error

$$a(u_0 - u_0^h, v^h) = \int_{\partial\Omega} v^h B_a u_0 ds = \int_{\partial\Omega} (\sum_{i,j=1}^{2} \nu_j a_{ij} \partial_i\, u_0 + \sum_{j=1}^{2} \nu_j a_{0j}\, u_0) v^h ds, \quad u_0 \in H^2(\Omega)\ \forall\ v^h \in \mathcal{V}_b^h. \tag{6.18}$$

This remains correct, if u_0 and u_0^h are replaced by an arbitrary $u \in H^2(\Omega)$ and u^h with $A^h u^h = Q'^h f_1 := Q'^h Au$, see (4.122). For the cases $\mathcal{V}_b = H^1(\Omega)$ and $\mathcal{V}_b = H_0^1(\Omega)$, see (3.19), (3.22), (4.114),(4.115), we have

$$\begin{aligned} &Q'^h : H^{-1}(\Omega) \to \mathcal{V}_b^{h'} : \langle Q'^h f, v^h \rangle_{H^{-1}(\Omega)\times H^1(\Omega)} = \langle f, v^h \rangle_{H^{-1}(\Omega)\times H^1(\Omega)} \\ &A^h := Q'^h A|_{\mathcal{U}_b^h} : \mathcal{U}_b^h \to \mathcal{V}_b^{h'} \text{ s.t. } \langle A^h u^h, v^h \rangle_{H^{-1}(\Omega)\times \mathcal{V}_b^h} = a(u^h, v^h)\ \forall v^h \in \mathcal{V}_b^h. \end{aligned} \tag{6.19}$$

To determine the *classical consistency error* we choose the interpolation operator, $P^h := I^h$, see (2.31) and

$$\text{test } A^h I^h u - Q'^h Au \quad \text{with } v^h \in \mathcal{V}_b^h \text{ for an arbitrary fixed } u \in \mathcal{U}_b.$$

We have to distinguish the two types of boundary conditions in (3.22) and combine (3.19), (4.13) to obtain

Theorem 6.2.1. *Let Q'^h and A^h be defined in (6.19) and choose $u \in \mathcal{U}_b \cap \in H^2(\Omega)$. Then the violated boundary conditions for the test functions $v^h \in \mathcal{V}_b^h$ are reflected in the* variational consistency error *and the* classical consistency error *as*

$$a(u - u^h, v^h) = \int_{\partial\Omega} v^h B_a u ds \text{ for } A^h u^h = Q'^h Au\ \forall\ v^h \in \mathcal{V}_b^h, \tag{6.20}$$

and man koennte auf $\int_{\partial\Omega} v^h B_a u ds$ verzichten und alles scheint korrekt zu bleiben [1]

[1] testing with $v^h \in \mathcal{V}_b^h$ always introduces the violated boundary or continuity terms. A direct estimation of $\langle A^h I^h u - Q'^h Au, v^h \rangle_{\mathcal{V}_b^{h'}\times\mathcal{V}_b^h}$ would introduce some type of $\int_{\partial\Omega} v^h \hat{B}_a (I^h u - u) ds$. This is more complicated to study than the terms in (6.21). The same is true for violated continuity.

$$\langle A^h I^h u - Q'^h A u, v^h \rangle_{\mathcal{V}_b^{h\prime} \times \mathcal{V}_b^h} = a(I^h u - u, v^h) - \int_{\partial\Omega} v^h B_a u ds. \quad (6.21)$$

They can be estimated as

$$\sup_{0 \neq v^h \in \mathcal{V}_b^h} \{|a(u - u^h, v^h)| / \|v^h\|_{\mathcal{V}}^h\} \leq \sup_{0 \neq v^h \in \mathcal{V}_b^h} |\int_{\partial\Omega} v^h B_a u ds| / \|v^h\|_{\mathcal{V}}^h, (6.22)$$

and

$$\begin{aligned} ||A^h I^h u - Q'^h A u\|_{\mathcal{V}_b^{h\prime}} &\leq \sup_{0 \neq v^h \in \mathcal{V}_b^h} \left(|a(I^h u - u, v^h)| + |\textstyle\int_{\partial\Omega} v^h B_a u ds|\right) / \|v^h\|_{\mathcal{V}}^h \\ &\leq C \|I^h u - u\|_{H^1(\Omega)} + \sup_{0 \neq v^h \in \mathcal{V}_b^h} |\textstyle\int_{\partial\Omega} v^h B_a u ds| / \|v^h\|_{\mathcal{V}}^h. \quad (6.23) \end{aligned}$$

Theorem 6.2.2. Natural boundary conditions: *Under the conditions of the last Theorem, the boundary terms in (6.22), (6.23) vanish, if u or the exact solution u_0 satisfies the natural boundary condition, hence*

$$a(u - u^h, v^h) = \int_{\partial\Omega} v^h B_a u ds = 0 \quad \textit{for} \quad B_a u|_{\partial\Omega} = 0 \ \forall \ v^h \in \mathcal{V}_b^h \subset \mathcal{V} (6.24)$$

For general $v^h \in \mathcal{V}_b^h$ it cannot vanish for Dirichlet boundary conditions (for u and $\mathcal{V}_b$), since the v^h do not satisfy this boundary condition exactly.

For satisfied natural boundary conditions, $B_a u|_{\partial\Omega} = 0$, the variational consistency error *is always* 0, *the classical consistency error is a $\mathcal{O}(\|I^h u - u\|_{H^1(\Omega)})$. This implies the same error estimate as for conforming FEMs,*

$$\|u_0 - u_0^h\|_{H^1(\Omega)} \leq C h^{m-1} \|u_0\|_{H^m(\Omega)}. \quad (6.25)$$

Estimates for $\|I^h u - u\|_{H^1(\Omega)}$ are known from Theorem 2.5.1. So, for natural boundary conditions the classical consistency results are equivalent to those for interpolation errors. Therefore, we are left with the problem to find

6.2.2 Consistency Estimates for Violated Dirichlet Conditions

For $v^h \in \mathcal{V}_b^h \not\subset \mathcal{V}_b$, hence $v^h|_{\partial\Omega} \not\equiv 0$, an estimate depends on how well v^h approximates $v^h|_{\partial\Omega} = 0$ and how this implies $|\int_{\partial\Omega} v^h B_a \ u_0 ds|$ to be small. The following steps involve quadrature errors and inverse estimates. First, we use Proposition 6.1.4 to estimate $|\int_{\partial\Omega} f ds - \sum_{P_j^\rho \in \partial\Omega} w_j(h) f(P_j^\rho)|$ for $f \in W_\infty^{m'}(\Omega)$. Here, the $P_j^\rho \in \partial\Omega$ indicate all quadrature points along all edges on $\partial\Omega$, or on a section of $\partial\Omega$ or its approximation. The $w_j(h)$ indicate the corresponding quadrature weights. For FE methods we only consider Legendre polynomials with $w \equiv 1$. Rescaling $[-1, 1]$ to $[0, h]$, a modified Proposition 6.1.4 reads as, see e.g., [18],

Proposition 6.2.3. *For all* $m' \in \mathbb{N}$ *there exists* $C = C_{m'}$ $\forall 0 < h < 1$, *s.t.*

$$\left|\int_0^h f(x)dx - h\sum_{j=0}^{m'-1} w_j f(h\xi_j)\right| \le C_{m'}\ h^{2m'+1-\rho}\|f^{(2m'-\rho)}\|_{L^\infty(0,h)}$$
$$\forall\ \ f \in W_\infty^{2m'-\rho}[0,h] \cap C[0,h], \quad (6.26)$$

with $\rho = 0, 1, 2$, *for Gauss, Gauss-Radau and Gauss-Lobatto quadrature points* $h\xi_j := h(y_j^\rho + 1)/2$ *and* y_j^ρ *the zeros of the Legendre polynomials* $P_{m'}^\rho$ *of degree* m', *see (6.14).*

These Gauss-type points are used for conforming and non conforming FEs, resp. Now we have to specify the $v^h \in \mathcal{V}_b^h$. We choose for each straight or curved boundary edge e a parameterization

$$e = \{x_b(s) : s \in [s_e, s_e + h_e], s = \text{ arc length}\}. \quad (6.27)$$

With the ξ_j in (6.26) we define the boundary nodes to be

$$x_b(P_j^e) := x_b(s_e + h_e\xi_j) \in e, j = 0, 1, \ldots, m' - 1$$
$$\forall\ \ e \in \mathcal{T}^h \text{ with } \sharp\{e \cap \partial\Omega\} = 2. \quad (6.28)$$

This allows to define

$$\mathcal{V}_b^h := \{v : \Omega \to \mathbb{R} : v|_T \in \mathcal{P}_{m'}\ \forall\ \ T \in \mathcal{T}^h,$$
$$v \text{ vanishes at all boundary nodes } x_b(P_j^e)\} \quad (6.29)$$

We apply (6.26) to the special case of $f = (v^h B_a u) \circ x_b$ with the above parameterization x_b for $e \subset \partial\Omega$ and $v^h B_a u \in W_\infty^{k+1}(\Omega)$. We use (2.4) for $p = \infty$, (6.2) and the product and chain rule to estimate

$$\|((v^h B_a\ u) \circ x_b)^{(k)}\|_{L^\infty(\partial\Omega)} \le C_{x_b}\|u\|_{W_\infty^{k+1}(\Omega)}\|v^h\|_{W_\infty^{\min\{k, m+\tau\}}(\Omega)}$$
$$\forall u \in W_\infty^{k+2}(\Omega),\ v^h \in \mathcal{V}^h\ , \quad (6.30)$$

with v^h locally of degree $\le m+\tau$ and C_{x_b} depending on powers of order $\le k$ of different derivatives $x_b^{(j)}, j = 0, \ldots, k$, and $j = 0, 1$ for straight edges. Finally, inverse estimates yield for (6.26) and Condition 6.1, see e.g., Proposition 2.5.3 and Theorem 2.5.4 with $n = 2$,

$$\|v^h\|_{W_p^j(\Omega)} \le C\ \ h^{l-j-(2/q-2/p)}\|v^h\|_{W_q^l(\Omega)} \text{ for FEMs and}$$
$$0 \le l \le j \le m + \tau, 1 \le q \le p \le \infty\ \ \forall v^h \in \mathcal{U}^h \text{ or } \in \mathcal{V}^h. \quad (6.31)$$

We need repeatedly the estimates for the interpolation errors: If Condition 6.1 and $v \in W_p^m(\Omega)$ are satisfied, then

$$\|v - I^h v\|^h_{W^s_p(\Omega)} \leq C\ h^{m-s}|v|_{W^m_p(\Omega)} \text{ and}$$
$$\|v - I^h v\|_{W^s_p(T)} \leq C\ (\ \text{diam}\ T)^{m-s}|v|_{W^m_p(T)} \text{ for } 0 \leq s \leq m. \quad (6.32)$$

A *quasi-uniform* subdivision in Condition 6.1 is necessary, if $1 \leq p, q \leq \infty$ should be intended in (6.31). We summarize the conditions for the boundary and the quadrature formulas defined by their boundary points, see (6.26):

Condition 6.2 Conditions for boundary and quadrature: *Let $\partial\Omega$ be a piecewise smooth boundary. We allow a curved boundary and introduce $\partial\Omega^h$ and the points P_j along $\partial\Omega$ as in Subsection 2.7.1, see (2.87). We delay isoparametric approximation to Section 6.4. Let the segments $e \subset \overline{\partial\Omega}$, see (6.27) be defined by these P_j, P_{j+1}, see (2.7.1). Choose a quadrature approximations according to Condition 6.1, (6.14), (6.15), (6.26) along the above closed segments e of the piecewise smooth $\partial\Omega \subset \mathbb{R}^2$. Impose Dirichlet boundary conditions in m' points $x_e(P^\rho_i) \in \partial\Omega \cap e$, $\rho = 0, 1, 2$, see (6.28), or close to $\partial\Omega \cap e$. These $P^\rho_i, i = 0, \ldots, m' - 1$, are defined as Gauss, Gauss-Radau and Gauss-Lobatto points according to the chosen quadrature approximation along the segment (, parametrized by x_b w.r.t. the arc-length, see (6.27)). These points, $x_b(P^\rho_i)$, possibly except $P_j = P^\rho_0$, $P_{j+1} = P^\rho_{m'-1}$, are additional to the above. For polygonal $\partial\Omega$ we have $m' \leq m + \tau$, for curved $\partial\Omega$ cases with $m' > m + \tau$ are not excluded.*

In fact, we have to distinguish two cases:

Polygonal $\partial\Omega$: The piecewise linear x_b implies $v^h \in \mathcal{P}_{m+\tau}$ and yields $v^h \circ x_b \in \mathcal{P}_{m+\tau}$. Now $m' > m + \tau$ boundary points with $v^h(x_b(P_j)) = 0, j = 0, \ldots, m' - 1$ implies $v^h \equiv 0, v^h \circ x_b \equiv 0$, hence the Dirichlet boundary conditions (D.b.cs) are satisfied exactly. Hence violated D.b.cs. imply $m' \leq m + \tau$. Whenever $p|_e \in \mathcal{P}^1_{m+\tau} \backslash \mathcal{P}^1_{m+\tau-1}$. This $m' \leq m + \tau$ is equivalent to violated D.b.cs., this is particularly correct for $\tau = -1$. Now, we assume the same number of functionals on every edge to define the FEs, see (2.50). Then violated boundary conditions for

$$m' \leq m + \tau \ \text{ and } p|_e \in \mathcal{P}^1_{m+\tau} \backslash \mathcal{P}^1_{m+\tau-1} \ \forall \ \ e \in \overline{T}$$

are equivalent to violated continuity. This is discussed in Section 6.3 with $a(\cdot,\cdot)$ replaced by $a^h(\cdot,\cdot)$. In fact it is possible for large enough $m(\geq 6)$, see Proposition 2.3.1, to have $p|_e, p|_{e_2} p|_{e_3} \in \mathcal{P}^1_{m+\tau} \backslash \mathcal{P}^1_{m+\tau-1}$, for the three edges, $e, e_1, e_2 \subset \overline{T}$ and still have $p \in \mathcal{P} \overset{\subset}{\neq} \mathcal{P}_{m+\tau}$.

Curved $\partial\Omega$ requires to discuss, for $e \in \partial\Omega$, every specific situation separately. In particular, $m' > m + \tau$ does not allow to conclude $v^h \circ x_b|_e \equiv 0$ as above. Thus, it does not exclude violated Dirichlet boundary conditions. Hence, see (2.50), violated boundary conditions do not necessarily imply violated continuity.

In this Subsection, we only study Dirichlet conditions and we need for the proof of the next Theorem the following Lemma 6.2.4. It discusses the influence of boundary values onto the necessary estimates.

Lemma 6.2.4. *Under the Condition 6.2 and for arbitrary $u \in W_\infty^{2m'+2-\rho}(\Omega)$, we choose A, $a(\cdot,\cdot)$ and Dirichlet boundary conditions as in (6.2). Furthermore, we define $\mathcal{V}_b^h$ as in (6.29). Then there exists a constant $C_\rho = C(\rho, m, \chi)$ s.t.*

$$|\int_{\partial\Omega} v^h B_a u \, ds| \le C_\rho \; h^\mu \|u\|_{W_\infty^{2m'+1-\rho}(\Omega)} \|v^h\|_{H^1(\Omega)}^h$$

$$\forall v^h \in \mathcal{V}_b^h \cap H^1(\Omega) \not\subset H_0^1(\Omega), \text{ for} \tag{6.33}$$

$$\mu := 2m' - \rho - \min\{2m' - \rho, m + \tau\} + 1/2. \tag{6.34}$$

(If $v^h \in \mathcal{V}_b^h \cap W_q^1(\Omega), 1 \le q < \infty$, then (6.33) still holds with a slightly changed exponent of h).

Proof The (6.26), (6.30)and (6.31) with $l = 1, j = m + \tau, p = \infty, q \ge 2$ are applied to every edge $e \in \overline{T} \in \mathcal{T}^h$ with length h_e and $\sharp(e \cap \partial\Omega) = 2$. With the parameterization x_b for e, see (6.27), this yields see (6.29), (6.30), (2.4) and with $q^{m'}(v^h B_a u) \circ x_b = 0 \; \forall v^h \in \mathcal{V}_b^h$ and $\mu' := 2m' - \rho$, we find

$$\begin{aligned}
|\int_e v^h B_a u| &= |\int_{s_e}^{s_e+h_e} (v^h B_a u) \circ x_b ds| \\
&\le C h_e^{\mu'+1} \|u\|_{W_\infty^{\mu'+1}(T)} \|v^h\|_{W_\infty^{\min\{\mu', m+\tau\}}(T)} \\
&\le C h_e^{\mu'+1} h_e^{1-(\min\{\mu', m+\tau\})-2/q} \|u\|_{W_\infty^{\mu'+1}(\Omega)} \|v^h\|_{W_q^1(T)} \qquad (6.35) \\
&= C h_e^{\mu'+2-\min\{\mu', m+\tau\}-2/q} \|u\|_{W_\infty^{\mu'+1}(\Omega)} \|v^h\|_{W_q^1(T)} \text{ or} \\
&\le C h_e^{\mu'+1-\min\{\mu', m+\tau\}} \|u\|_{W_\infty^{\mu'+1}(\Omega)} \cdot \|v^h\|_{H^1(T)}, \text{ for } q = 2.
\end{aligned}$$

Summation over all edges e along $\partial\ \Omega$ yields, for $q = 2$,

$$\begin{aligned}
|\int_{\partial\ \Omega} v^h B_a u ds| &\le \sum_e |\int_e v^h B_a u ds| \qquad (6.36) \\
&\le C h^{\mu'+1-\min\{\mu', m+\tau\}-1/2} \|u\|_{W_\infty^{\mu'+1}(\Omega)} \left(\sum_e h_e^{1/2} \|v^h\|_{H^1(T)} \right) \\
&\le C h^{\mu'-\min\{\mu', m+\tau\}+1/2} \|u\|_{W_\infty^{\mu'+1}(\Omega)} \|v^h\|_{H^1(\Omega)}^h,
\end{aligned}$$

since, by the Hoelder inequality and (2.17),

$$\sum_e h_e^{1/2} \|v^h\|_{H^1(T)} \le \left(\sum_e h_e \right)^{1/2} \|v^h\|_{H^1(\Omega)}^h \text{ and } \sum_e h_e \le 2 \times (\text{arc-length of } \partial\Omega)$$

for small enough h. So we obtain (6.33) with the above μ. ■

Before we formulate the next Theorem we have to study the behavior of the complicated μ in (6.34). We have seen that

$$m' = m + \tau - k,\ k \in \mathbb{Z} \text{ is possible,} \tag{6.37}$$

with $k \geq 0$ for polyhedral $\partial\Omega$ and $k \in \mathbb{Z}$ for curved $\partial\Omega$. This shows with (6.37)

$$\min\{2m' - \rho, m + \tau\} = \begin{cases} m + \tau & \text{for } m' - \rho - k > 0, \\ 2m' - \rho & \text{for } m' - \rho - k \leq 0, \end{cases}$$

s.t. (6.37) implies

$$1/2 \leq \mu = \begin{cases} m' - \rho - k + 1/2 > 1/2 & \text{for } m' - \rho - k > 0, \\ = 1/2 & \text{for } m' - \rho - k \leq 0. \end{cases}$$

We want to estimate $\|u_0^h - u_0\|_{H^1(\Omega)}$ in Lemma 5.1.2 and the variational and classical consistency errors. We combine the conditions for u and the exponents of h in Theorem 2.5.1 $\|v - I^h v\|^h_{W_p^1(\Omega)} \leq C\ h^{m-1}|v|_{W_p^m(\Omega)}$, with those for $\inf_{u^h \in \mathcal{U}_b^h} \|u^h - u_0\|_{H^1(\Omega)}$ and in (6.33), (6.34): We need $u \in W_\infty^{\max\{m, 2m'+2-\rho\}}$ and get $h^{\min\{\mu, m-1\}}$. Since we aim for convergence of the form $\|u_0^h - u_0\|^h_{W_p^1(\Omega)} \leq C\ h^\alpha,\ \alpha > 0$, we assume $m \geq 2$. Again as above (6.37) implies

$$\max\{m, 2m' + 2 - \rho\} = \begin{cases} m & \text{for } m' + \tau + 2 - \rho - k \leq 0, \\ 2m' + 2 - \rho & \text{for } m' + \tau + 2 - \rho - k > 0. \end{cases} \tag{6.38}$$

Finally, with (6.38) and by considering the two cases $m' - \rho - k > 0$ and ≤ 0 we find, for $m \geq 2$,

$$1/2 \leq \min\{\mu, m-1\} = \begin{cases} m - 1 & \text{for } m' - \rho - k > 0,\ 3/2 + \tau - \rho - 2k > 0, \\ \mu & \text{for } m' - \rho - k > 0,\ 3/2 + \tau - \rho - 2h \leq 0, \\ 1/2 & \text{for } m' - \rho - k \leq 0. \end{cases}$$

For the proof we need Remark 6.1.3.

Theorem 6.2.5. *Under the conditions (6.2), (6.2), Condition 6.1, (6.19), and Condition 6.2 with* Dirichlet conditions, *let $u \in H^1(\Omega)$. Then $A^h = Q'^h A|_{\mathcal{U}_b^h}$ is consistent with A in u. Let u^h for $A^h u^h = Q'^h A u$ exist. For $u \in W_\infty^{\max\{m, 2m'+2-\rho\}}(\Omega)$ the* variational *and* classical consistency errors *for u vanish of orders μ and $\min\{m-1,\ \mu\}$, resp., see (6.34), (6.38). They can be estimated, with a h -independent $C_\rho = C_{(m,n,\rho,x_b)}$, by*

$$\sup_{0 \neq v^h \in \mathcal{V}_b^h} \frac{|a(u - u^h, v^h)|}{\|v^h\|^h_{H^1(\Omega)}} \leq C_\rho h^\mu \|u\|_{W_\infty^{2m'+1-\rho}(\Omega)}, \text{ and} \tag{6.39}$$

$$\|A^h I^h u - Q'^h A u\|_{\mathcal{V}_b^{h\prime}} \leq C h^{\min\{\mu, m-1\}} \|u\|_{W_\infty^{\max\{m, 2m'+2-\rho\}}}.$$

Then a $\mathcal{U}_b$-coercive $a(\cdot,\cdot)$ implies, for $\mathcal{U}_b^h = \mathcal{V}_b^h$, again a $\mathcal{U}_b^h$-coercive $a(\cdot,\cdot)$. For $\mathcal{U}_b^h \neq \mathcal{V}_b^h$ and for $a_s(\cdot,\cdot)$, the uniform discrete inf–sup– condition is valid

for $a(\cdot,\cdot)$. This implies the stability of the corresponding A^h and the existence of u^h, u_0^h.

The weak solution u_0^h converges to u_0 according to

$$\|u_0^h - u_0\|_{H^1(\Omega)} \leq Ch^{\min\{\mu,m-1\}}\|u_0\|_{W_\infty^{\max\{m,2m'+1-\rho\}}(\Omega)}. \tag{6.40}$$

These results, see Theorem 6.1.2, remain correct for the strong bilinear forms, $a_s(\cdot,\cdot)$, $a_s^h(\cdot,\cdot)$, introduced in Chapter 4. This statement requires the condition that enough points on every edge, e, have been chosen to guarantee differences in (6.3) to vanish with h. For the strong solutions (for $a_s(\cdot,\cdot)$) the (6.40) has to be modified as $\|u_0^h - u_0\|_{H^2(\Omega)} \leq Ch^{\min\{\mu,m-2\}}\|u_0\|_{W_\infty^{\max\{m+1,2m'+1-\rho\}}(\Omega)}$.

Proof Remark 6.1.3 shows that we can assume $u := u_s \in W_\infty^{\max\{m,2m'+2-\rho\}}(\Omega)$ approximating the original $u \in W_\infty^m(\Omega)$. This argument has to be repeated for the other convergence results as well, without mentioning it every time. The estimate (6.39) is an immediate consequence of (6.18) and Theorem 6.2.1, Lemma 6.2.4 and Remark 6.1.3, (6.40) is obtained either by combining the stability result in Theorem 7.2.3 with the consistency estimate in (6.39) or by a combination with Lemma 5.1.2, see (5.8). ■

6.3 Violated Continuity

Here we consider all types of non conformity caused by $\mathcal{U}^h \not\subset \mathcal{U}$ still with $n = 2$. This includes violated continuity or smoothness. Then often the boundary conditions will be violated as well. For a detailed discussion, see the discussion following Condition 6.2. Violated boundary conditions are partially included in the following discussion for polygonal $\partial\Omega$, however, it yields results of the same quality as for violated continuity. They are slightly worse than in Subsection 6.2. For curved boundary and isoparamtric FEs the techniques of Subsections 6.2 and 6.4, resp., and the results of this Section can and have to be combined.

We have seen already, that for the same number of functionals on every edge defining FEs in (2.50), for polygonal $\partial\Omega$, violated boundary conditions are equivalent to violated continuity. Again, for curved $\partial\Omega$, every specific situation has to be discussed, see Condition 6.2 ff.

For noncontin uous $\mathcal{U}_b^h \not\subset \mathcal{U}$, the $a(\cdot,\cdot), \|\cdot\|_{\mathcal{U}}$, sometimes even $f(\cdot)$, have to be generalized. We delay the case $f^h \neq f$ to Section 6.5 and assume here $f^h = f$. With the norm $\|v^h\|^h$ defined in (2.17) and the generalized a^h, compare (4.16), (6.2), we obtain with partial integration as in (3.20) and (6.2), (6.3),

$$a^h(u^h,v^h) := \sum_{T\in\mathcal{T}^h}(\int_T \sum_{i,j=1}^{2} a_{ij}\partial_i\, u^h\partial_j\, v^h + \sum_{i=1}^{2} a_{i0}\,(\partial_i\, u^h)v^h$$
$$+\sum_{j=1}^{2} a_{0j}\, u^h\,(\partial_j\, v^h) + a_{00}u^h v^h dx) \qquad (6.41)$$
$$= \sum_{T\in\mathcal{T}^h} a_T(u^h,v^h) = \sum_{T\in\mathcal{T}^h}(\int_T (A_s\, u^h)v^h dx + \int_{\partial\, T} v^h B_a\, u^h ds),$$

with B_a as in (6.2). The $a(\cdot,\cdot)$ and the generalized forms $a^h(\cdot,\cdot)$ are related, see (6.2), and the exact and discrete solutions u_0 and u_0^h are defined as

$$a(u_0,v) = f(v)\ v\in\mathcal{V},\ \text{and}\ a^h(u_0^h,v^h) = f(v^h)\ \ \forall v^h\in\mathcal{V}^h, \qquad (6.42)$$

resp. We collect the conditions for the actual case in

Condition 6.3 Conditions for interpolation points on edges: *We require Condition 6.1, (6.42) and a modified Condition 6.2: Replace the Dirichlet conditions by interpolation conditions on edges, e , in the above $P_i^\rho\in\overline{e}, i = 0,\ldots,m'-1$. This implies continuity of the FEs in these points P_i^ρ but not in Ω. Choose an $m'\le m+\tau$ s.t. $v^h\in\mathcal{V}^h$ and $v^h(P_i^\rho)=0, i=0,\ldots,m'-1$, does not imply $v^h|e\equiv 0$, hence, v^h is not continuous across e, see the discussion following Condition 6.2. If we consider a problem with* natural boundary conditions *we do not have to impose boundary conditions. If we want to include* Dirichlet boundary conditions *we additionally impose the original Condition 6.2 along the boundary $\partial\Omega$ which we assume to be polygonal in this case.*

Now, let e be a joint edge of T_1 and $T_2\in\mathcal{T}^h$. Let ν_e be the unit normal vector oriented into T_2. Define the jump along e as $[v^h] := v^h|_{T_1} - v^h|_{T_2}$.

To include violated boundary conditions once more we have to distinguish natural and Dirichlet conditions. We have seen already in Section 6.2, see Theorem 6.2.1, that for natural boundary conditions $\int_{\partial\ \Omega} v^h B_a u ds = 0\ \forall\ B_a u = 0$. So, we can restrict the discussion to the Dirichlet case: We choose, for $\overline{T}_1\cap\partial\ \Omega\supset\{P_1,P_2\}$, for the boundary edge, e., a $T_2\subset\mathbb{R}^2\backslash\Omega, e\subset\overline{T}_1\cap\overline{T}_2$ satisfying Condition 6.1. Crossing a *straight* boundary edge, e, we define an extension of v^h outside of Ω essentially as negative flip, that is: For the outer unit normal vector, ν_e, $\delta\ll h$ and a point $P\in\partial\ \Omega$ let $v^h(P+\delta\nu_e) := -v^h(P-\delta\nu_e)$. Obviously, we have for this extended v^h the relation $[v^h] = [v^h - c]\ \forall c\in\mathbb{R}$.

Combining (6.41) with $a^h(u_0^h,v^h) = f(v^h) = a^h(u_0,v^h)\ \forall v^h\in\mathcal{V}_b^h$ by (6.42), we obtain for the *variational discretization error* for u_0 or a general u, u^h with

$$A^h := Q'^h A_h|_{\mathcal{U}_b^h} \ , \ A^h u^h \ = \ Q'^h f_1, \ f_1 := Au, \ u \in H^1(\Omega),$$

$$a^h(u-u^h, v^h) = \sum_{T\in\mathcal{T}^h} \left[\int_T v^h(Au - f_1)dx + \int_{\partial\ T} v^h B_a u ds \right]$$

$$= \sum_{T\in\mathcal{T}^h} \int_{\partial\ T} v^h B_a u ds = \sum_{e\in\mathcal{T}^h} \int_e [v^h] B_a u ds. \qquad (6.43)$$

Obviously, we find for boundary edges $\partial\Omega$ for both types of boundary conditions $\int_e [v^h] B_a u ds = 2\int_e v^h B_a u ds$. So, for the above extension to $T_2 \subset \mathbb{R}^2 \backslash \Omega$, this approach allows to include the discussion of Section 6.2 as well, however it yields slightly worse results.
Now we proceed very similarly to Section 6.2. With $\langle Q'^h f, v^h\rangle_{\mathcal{V}_b^{h'}\times\mathcal{V}_b^h} = \langle Q'^h f, v^h\rangle_{H_h^{-1}(\Omega)\times H_h^1(\Omega)} = \langle f, v^h\rangle_{H_h^{-1}(\Omega)\times H_h^1(\Omega)}$ as in (6.19) we define, compare (6.19), and Notation (4.1),

$$A^h : \mathcal{U}_b^h \to \mathcal{V}_b^{h'} \text{ as } u^h \to A^h u^h = Q'^h A_h u^h \text{ s.t.} \qquad (6.44)$$
$$\langle A^h u^h, v^h\rangle_{H_h^{-1}(\Omega)\times H_h^1(\Omega)} = a^h(u^h, v^h) \ \ \forall\ v^h \in \mathcal{V}_b^h.$$

For the proof we again need Remark 6.1.3.

Theorem 6.3.1. *Let Ω and the FEs satisfy Condition 6.1, (6.42), Condition 6.3, $f \in L^2(\Omega)$. We choose $A, a(\cdot,\cdot), B_a$ and $a^h(\cdot,\cdot), A^h$ as in (6.2), (6.2) and (6.43) and assume u^h to be the discrete solution of $A^h u^h = Q'^h Au$. We consider* natural or Dirichlet boundary conditions. *In the latter case, Condition 6.2 has to be required additionally.* [2] *The* classical *and* variational consistency errors *are*

$$\langle A^h I^h u - Q'^h A_h u, v^h\rangle_{\mathcal{V}_b^{h'}\times\mathcal{V}_b^h} = a^h(I^h u - u, v^h) - \sum_{e\in\mathcal{T}^h} \int_e [v^h] B_a u ds \text{ and}$$

$$\sup \frac{|a^h(u-u^h, v^h)|}{\|v^h\|^h_{H^1(\Omega)}} = \sup \frac{|\sum_{e\in\mathcal{T}^h} \int_e [v^h] B_a u ds|}{\|v^h\|^h_{H^1(\Omega)}}, \text{ resp.}$$

For $u \in H^1(\Omega)$ the A^h is consistent with A in u. For $u \in W_\infty^{\max\{m,\ \mu'+2\}}(\Omega)$, μ' in (6.46), the classical and variational consistency errors vanish of orders $\min\{m-1,\ \mu'\}$ *and μ'. They can be estimated with h -independent $C_\rho = C_{(m,n,\rho,x_b)}$, by*

$$\|A^h I^h u - Q'^h A_h u\|_{\mathcal{V}_b^{h'}} \le C h^{m-1} |u|_{H^m(\Omega)} + C_\rho h^{\mu'} \|u\|_{H^{\mu'+2}(\Omega)}, \qquad (6.45)$$

$$\sup \frac{|a^h(u-u^h, v^h)|}{\|v^h\|^h_{H^1(\Omega)}} \le C_\rho h^{\mu'} \|u\|_{H^{\mu'+2}(\Omega)}. \text{ with}$$

$$\mu' = 2m' - 1 - \rho - (m+\tau). \qquad (6.46)$$

[2] If we want to include violated boundary conditions independent of Section 6.2, it is convenient, to require a polygonal $\partial\Omega$. Otherwise, we have to add the modified error terms for violated boundary conditions in Theorems 6.2.5 and 6.4.2.

For a $\mathcal{U}_b$-coercive $a(\cdot,\cdot)$ this implies, for $\mathcal{U}_b^h = \mathcal{V}_b^h$, again a $\mathcal{U}_b^h$-coercive $a^h(\cdot,\cdot)$. For $\mathcal{U}_b^h \neq \mathcal{V}_b^h$, the uniform discrete inf–sup– condition is valid for $a^h(\cdot,\cdot)$. This implies the stability of the corresponding A^h and the existence of u^h, u_0^h.

The weak solutions (for $a^h(\cdot,\cdot)$) converge according to

$$\|u_0^h - u_0\|_{H^1(\Omega)} \leq C h^{\min\{m-1,\mu'\}} \|u_0\|_{H^{\max\{m,\mu'+2\}}(\Omega)}. \tag{6.47}$$

These results, see Theorem 6.1.2, remain correct for the strong bilinear forms, $a_s^h(\cdot,\cdot)$, introduced in Chapter 4. This statement requires the condition that enough points on every edge, e, have been chosen to guarantee the differences in (6.3) to vanish with h. In the convergence result for the strong solutions (for $a_s^h(\cdot,\cdot)$) the (6.40) has to be modified as $\|u_0^h - u_0\|_{H^2(\Omega)} \leq C h^{\min\{\mu',m-2\}} \|u_0\|_{H^{\max\{m,\mu'+2\}}(\Omega)}$.

For the proof we need

Lemma 6.3.2. *Let T_1, T_2 be two neighboring triangles $T_1 \in \mathcal{T}^h, T_2 \in \mathcal{T}^h$ (or $\subset \mathbb{R}\backslash\Omega$) with* diam $T_i \leq h, i = 1,2$ *and a joint* straight *edge, $e \in \overline{T_1 \cup T_2}$. Furthermore, let (6.2), (6.2), and Condition 6.3, $u \in H^{\mu'+2}(\overline{T_1 \cup T_2})$, and $v^h \in \mathcal{U}_b^h$ or $\mathcal{V}_b^h$ be satisfied. Then, with μ' in (6.46),*

$$\left| \int_e [v^h] B_a \, u ds \right| \leq C_{m,n,\gamma} h^{\mu'} \left(|v^h|_{H^1(T_1)} + |v^h|_{H^1(T_2)} \right) |B_a \, u|_{H^{\mu'+1}(T_1 \cup T_2)}. \tag{6.48}$$

Proof To apply a technique as in Proposition 6.1.4 to our problem, we chose the two triangles (or rectangles) T_1, T_2 as above and let $G := \overline{T_1 \cup T_2}$ with the common edge e of length $h_e = h$. The substitution $t = h(s+1)/2$ then transforms (6.15), with $y_j = y_j^\rho, w_j = w_j^\rho$, into

$$\int_e q(t)dt = \int_0^h q(t)dt = h/2 \sum_{j=0}^{m'-1} w_j q(x_j) \quad \forall \; q \in P^{2m'-1-\rho};$$
$$x_j := h(y_j+1)/2, j = 0, \ldots, m'-1. \tag{6.49}$$

This result is only valid along a joint *straight* edge, $e \subset \overline{T_1 \cup T_2}$, since otherwise the parameterization usually destroys the polynomial structure of $q(t)$. We apply (6.26) to $w[v^h], w = B_a u$. With the above definition of $[v^h] = v^h|_{T_1} - v^h|_{T_2}$ along e we have $[v^h - c] = [v^h] \in P^{m+\tau}$ for every $c \in \mathbb{R}$. Now we study the single terms $\int_e [v^h] B_a u ds$ in (6.18). So, by (6.15), (6.49),

$$\int_e z[v^h]ds = \int_e z[v^h - c]ds = 0 \text{ for } [v^h] \in P^{m+\tau}, \; z \in P^{\mu'} \text{ with}$$
$$\mu' = 2m' - 1 - \rho - (m+\tau). \tag{6.50}$$

Hence

$$
\begin{aligned}
|\int_e w[v^h]ds| = |\int_e w[v^h - c]ds| &= |\int_e (w-z)[v^h - c]ds| \ \forall\ z \in P^{\mu'} \\
&\le \|w - z\|_{L^2(e)} \cdot \|v^h - c\|_{L^2(e)} \text{ by Cauchy-Schwarz-inequality} \\
&\le C\|w - z\|_{H^1(G)}(\|v^h - c\|_{H^1(T_1)} + \|v^h - c\|_{H^1(T_2)}) \ \forall\ z \in P^{\mu'};
\end{aligned}
$$

this last inequality is obtained from the Trace Theorem 2.1.3, for our case and $p = 2$ by

$$
\|v\|_{L^2(e)} \le C\|v\|^{1/2}_{L^2(G)}\|v\|^{1/2}_{H^1(G)} \le C\|v\|_{H^1(G)} \quad \forall v \in H^1(G).
$$

Since the v^h match in the P_i^ρ we can choose an appropriate c, s.t.

$$
(\|v^h - c\|_{H^1(T_1)} + \|v^h - c\|_{H^1(T_2)}) = (|v^h|_{H^1(T_1)} + |v^h|_{H^1(T_2)})
$$

with the usual semi norms for $H^1(T_i), i = 1, 2$. Using (6.50) and the averaged Taylor polynomials, see (2.8), $z := Q^{\mu'+1}w \in \mathcal{P}^{\mu'}$ we find,

$$
|\int_e w[v^h]ds| \le C\|w - Q^{\mu'+1}w\|_{H^1(G)}\big(|v_h|_{H^1(T_1)} + |v_h|_{H^1(T_2)}\big). \quad (6.51)
$$

This yields finally with the Bramble-Hilbert Lemma 2.1.5,

$$
\begin{aligned}
|\int_e w[v_h]ds| \le C_{m,n,\gamma}h^{\mu'}|w|_{H^{\mu'+1}(G)}\big(|v_h|_{H^1(T_1)} + |v_h|_{H^1(T_2)}\big) \\
\text{and } h = \max\{\ \text{diam } T_1, \quad \text{diam } T_2\} \ge h_e. \quad (6.52)
\end{aligned}
$$

With the above convention concerning $T_2 \subset \mathbb{R}^2 \setminus \Omega$ this result (6.52) remains valid for $T_2 \subset \mathbb{R}^2 \setminus \Omega$ as well. Now, we apply (6.52) to the case $w = B_a\, u$. We compare it, see (6.51), with $Q^{\mu'+1}B_a u \in \mathcal{P}^{\mu'}$ to verify (6.48). ■

Remark 6.3.3. Indeed, the last Lemma and its proof apply to the case of violated boundary conditions as well, if the above conditions are satisfied. We obtain worse results here, since the exponents $\mu = 2m' + 1/2 - \rho - (m + \tau) > \mu' = 2m' - 1 - \rho - (m + \tau)$. This yields better convergence for the approach in Lemma 6.2.4 and Theorem 6.2.5.

Proof (for Theorem 6.3.1) Similarly to Theorem 6.2.5 the consistency errors are immediate consequences of Lemmas 6.3.2 and 6.3.2, for Dirichlet boundary conditions. If the FEs satisfy Conditions 6.3 and 6.2 for Dirichlet boundary conditions. Then we obtain by summation over all $e \in \mathcal{T}^h$ in Condition 6.1 and with $G_e := G$ indicating the appropriate G for e,

$$
\begin{aligned}
|\sum_{e\in\mathcal{T}^h}\int_e [v^h]B_a u_0 ds| &\le Ch^{\mu'}\sum_{e\in\mathcal{T}^h}|v^h|_{H^1(G_e)}|B_a u|_{H^{\mu'+1}(G_e)} \text{ since } {}^3 \\
&\le Ch^{\mu'}\sum_{T\in\mathcal{T}^h}|v^h|_{H^1(T)}|B_a u|_{H^{\mu'+1}(T)} \text{ (by Hölder)} \\
&\le Ch^{\mu'}|v^h|^h_{H^1(\Omega)} \cdot |B_a u|_{H^{\mu'+1}(\Omega)}. \quad \blacksquare \quad (6.53)
\end{aligned}
$$

Remark 6.3.4. It is important to realize that (6.48) is strictly local with the local step size h and the locally appropriate $|B_a\ u|_{H^{\mu'+1}(G)}$ and holds for all $u \in H^{\mu'+2}(G)$. This can be used into two directions. Either locally large $B_a\ u$ and discretization errors are compensated by mesh refinements or the so called hp-methods are employed. For smooth u_0 and $B_a\ u_0$, high values of orders $p = m-1$ and large h_e, for unpleasant u_0 and $B_a\ u_0$, small orders $p = m-1$ and small h_e have to be combined. We do not want to pursue this point here any more to avoid too many technicalities. □

6.4 Isoparametric Violation of Boundary Conditions

In this Section we only consider the case of FEs in $\mathcal{P} = \mathcal{P}_{m-1}$. Motivated by Theorem 6.2.1 we only discuss *Dirichlet boundary conditions.*

For the *consistency estimates* we recall the original and approximating polygonal domains Ω and Ω^h, the functions $F^h : \Omega^h \to \Omega_c^h := F^h(\Omega^h)$, $F_c : \Omega^h \to \Omega$, and $\phi^h : \Omega \to F^h(\Omega^h) = \Omega_c^h$, $\phi^h := F^h \circ F_c^{-1}$ introduced in (2.93), (2.95) and (2.96), resp. The FEs $\mathcal{U}_b^h : \Omega^h \to \mathbb{R}$ satisfy the interpolation conditions along the boundary, see (2.94). We recall the definition of the $\hat{u}^h$, $\widehat{\mathcal{U}}_b^h$, $\check{f}, \phi^h$ in (2.97), (2.98)

$$
\begin{aligned}
&u^h\ :\ \Omega_c^h = F^h(\Omega^h) \to \mathbb{R},\ \hat{u}^h : \Omega \to \mathbb{R},\ \hat{u}^h(t) = u^h(\phi^h(t)) = (u^h \circ \phi^h)(t),\\
&\widehat{\mathcal{U}}_b^h = \{\hat{u}^h : u^h \in \mathcal{U}_b^h\}, \text{and analogous}\\
&\widehat{\mathcal{V}}_b^h\ ,\ \check{f} : \Omega_c^h \to \mathbb{R},\ \check{f}(x) := f((\phi^h)^{-1}(x)) = (f \circ (\phi^h)^{-1})(x) \text{ and}\\
&\qquad t - \phi^h(t) = \mathcal{O}(h^m) \text{ implying} \qquad (6.54)\\
&\qquad (\phi^h)' - Id_\Omega = \mathcal{O}(h^{m-1}),\ ((\phi^h)^{-1})' - Id_{F^h(\Omega^h)} = \mathcal{O}(h^{m-1})\ .
\end{aligned}
$$

For simplicity we again formulate the results for $n = 2$. However, in [41] they are presented for $n \leq 3$. Note that $\widehat{\mathcal{U}}_b^h \subset \mathcal{U}_b$, that is, the *Dirichlet boundary conditions* are *satisfied exactly* for $\hat{u}^h \in \widehat{\mathcal{U}}_b^h$ and similarly for $\widehat{\mathcal{V}}_b^h$. For the $u^h \in \mathcal{U}_b^h, v^h \in \mathcal{V}_b^h$ in (2.94), with $u^h, v^h : F^h(\Omega^h) \approx \Omega \to \mathbb{R}$ we define the approximate bilinear and linear form, norm, equation [4] and solution u_0^h as

[3] for $0 \leq a_1 \leq a_2, 0 \leq b_1 \leq b_2$ we have $(a_1 + a_2)(b_1 + b_2) \leq a_1 b_1 + 3a_2 b_2$ and every $T \in \mathcal{T}^h$ contains only a few edges, e.g., 3 for triangulations.

[4] for simplicity it is often assumed that $f \in L_2(F^h(\Omega^h))$. In this case the definition $f^h(v^h) := \int_{F^h(\Omega^h)} f(x)v^h(x)dx$ is used. We will show below, see (6.60), that replacing $|det(\phi^h)'(t)|$ by $det(\phi^h)'(t)$ in the integral in (6.55) is correct.

$$
\begin{aligned}
a^h(u^h, v^h) &= \int_{F^h(\Omega^h)} \nabla u^h(x) \cdot \nabla v^h(x) dx, \\
a^h(u_0^h, v^h) &= f^h(v^h) := \int_{F^h(\Omega^h)} \check{f}(x) v^h(x) dx \qquad (6.55) \\
&= \int_\Omega f(t) \hat{v}^h(t) det(\phi^h)'(t) dt \ \forall v^h \in \mathcal{V}_b^h, \text{with} \ \ f = \hat{f} \circ \phi^h, \\
\|u^h\|_{\mathcal{U}^h}^h &= \|u^h\|_{W_q^k(F^h(\Omega^h))}^h \ = \ \|\hat{u}^h\|_{W_q^k(\Omega)} (1 + \mathcal{O}(h^{m-1})),
\end{aligned}
$$

see (6.61) below, and replace $\mathcal{T}^h$ by $\mathcal{T}_c^h$ in (2.17); here $\nabla u^h(x) \cdot \nabla v^h(x) = (\nabla u^h(x))^T \nabla v^h(x)$. For the general case, we have correspondingly

$$
\begin{aligned}
a^h(u^h, v^h) : = \int_{F^h(\Omega^h)} (& \sum_{i,j=1}^{2} a_{ij} \partial_i \ u^h \ \partial_j \ v^h \qquad (6.56) \\
& + \sum_{j=1}^{2} a_{0j} \ u^h \ \partial_j \ v^h + \sum_{i=1}^{2} a_{i0} (\partial_i \ u^h) v^h + a_{00} u^h \ v^h \big)(x) dx.
\end{aligned}
$$

Now we define the appropriate Q'^h by

$$
\begin{aligned}
Q'^h : \quad & H^{-1}(\Omega) \to \mathcal{V}_b^{h'} : \langle Q'^h f - \check{f}, v^h \rangle_{\mathcal{V}_b^{h'} \times \mathcal{V}_b^h} \\
&= \int_{F^h(\Omega^h)} (Q'^h f - \check{f})(x) v^h(x) dx = \\
&= \int_\Omega (\widehat{Q'^h f} - f)(t) \hat{v}^h(t) det(\phi^h)'(t) dt \\
&= \langle (\widehat{Q'^h f} - f), \hat{v}^h det(\phi^h)' \rangle_{H^{-1}(\Omega) \times H_0^1(\Omega)} = 0 \ \ \forall v^h \in \mathcal{V}_b^h. \qquad (6.57)
\end{aligned}
$$

In a first step prove the *discrete coercivity*:

Theorem 6.4.1. *Let $\Omega \subset \mathbb{R}^2$ be bounded and have a piecewise smooth boundary. For the isoparametric FEs as introduced in Section 2.7.2. Use the $a^h(\cdot,\cdot), f^h(\cdot)$ and Q'^h as in (6.55) and (6.57), resp. If $a(\cdot,\cdot)$ is $\mathcal{U}_b-$coercive, see (6.8), then $a^h(\cdot,\cdot)$ is $\mathcal{U}_b^h-$coercive as well or, for $\mathcal{U}_b^h \neq \mathcal{V}_b^h$ the $a^h(\cdot,\cdot)$ satisfies the uniform discrete inf–sup– condition, compare Theorems 6.1.1 and 2.1.7.*

Proof Again we essentially restrict the presentation to the Laplace operator and use the substitution rule in

$$
\begin{aligned}
a^h(u^h, v^h) &= \int_{F^h(\Omega^h)} \nabla u^h(x) \cdot \nabla v^h(x) (dx) \qquad (6.58) \\
&= \int_\Omega \nabla u^h(\phi^h(t)) \cdot \nabla v^h(\phi^h(t)) \det(\phi^h)'(t) dt, \qquad (6.59)
\end{aligned}
$$

with $\nabla u^h(x) \cdot \nabla v^h(x) = (\nabla u^h(x))^T \nabla v^h(x)$. The usual $|\det(\phi^h)'(t)|$ is replaced by $\det(\phi^h)'(t)$, since

$$(\phi^h)'(t) = id + \mathcal{O}(h^{m-1}),\ \text{hence},\ \det(\phi^h)'(t) = 1 + \mathcal{O}(h^{m-1}) > 0 \tag{6.60}$$

for sufficiently small h. For the above $\hat{u}^h(t) = u^h(\phi^h(t))$ and $u^h(x)$ the derivatives ∂_i^t, ∇^t and ∂_i^x, ∇^x always are understood w.r.t. variable t for $\hat{u}^h$ and x for u^h, resp. Then we find with the unit vector e_i

$$\begin{aligned}
\partial_i \hat{u}^h(t) = \partial_i^t \hat{u}^h(t) &= \partial(u^h(\phi^h(t)))/\partial t_i = \nabla^x u^h(\phi^h(t))\partial_i \phi^h(t) \\
&= \nabla^x u^h(x)(\phi^h)'(t)e_i \ \text{and} \\
\nabla \hat{u}^h(t) = \nabla^t \hat{u}^h(t) &= (\nabla^x u^h)(\phi^h(t))(\phi^h)'(t) = \nabla^x u^h(x)(\phi^h)'(t) \ \text{or} \\
\partial_i u^h(x) = \partial_i^x u^h(x) &= \partial(\hat{u}^h(\phi^h)^{-1}(x))/\partial x_i \\
&= \nabla^t \hat{u}^h(t)\partial_i^x\big((\phi^h)^{-1}(x)\big) \\
&= \nabla^t \hat{u}^h(t)((\phi^h)^{-1})'(x)e_i \ \text{and} \\
\nabla u^h(\phi^h(t)) = \nabla^x u^h(\phi^h(t)) &= \nabla^t(\hat{u}(\phi^h)^{-1}(x))((\phi^h)^{-1})'(x) \\
&= \nabla^t \hat{u}(t)((\phi^h)^{-1})'(x)
\end{aligned} \tag{6.61}$$

Applied to the above (6.59) we find with (2.96)

$$\begin{aligned}
a^h(u^h, v^h) &= \int_\Omega \big((\phi^h)'(t)^T\big)^{-1}\nabla \hat{u}^h(t) \cdot \nabla \hat{v}^h(t)((\phi^h)'(t))^{-1} \det(\phi^h)'(t)dt \\
&= a(\hat{u}^h, \hat{v}^h) + \mathcal{O}(h^{m-1})\|\hat{u}^h\|_{H^1(\Omega)} \cdot \|\hat{v}^h\|_{H^1(\Omega)}.
\end{aligned} \tag{6.62}$$

Now, by (6.60), $\|\hat{v}^h\|_{H^1(\Omega)} = \|v^h\|_{\mathcal{V}^h}^h(1+\mathcal{O}(h^{m-1}))$ and similarly for $\|\hat{u}^h\|_{H^1(\Omega)}$, so with (6.62) and $\hat{\mathcal{U}}_b^h \subset \mathcal{U}_b$, the $\mathcal{U}_b$ coercivity of $a(\hat{u}^h, \hat{u}^h)$ implies the $\mathcal{U}_b^h$ coercivity of $a^h(u^h, u^h)$. Hence, there exists a unique solution u_0^h for $a^h(u_0^h, v^h) = f^h(v^h)\ \ \forall\, v^h \in \mathcal{V}_b^h$, see (6.55). For the case $\mathcal{U}_b^h \neq \mathcal{V}_b^h$ we proceed as in Theorem 6.1.1.

For the general case in (6.56) we find, for smooth enough $a_{ij},\ a_{i0},\ a_{0j},\ a_{00}$,

$$\begin{aligned}
a^h(u^h, v^h) = \int_\Omega &\sum_{i,j=1}^{2} a_{ij}(t)\nabla \hat{u}^h(t)\partial_i\ \big((\phi^h)^{-1}(x)\big) \cdot \nabla \hat{v}(t)\partial_j\big((\phi^h)^{-1}(x)\big) \\
&+ \sum_{j=1}^{2} a_{0j}(t)\ \hat{u}^h(t)\nabla \hat{v}^h(t)\ \partial_j\big((\phi^h)^{-1}(x)\big) \\
&+ \big(\sum_{i=1}^{2} a_{i0}(t)\nabla \hat{u}^h(t)\partial_i\big((\phi^h)^{-1}(x)\big) + a_{00}(t) \cdot \hat{u}^h(t)\big)\ \cdot\ \hat{v}^h(t)dt \\
= a(\hat{u}^h, \hat{v}^h) &+ \mathcal{O}(h^{m-1})\|\hat{u}^h\|_{H^1(\Omega)} \cdot \|\hat{v}^h\|_{H^1(\Omega)}.
\end{aligned} \tag{6.63}$$

An immediate consequence is

$$\sup_{0 \neq v^h \in \mathcal{V}_b^h} |a^h(u^h, v^h) - a(\hat{u}^h, \hat{v}^h)|\ /\ \|v^h\|_{H^1(\Omega)} \leq Ch^{m-1}\|u\|_{H^m(\Omega)}, \tag{6.64}$$

which we use in the next proof again. □

As second step we estimate the *consistency errors*:

Theorem 6.4.2. *Let $\Omega \subset \mathbb{R}^2$ be bounded and have a piecewise smooth boundary. We use the isoparametric FEs and $a^h(\cdot,\cdot), f^h(\cdot)$ as introduced in Section 2.7.2. Choose Q'^h as in (6.55) and (6.57), resp. and $A^h = Q'^h A_h|_{\mathcal{U}_b^h}$, with*

$$< A^h u^h, v^h >_{H^{-1}(F^h(\Omega^h))\times H^1(F^h(\Omega^h))} = a^h(u^h, v^h) \quad \forall\, u^h \in \mathcal{U}_b^h, v^h \in \mathcal{V}_b^h.$$

For an arbitrary u, let the discrete solution u^h with $f = Au$ and $A^h u^h = Q'^h Au$, resp., exist and assume that A is boundedly invertible [5].

For $u \in H^1(\Omega)$ the A^h is consistent with A in u. For $u \in H^m(\Omega)$, $m \geq 1$, the classical *and* variational consistency errors *for u vanish of order $m-1$ and* [6] *can be estimated, with h independent C, by*

$$\sup_{0\neq v^h \in \mathcal{V}_b^h} |a^h(u^h, v^h) - a(u, \hat{v}^h)| / \|v^h\|_{H^1(\Omega)} \leq C h^{m-1} \|u\|_{H^1(\Omega)}, \quad (6.65)$$

$$\|A^h I^h u - Q'^h Au\|_{\mathcal{V}_b^{h\prime}} \leq C h^{m-1} \|u\|_{H^m(\Omega)}, \quad (6.66)$$

$$\sup_{0\neq v^h \in \mathcal{V}_b^h} |f^h(v^h) - \check{f}(\hat{v}^h)| / \|v^h\|_{H^1(\Omega)} \leq C h^{m-1} \|f\|_{H^{-1}(\Omega)}$$

$$\leq C h^{m-1} \|u\|_{H^1(\Omega)} \quad (6.67)$$

(The invertability condition for A can be avoided, if $\|u\|_{H^1(\Omega)}$ is replaced by $\|f\|_{H^{-1}(\Omega)}$ with $f = Au$.)

Let $a(\cdot,\cdot)$ be coercive, u_0 be the exact solution of $Au_0 = f$ and u_0^h the discrete solution of (6.55). Then, for sufficiently small h,

$$\|u_0 - \hat{u}_0^h\|_{H^1(\Omega)} \leq C h^{m-1} \left(\|u_0\|_{H^m(\Omega)} + \|u_0\|_{W_\infty^1(\Omega)} \right).$$

Proof The relations (6.63) and (6.64) imply

$$\sup_{v^h \in \mathcal{V}_b^h \setminus \{0\}} \frac{|a^h(u^h, v^h) - a(\hat{u}^h, \hat{v}^h)|}{\|v^h\|_{H^1(\Omega)}} \leq C h^{m-1} \|\hat{u}^h\|_{H^1(\Omega)}, \text{ hence } (6.65)(6.68)$$

The error term for f can be estimated similarly:

[5] this condition can be avoided, if in the following estimates the $\|u\|_{H^m(\Omega)}$ is replaced by $\|f\|_{H^{m-2}(\Omega)}$

[6] if this $f \in L_2(F^h(\Omega^h)) \cap L_2(\Omega)$ then $f^h(v^h) = \int_{F^h(\Omega^h)} f(x) v^h(x) dx$ can be used. We find $\left| (f, \hat{v}^h) - \int_{F^h(\Omega^h)} f v^h dt \right| = \left| \int_\Omega \left(f(t) - f(\phi^h(t)) \det(\phi^h)'(t) \right) \hat{v}^h(t) dt \right|$ $\leq \left| \int_\Omega \left(f(t) - f(\phi^h(t)) \right) \det(\phi^h)'(t) \hat{v}^h(t) dt \right| + \left| \int_\Omega f(t)(1 - \det(\phi^h)'(t)) \hat{v}^h(t) dt \right|$ $\leq C \left(h^m \|f\|_{W_\infty^1(\Omega)} + h^{m-1} \|f\|_{L^2(\Omega)} \right) \|v^h\|_{L^2(\Omega)} \leq$ $C \left(h^m \|u_0\|_{W_\infty^2(\Omega)} + h^{m-1} \|u_0\|_{H^1(\Omega)} \right) \|v^h\|_{L^2(\Omega)}$ for $f = Au_0$.

$$\begin{aligned}\left|(f,\hat{v}^h)-\int_{F^h(\Omega^h)}\check{f}v^h dx\right| &= \left|\int_\Omega \left(f(t)-f(t)\det((\phi^h)^{-1})'(t)\right)\hat{v}^h(t)dt\right| \\ &\le \left|\int_\Omega f(t)(1-\det((\phi^h)^{-1})'(t))\hat{v}^h(t)dt\right| \\ &\le Ch^{m-1}\|f\|_{H^{-1}(\Omega)}\|\hat{v}^h\|_{L^2(\Omega)} \qquad (6.69)\\ &\le Ch^{m-1}\|u\|_{H^1(\Omega)}\|v^h\|_{L^2(\Omega)} \text{ for } f=Au.\end{aligned}$$

Similarly, to (6.18) we have to consider $a^h(u-u^h,v^h)$ with $A^hu^h = Q'^hAu$. However, the $a^h(u,v^h)$ in (6.55) is not defined. So we use, see Lemma 5.1.3, the $z^h := I^hu$ and the corresponding $\hat{z}^h \in \hat{\mathcal{U}}_b^h$ with $\|\hat{z}^h-u\|_{H^1(\Omega)} \le Ch^{m-1}\|u\|_{H^m(\Omega)}$, see Theorem 2.7.1, and estimate instead $a^h(u^h,v^h) - a^h(z^h,v^h)$. For the case (6.55) we find by (6.62), and again with $f=Au$,

$$\begin{aligned}a^h(u^h-I^hu,v^h) &= a^h(u^h,v^h)-\textstyle\int_\Omega((\phi^h)'(t)^T)^{-1}\nabla\hat{z}^h(t)\\ &\quad\cdot\nabla\hat{v}^h(t)((\phi^h)'(t))^{-1}\det(\phi^h)'(t)dt\\ &= \textstyle\int_{F^h(\Omega^h)}\check{f}v^hdx-\int_\Omega\nabla u(t)\cdot\nabla\hat{v}^h(t)dt\\ &\quad+\mathcal{O}(h^{m-1})\|u\|_{H^m(\Omega)}\|\hat{v}^h\|_{H^1(\Omega)}\\ &= f(\hat{v}^h\det(\phi^h)')-\textstyle\int_\Omega-\Delta u\hat{v}^hdt-\int_{\partial\Omega}\frac{\partial u}{\partial\nu}\hat{v}^hds\\ &\quad+\mathcal{O}(h^{m-1})\|u\|_{H^m(\Omega)}\|\hat{v}^h\|_{H^1(\Omega)}\\ &= f(\hat{v}^h\det(\phi^h)')-f(\hat{v}^h)+\tilde{a}^h\mathcal{O}(h^{m-1})\|u\|_{H^m(\Omega)},\\ &\quad\|\hat{v}^h\|_{H^1(\Omega)} \text{ by (6.69) and } \hat{v}^h|_{\partial\Omega}=0,\\ &= \mathcal{O}(h^{m-1})\|u\|_{H^m(\Omega)}\|v^h\|^h_{H^1(\Omega)}, \text{ since } \|u\|_{H^1(\Omega)}\\ &\le C\|f\|_{H^{-1}(\Omega)}.\end{aligned}$$

This directly yields

$$|a^h(u^h-I^hu,v^h)| \le Ch^{m-1}\|u\|_{H^m(\Omega)}\|\hat{v}^h\|^h_{H^1(\Omega)} \ \forall u \in H^m(\Omega)\cap\mathcal{U} \qquad (6.70)$$

If we consider instead of $A_su=-\Delta u=f$ the general case (6.2) - (6.3), we obtain the same result with $\partial\hat{z}^h/\partial\nu$ and $\partial u/\partial\nu$ replaced by $B_a\hat{z}^h$ and B_au, resp., again annihilated by $\hat{v}^h|_{\partial\Omega}=0$. To obtain (6.65) let again $f=Au$ and $A^hu^h=Q'^hAu$. Then

$$\begin{aligned}a^h(u^h,v^h)-a(u,\hat{v}^h) &= a^h(u^h,v^h)-\textstyle\int_\Omega\nabla u\cdot\nabla\hat{v}^h(x)dx\\ &= \textstyle\int_{F^h(\Omega^h)}fv^hdx-\int_\Omega\nabla u(t)\cdot\nabla\hat{v}^h(t)dt \quad \text{by (6.69)}\\ &= f(\hat{v}^h)-\textstyle\int_\Omega-\Delta u\hat{v}^hdt-\int_{\partial\Omega}\frac{\partial u}{\partial\nu}\hat{v}^hds\\ &\quad+\mathcal{O}(h^{m-1})\|u\|_{H^1(\Omega)}\|\hat{v}^h\|^h_{H^1(\Omega)}\\ &= f(\hat{v}^h)-f(\hat{v}^h)+\mathcal{O}(h^{m-1})\|u\|_{H^1(\Omega)}\|\hat{v}^h\|^h_{H^1(\Omega)}\\ &\quad\text{by } \hat{v}^h|_{\partial\Omega}=0\\ &= \mathcal{O}(h^{m-1})\|u\|_{H^1(\Omega)}\|v^h\|^h_{\mathcal{V}}.\end{aligned}$$

To relate these estimates to the general definition of classical consistency in [57], we combine the last equality with (6.70) and the continuity of $a(\cdot,\cdot)$ to see

$$|a^h(I^h u, v^h) - a(u, \hat{v}^h)| \leq C h^{m-1} \|u\|_{H^m(\Omega)} \|\hat{v}^h\|_{H^1(\Omega)}.$$

Now, with $\hat{v}^h \in \hat{\mathcal{V}}_b^h \subset \mathcal{V}_b$

$$a(u, \hat{v}^h) = \langle Au, \hat{v}^h \rangle_{H^{-1}(\Omega) \times \hat{\mathcal{V}}_b^h} = \langle Q'^h Au, v^h \rangle_{\mathcal{V}_b^{h'} \times \mathcal{V}_b^h} (1 + \mathcal{O}(h^{m-1})\|u\|_{H^m(\Omega)}) \cdot \|v^h\|_{\mathcal{V}^h} \text{ and } \langle A^h u^h, v^h \rangle_{\mathcal{V}_b^{h'} \times \mathcal{V}_b^h} := a^h(u^h, v^h) \qquad (6.71)$$

the combination with (6.57) shows (6.66). The last estimate for $\|u_0 - \hat{u}_0^h\|_{H^1(\Omega)}$ is an immediate consequence of the stability combined with the above consistency errors. □

Remark 6.4.3. 1) For isoparametric FEs the above nonlinear ϕ^h, see (6.54) determines the discretization. However, since $(\phi^h)' - Id = \mathcal{O}(h^m)$ the discrete A^h and G^h for the isoparametric case is still $k-$ times consistently differentiable with A and G in u_0, resp.
2) For a $\mathcal{U}_b$ coercive $a(\cdot,\cdot)$ we obtain, for $\mathcal{U}_b^h = \mathcal{U}_b^h$ the $\mathcal{U}_b^h$ coercivity of $a^h(\cdot,\cdot)$. Otherwise we get the discrete inf–sup– condition for $a^h(\cdot,\cdot)$, $a_s^h(\cdot,\cdot)$. This implies for all cases stability and convergence as above. We include this result in this Remark, since we want to avoid to go through all technical details of the proofs.

6.5 Approximate Operators, Bilinear and Linear Forms

In this and the next subsection we consider mainly quadrature approximations. They are defined with function values. Hence, the functions $u, u_0, f, a_{ij}, \ldots,$ have to be chosen smooth enough to allow function evaluations. So, we impose the following condition:

$$\text{Let } \Omega \text{ be bounded and } u, u_0, f, a_{ij}, \ldots, \in W_p^m(\Omega) \subset C(\Omega), \quad \text{e.g., for } m \geq 1,\ mp > n. \qquad (6.72)$$

In the preceeding Sections we considered the more complicated variational crimes for FEs. They are caused by the violated boundary conditions and continuity. Now, we return to general n and to the more straight forward variational crimes for FE- and spectral methods. They are obtained by replacing the bilinear form $a(\cdot,\cdot)$ and $a^h(\cdot,\cdot)$ and the corresponding operators A and A^h by quadrature approximations. To avoid a duplication of formulas and according to the Notation 4.1 in Section 4.2, we denote the original bilinear forms and operators in this Section as $a^h(\cdot,\cdot)$ and A^h, resp. We will show that

$$\tilde{a}^h(\cdot,\cdot) : \mathcal{U}_b^h \times \mathcal{V}_b^h \to \mathbb{R} \text{ and}$$
$$\sup_{0 \neq v^h \in \mathcal{V}_b^h} \frac{|a^h(u^h, v^h) - \tilde{a}^h(u^h, v^h)|}{\|v^h\|_{\mathcal{V}}^h} \to 0 \text{ for } h \to 0 \ \forall u^h \in \mathcal{U}_b^h. \qquad (6.73)$$

Similarly, we replace f or f^h, both denoted as f^h, by $\tilde{f}^h$ with

$$\tilde{f} : \mathcal{V}_b^h \to \mathbb{R} \text{ and } \sup_{0 \neq v^h \in \mathcal{V}_b^h} \frac{|\tilde{f}^h(v^h) - f^h(v^h)|}{\|v^h\|_{\mathcal{V}}^h} \to 0 \text{ for } h \to 0 \,. \quad (6.74)$$

These $\tilde{a}^h(u,v), \tilde{f}^h(v)$ are usually not defined for the original $u \in \mathcal{U}, v \in \mathcal{V}$, since quadrature formulas require $u(P_j)$, see (6.26), to be defined. However, convergence (and consistency) is only to expected for smooth enough u_0, u. For these u_0, u the $u_0(P_j), u(P_j)$ have to be defined.

We proceed as in Sections 6.2 - 6.4: In a first sequence of results we prove the $\mathcal{U}_b^h$ coercivity and inf–sup–condition of $\tilde{a}^h(\cdot,\cdot)$ for $\mathcal{U}_b$ coercive $a(\cdot,\cdot)$. Then we estimate the variational consistency errors as indicated in (6.73) and (6.74).

More generally than in (6.26) we assume to have a quadrature formula $q^h(w)$ satisfying, for [7] some $\ell,\ k$,

$$|\int_\Omega w(x)dx - q^h(w)| \leq Ch^\ell \sum_{T \in \mathcal{T}^h} \|w\|_{W_\infty^k(T)} \quad \forall w \in C(\Omega). \quad (6.75)$$

We may choose $w \in \mathcal{U}, \mathcal{U}^h, \mathcal{V}, \mathcal{V}^h$, or products of those functions. Usually, we apply (6.75) to the sub triangles, T. Then we find (, often even for $C = 1$, if all weights are positive)

$$q^h(w) = \sum_{T \in \mathcal{T}^h} q_T^h(w) \text{ with } q_T^h(w) := \sum_{P_{j,T} \in \overline{T}} w_{j,T} w(P_{j,T}) \text{ and}$$
$$|q_T^h(w)| \leq C \text{ meas } (T) \|w\|_{L^\infty(T)} \quad \forall T \in \mathcal{T}^h. \quad (6.76)$$

We want to recall the Remark from Section 4.1

Remark 6.5.1. In these quadrature approximations (6.76), (6.77) the quadrature points $P_{j,T}$ may be chosen totally independent of the interpolation points for the FEs. This contrasts to the situation for collocation methods. Here quadrature and interpolation= collocation points coincide, see (6.98), below.

The quadrature error in (6.75) has the structure

$$|\sum_{T \in \mathcal{T}^h} (\int_T w(x)dx - q_T^h(w))| \leq C\ h^\ell \sum_{T \in \mathcal{T}^h} \|w\|_{W_\infty^k(T)} = C\ h^\ell \|w\|_{W_\infty^k(\Omega)}^h \quad (6.77)$$

We apply this $q^h(w)$ to $w = fv$ and to the terms in $a^h(u^h, v^h)$ and $a_s^h(u^h, v^h)$ to define

[7] since errror estimates for quadrature formulas nearly always are formulated w.r.t. the sup norm $\|\cdot\|_{W_\infty^k(T)}$ we restrict the discussion to this case

$$\tilde{f}^h(v^h) := \sum_{T\in\mathcal{T}^h} q_T^h(f\ v^h) = \sum_{T\in\mathcal{T}^h}\sum_{P_{j,T}\in\overline{T}} w_{j,T}(fv^h)(P_{j,T}) \text{ and} \tag{6.78}$$

$$\begin{aligned}\tilde{a}^h(u^h,v^h) &:= <\tilde{A}^h u^h,\ v^h\tilde{>}^h := \sum_{T\in\mathcal{T}^h} q_T^h\Big(\sum_{i,j=1}^n a_{ij}\partial_i\ u^h\partial_j\ v^h \\ &\quad + \sum_{i=1}^n a_{i0}\ (\partial_i\ u^h)v^h + \sum_{j=1}^n a_{0j}\ u^h\ (\partial_j\ v^h) + a_{00}u^h v^h\Big) \\ &= \sum_{T\in\mathcal{T}^h}\sum_{P_{k,T}\in\overline{T}} w_{k,T}\Big(\sum_{i,j=1}^n a_{ij}\partial_i\ u^h\partial_j\ v^h \\ &\quad + \sum_{i=1}^n a_{i0}\ (\partial_i\ u^h)v^h + \sum_{j=1}^n a_{0j}\ u^h\ (\partial_j\ v^h) + a_{00}u^h v^h\Big)(P_{j,T}),\ \text{and}\end{aligned} \tag{6.79}$$

$$\begin{aligned}\tilde{a}_s^h(u^h,v^h) &:= (\tilde{A}_s^h u^h,\ v^h\tilde{)}^h := \sum_{T\in\mathcal{T}^h} q_T^h\Big(\Big(\sum_{i,j=1}^n -a_{ij}\partial_i\ \partial_j\ u^h \\ &\quad + \sum_{i=1}^n a_{i0}\ (\partial_i\ u^h) + \sum_{j=1}^n -\partial_j\ (a_{0j}\ u^h) + a_{00}u^h\Big)v^h\Big) \\ &= \sum_{T\in\mathcal{T}^h}\sum_{P_{k,T}\in\overline{T}} w_{k,T}\Big(\Big(\sum_{i,j=1}^n -a_{ij}\partial_i\ \partial_j\ u^h \\ &\quad + \sum_{i=1}^n a_{i0}\ (\partial_i\ u^h) + \sum_{j=1}^n -\partial_j\ (a_{0j}\ u^h) + a_{00}u^h\Big)v^h\Big)\Big)(P_{j,T}).\end{aligned} \tag{6.80}$$

Now we are able to define the weak and strong approximate solutions. Determine $\tilde{u}_0^h$ s.t.

$$\tilde{a}^h(u_0^h,v^h) = \tilde{f}^h(v^h) \text{ and } \tilde{a}^h(u_0^h,v^h) = \tilde{f}^h(v^h)\ \forall v^h\in\mathcal{V}_b^h. \tag{6.81}$$

The inverse estimate (2.44) for the piecewise polynomial v^h of degree $\le m+\tau$ yields with $n\ge 2$

$$\|v^h\|_{W_\infty^j(\Omega)}^h \le Ch^{-j+1-n/2}\|v^h\|_{H^1(\Omega)}^h \text{ for } 1\le j\le m+\tau. \tag{6.82}$$

We use it to estimate, e.g.,

$$
\begin{aligned}
|f(v^h) - \tilde{f}^h(v^h)| &= |\int_\Omega f v^h dx - q^h(f v^h)| \\
&\quad \text{by } \|fv\|_{W^k_\infty(T)} \le \|f\|_{W^k_\infty(T)} \cdot \|v^h\|_{W^{\min\{m+\tau,k\}}_\infty(T)} \\
&\le Ch^\ell \sum_{T\in\mathcal{T}^h} \|f\|_{W^k_\infty(T)} \|v^h\|_{W^{\min\{m+\tau,k\}}_\infty(T)} \text{ by (6.75)} \\
&\le Ch^\ell \|f\|^h_{W^k_\infty(\Omega)} \cdot \|v^h\|^h_{W^{\min\{m+\tau,k\}}_\infty(\Omega)} \qquad (6.83) \\
&\quad \text{by } \sum_i a_i b_i \le \sum_i a_i \sum_i b_i \text{ for } a_i, b_i \ge 0 \ \forall\ i \\
&\le Ch^\ell \|f\|^h_{W^k_\infty(\Omega)} \cdot h^{-\min\{m+\tau,k\}+1-n/2} \|v^h\|^h_{H^1(\Omega)} \text{ by (6.82)} \\
&\le Ch^{\ell-\min\{m+\tau,k\}+1-n/2} \|f\|^h_{W^k_\infty(\Omega)} \cdot \|v^h\|^h_{H^1(\Omega)}.
\end{aligned}
$$

$\ell+1-n/2-\min\{m+\tau,k\} > 0$ diskutieren, insbesondere so, dass quadraturformel = integral fuer FEe

Condition 6.4 *Let Ω and the FEs satisfy Condition 6.1. We choose A, A_s, $a(\cdot,\cdot)$, $a_s(\cdot,\cdot)$, $f(\cdot)$, B_a and the quadrature approximations $\tilde{A}^h$, $\tilde{A}^h_s$, $\tilde{a}^h(\cdot,\cdot)$, $\tilde{a}^h_s(\cdot,\cdot)$, $\tilde{f}^h(\cdot)$ as in (6.2), (6.3), and (6.5), (6.78), (6.79), (6.80).Let the estimates (6.75), (6.76), (6.77) for quadrature errors be satisfied, with ℓ, k defined there. The functions $u, u_0, f, a_{ij}, \ldots$, have to be chosen smooth enough to allow function evaluations. Hence, see (6.72) above, we always require $W^m_p(\Omega) \subset C(\Omega)$, e.g. for $m \ge 1$, $mp > n$.*

Theorem 6.5.2. *Under the Condition 6.4 let, for* Dirichlet boundary conditions *or* natural boundary conditions *as in (6.2), the $\ell - 2(\min\{m+\tau, k+1\} - 1 + n/2) > 0$, $\tau \ge -1$, see (2.34). Then a $\mathcal{U}_b$-coercive $a(\cdot,\cdot)$ implies, for $\mathcal{U}^h_b = \mathcal{V}^h_b$, again a $\mathcal{U}^h_b$-coercive $\tilde{a}^h(\cdot,\cdot)$. For $\mathcal{U}^h_b \ne \mathcal{V}^h_b$, the uniform discrete inf–sup– condition is valid for $\tilde{a}^h(\cdot,\cdot)$. This implies the stability of the corresponding $\tilde{A}^h$.*

If $\mathcal{V}^h_b \not\subset \mathcal{V}_b$ and $\mathcal{V}^h_b \not\subset \mathcal{V}$ the violated boundary and interior boundary error terms still guarantee the inf–sup– condition or stability for the corresponding $a(\cdot,\cdot)$, $a^h(\cdot,\cdot)$, $\tilde{a}^h(\cdot,\cdot)$, or the stability of the A^h, $\tilde{A}^h$.

These results, see Theorem 6.1.2, remain correct for the approximations $\tilde{a}^h_s(\cdot,\cdot)$, of the strong bilinear forms, $a_s(\cdot,\cdot)$, $a^h_s(\cdot,\cdot)$, introduced in Chapter 4. This statement requires the condition that the quadrature approximations and enough points on every edge, e, have been chosen to guarantee quadrature errors and differences in (6.3) vanishing with h.

Proof (6.77) yields, with $\sigma := \min\{m+\tau, k+1\}$,

$$
\begin{aligned}
&| a^h(u^h,v^h) - \tilde{a}^h(u^h,v^h)| \\
&\quad = \Big| \sum_{T\in\mathcal{T}^h} \Big(\int_T \Big(\sum_{i,j=1}^n a_{ij}\partial_i\, u^h \partial_j\, v^h + \sum_{i=1}^n a_{i0}\, (\partial_i\, u^h) v^h \\
&\qquad + \sum_{j=1}^n a_{0j}\, u^h\, (\partial_j\, v^h) + a_{00} u^h v^h dx\Big) \\
&\qquad - q_T^h \Big(\sum_{i,j=1}^n a_{ij}\partial_i\, u^h \partial_j\, v^h + \sum_{i=1}^n a_{i0}\, (\partial_i\, u^h) v^h \\
&\qquad + \sum_{j=1}^n a_{0j}\, u^h\, (\partial_j\, v^h) + a_{00} u^h v^h dx\Big)\Big)\Big| \qquad (6.84)\\
&\quad \le Ch^\ell \max_{i,j=0}^n \|a_{ij}\|_{W_\infty^k(\Omega)} \sum_{T\in\mathcal{T}^h} \|u^h\|_{W_\infty^\sigma(T)} \|v^h\|_{W_\infty^\sigma(T)} \\
&\quad \le Ch^\ell \max_{i,j=0}^n \|a_{ij}\|_{W_\infty^k(\Omega)} \cdot \|u^h\|^h_{W_\infty^\sigma(\Omega)} \|v^h\|^h_{W_\infty^\sigma(\Omega)}.
\end{aligned}
$$

For a corresponding estimate for the strong bilinear forms, we would find

$$
\begin{aligned}
&| a_s^h(u^h,v^h) - \tilde{a}_s^h(u^h,v^h)| \\
&\qquad \le Ch^\ell \max_{i,j=0}^n \|a_{ij}\|_{W_\infty^k(\Omega)} \cdot \|u^h\|^h_{W_\infty^{\sigma+1}(\Omega)} \|v^h\|^h_{W_\infty^{\sigma-1}(\Omega)}. \qquad (6.85)
\end{aligned}
$$

We employ (6.82) for u^h, v^h and find, for $j = \sigma, \sigma+1, \sigma-1$,

$$
\begin{aligned}
&|a^h(u^h,v^h) - \tilde{a}^h(u^h,v^h)| \le Ch^{\ell-2(\sigma-1+n/2)} \\
&\quad \cdot \max_{i,j=0}^n \|a_{ij}\|_{W_\infty^k(\Omega)} \cdot \|u^h\|^h_{H^1(\Omega)} \|v^h\|^h_{H^1(\Omega)}. \qquad (6.86)
\end{aligned}
$$

and

$$
\begin{aligned}
&|a_s^h(u^h,v^h) - \tilde{a}_s^h(u^h,v^h)| \le Ch^{\ell-2(\sigma-1+n/2)} \\
&\quad \cdot \max_{i,j=0}^n \|a_{ij}\|_{W_\infty^k(\Omega)} \cdot \|u^h\|^h_{H^2(\Omega)} \|v^h\|^h_{L^2(\Omega)}. \qquad (6.87)
\end{aligned}
$$

As in Theorem 6.1.1 and with the estimates for $|a_s^h(u^h,v^h) - a^h(u^h,v^h)|$ this shows the claim.

If $\mathcal{V}_b^h \subset \mathcal{V}_b$ and $\mathcal{V}_b^h \subset \mathcal{V}$ should be violated, the vanishing boundary and interior boundary error terms again yield convergence. □

We find, see Lemma 5.1.4,

Theorem 6.5.3. *Under the Condition 6.4 assume conforming FE-spaces* $\mathcal{U}_b^h$, $\mathcal{V}_b^h$. *Both, the* classical *and the* variational discretization errors *vanish for* $a(\cdot,\cdot)$ *with* $u \in H^1(\Omega)$ *and appropriate* ℓ, k.

With h *-independent* $C = C_{(m,n,\rho,\chi)}$, *the* classical *and* variational discretization error *can be estimated by*

$$
\|\tilde{A}^h I^h u - \tilde{Q}'^h A u\|_{H^{-1}(\Omega)} \le Ch^{m-1-n/2} \max_{i,j=0}^n \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|u\|_{W_\infty^m(\Omega)} \qquad (6.88)
$$

and

$$\sup_{0\neq v^h\in\mathcal{V}_b^h}\frac{|\tilde{a}^h(u_0^h-u_0,v^h)|}{\|v^h\|^h_{H^1(\Omega)}}\ \textit{and} \tag{6.89}$$

$$\sup_{0\neq v^h\in\mathcal{V}_b^h}\frac{|a^h(u_0,v^h)-\tilde{a}^h(u_0,v^h)|}{\|v^h\|^h_{H^1(\Omega)}}+\sup_{0\neq v^h\in\mathcal{V}_b^h}\frac{|f^h(v^h)-\tilde{f}^h(v^h)|}{\|v^h\|^h_{H^1(\Omega)}}$$
$$\leq C\big(h^{\ell+1-n/2-\min\{m+\tau,k+1\}}\max_{i,j=0^n}\|a_{ij}\|_{W^k_\infty(\Omega)}\cdot\|u_0\|_{W^{k+1}_\infty(\Omega)}$$
$$+h^{\ell+1-n/2-\min\{m+\tau,k\}}\|f\|_{W^k_\infty(\Omega)}\big),$$

based upon (6.76) and upon (6.77), with consistency of order $m-1-n/2\geq 0$ *and* $\ell+1-n/2-\min\{m+\tau,k+1\}\geq 0$*, resp. The regularity conditions for* a_{ij},u,f *are indicated by the norms, e.g.,* $\|u\|_{W^m_\infty(\Omega)}$.

If $\mathcal{V}_b^h\subset\mathcal{V}_b$ *and* $\mathcal{V}_b^h\subset\mathcal{V}$ *should be violated, we have to add the boundary and interior boundary error terms as discussed in the last Sections.*

Remark 6.5.4. 1)The first and second line in (6.89) emphazise the consistency and quadrature error, resp.

2)The estimate in (6.89) remains correct if $u_0,\|u_0\|_{W^{k+1}_\infty(\Omega)}$ are replaced by u_0^h, $\|u_0^h\|_{W^{\min\{m+\tau,k+1\}}_\infty(\Omega)}$ or if u_0 and u_0^h are replaced by any $u\in\mathcal{U}_b$ and the discrete solution u^h of $\tilde{A}^h u_0^h=\tilde{Q}'^h Au$, resp.

3) The estimates in (6.88) have to be compared with the estimates in (6.89) and with $\|I^hu-u\|^h_{H^1(\Omega)}=\mathcal{O}(h^{m-1})\|u\|^h_{W^m_\infty(\Omega)}<\mathcal{O}(h^{m-1-n/2})$. For small enough h we have lost $h^{-n/2}$ compared to the optimal possible result. Obviously the loss of $h^{-n/2}$ is due to the fact, that the estimate (6.76) requires $\|w\|_{L^\infty(T)}$ instead of $\|v^h\|_{H^1(T)}$ required in (6.88). Optimal results could be obtained, by combining $\|I^hu-u\|^h_{H^1(\Omega)}$ with the estimates in (6.88), (6.89) under the following very strong condition: Choose quadrature formulas (6.75) with high accuracy ℓ s.t. $l-(\min\{m+\tau,k+1\}-1+n/2)\geq m-1$.

Proof To determine the classical consistency error we use the above (6.84) for FEs of local degree, $m-1\leq$ degree $\leq m+\tau$. It implies with (2.35), (2.44), and by Notation 4.1, (4.109), (6.76) as in (6.83) for $\mathcal{V}_b^h\subset\mathcal{V}_b$ and (6.82) for $j=1$

$$
\begin{aligned}
&\left| < \tilde{A}^h\ I^h u - \tilde{Q}^{'h} A u,\ v^h \tilde{>}^h \right| = \\
&= \left| \sum_{T\in\mathcal{T}^h} q_T^h \Big(\sum_{i,j=1}^n a_{ij}\partial_i\ (I^h u - u)\partial_j\ v^h + \sum_{i=1}^n a_{i0}\ (\partial_i\ (I^h u - u))v^h \right. \\
&\quad \left. + \sum_{j=1}^n a_{0j}\ (I^h u - u)\ (\partial_j\ v^h) + a_{00}(I^h u - u)v^h dx \Big) \right| \qquad (6.90) \\
&\leq C\ \text{meas}\ (\Omega) \max_{i,j=0}^n \|a_{ij}\|_{L_\infty(\Omega)} \sum_{T\in\mathcal{T}^h} \|I^h u - u\|_{W^1_\infty(T)} \|v^h\|_{W^1_\infty(T)} \\
&\leq C \max_{i,j=0}^n \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|I^h u - u\|^h_{W^1_\infty(\Omega)} \|v^h\|^h_{W^1_\infty(\Omega)} \\
&\leq C h^{m-1-n/2} \max_{i,j=0}^n \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|u\|^h_{W^m_\infty(\Omega)} \|v^h\|^h_{H^1(\Omega)},
\end{aligned}
$$

hence (6.88).

To obtain (6.89) we use (6.84) and replace in (6.84) the u^h and $\|u^h\|_{W_\infty^{\min\{m+\tau,k+1\}}(T)}$ by u_0 and $\|u_0\|_{W_\infty^{k+1}(T)}$, resp. This yields, with the inverse estimate (6.82) for $j = \min\{m+\tau, k+1\}$, applied only to $\|v^h\|^h_{W_\infty^{\min\{m+\tau,k+1\}}(\Omega)}$,

$$
\begin{aligned}
|a^h(u_0, v^h)\ &- \tilde{a}^h(u_0, v^h)| \qquad (6.91) \\
&\leq C h^\ell \max_{i,j=0}^n \|a_{ij}\|_{W^k_\infty(\Omega)} \cdot \|u_0\|^h_{W^{k+1}_\infty(\Omega)} \|v^h\|^h_{W_\infty^{\min\{m+\tau,k+1\}}(\Omega)} \\
&\leq C h^{\ell+1-n/2-\min\{m+\tau,k+1\}} \max_{i,j=0}^n \|a_{ij}\|_{W^k_\infty(\Omega)} \cdot \|u_0\|^h_{W^{k+1}_\infty(\Omega)} \|v^h\|^h_{H^1(\Omega)}.
\end{aligned}
$$

The f^h contribution is estimated in (6.83)

$$
\begin{aligned}
|f(v^h) - \tilde{f}^h(v^h)| &\leq C h^\ell \|f\|^h_{W^k_\infty(\Omega)} \cdot \|v^h\|^h_{W_\infty^{\min\{m+\tau,k\}}(\Omega)} \\
&\leq h^{\ell+1-n/2-\min\{m+\tau,k\}} \|f\|^h_{W^k_\infty(\Omega)} \cdot \|v^h\|^h_{H^1(\Omega)} \qquad (6.92)
\end{aligned}
$$

Combining both estimates yields the second line in (6.89). To prove the first line in (6.89), we combine (6.84), the second line in (6.89) and (6.92):

$$
\begin{aligned}
|\tilde{a}^h(u_0^h - u_0, v^h)| &= |\tilde{f}^h(v^h) - \tilde{a}^h(u_0, v^h)| \\
&= |\tilde{f}^h(v^h) - a^h(u_0, v^h) + (a^h(u_0, v^h) - \tilde{a}^h(u_0, v^h))| \\
&\leq |\tilde{f}^h(v^h) - f(v^h)| + |(a^h(u_0, v^h) - \tilde{a}^h(u_0, v^h))|
\end{aligned}
$$

hence again (6.89). □

Theorem 6.5.5. *Under the Condition 6.4 assume conforming FE-spaces* $\mathcal{U}_b^h$, $\mathcal{V}_b^h$. *Both, the* classical *and the* variational discretization errors *vanish for* $u \in H^2(\Omega)$ *and appropriate* ℓ, k.

Again, the classical discretization error *can be estimated by*

$$\|\tilde{A}_s^h I^h u - \tilde{Q}'^h A_s u\|_{H_h^{-1}(\Omega)} \le C h^{m-2-n/2} \max_{i,j=0}^{n} \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|u\|^h_{W^m_\infty(\Omega)} \tag{6.93}$$

and

$$\sup_{0\neq v^h\in\mathcal{V}_b^h} \frac{|\tilde{a}^h(u_0^h - u_0, v^h)|}{\|v^h\|^h_{H^1(\Omega)}} \text{ and} \tag{6.94}$$

$$\sup_{0\neq v^h\in\mathcal{V}_b^h} \frac{|a^h(u_0, v^h) - \tilde{a}^h(u_0, v^h)|}{\|v^h\|^h_{H^1(\Omega)}} + \sup_{0\neq v^h\in\mathcal{V}_b^h} \frac{|f^h(v^h) - \tilde{f}^h(v^h)|}{\|v^h\|^h_{H^1(\Omega)}}$$
$$\le C\big(h^{\ell-n/2-\min\{m+\tau,k\}} \max_{i,j=0}^{n} \|a_{ij}\|_{W^k_\infty(\Omega)} \cdot \|u_0\|^h_{W^{k+2}_\infty(\Omega)}$$
$$+h^{\ell-n/2-\min\{m+\tau,k\}} \|f\|_{W^k_\infty(\Omega)}\big),$$

based upon (6.76) and (6.77) with consistency of order $m-2-n/2 \ge 0$ *and* $\ell - n/2 - \min\{m+\tau, k\} \ge 0$, *resp.*

If $\mathcal{V}_b^h \subset \mathcal{V}_b$ *and* $\mathcal{V}_b^h \subset \mathcal{V}$ *should be violated, we have to add the boundary and interior boundary error terms as discussed in the last Sections.*

Proof : To discuss the strong terms we modify (6.90) and find

$$\big|(\tilde{A}_s^h I^h u - \tilde{Q}'^h A_s u,\ \tilde{v^h})^h\big| \tag{6.95}$$
$$\le C \max_{i,j=0}^{n} \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|I^h u - u\|^h_{W^2_\infty(\Omega)} \|v^h\|^h_{L^\infty(\Omega)}$$
$$\le C h^{m-2-n/2} \max_{i,j=0}^{n} \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|u\|^h_{W^m_\infty(\Omega)} \|v^h\|^h_{L^2(\Omega)},$$

hence (6.93). Similarly, we find for the strong bilinear forms, see (6.91),

$$|a_s^h(u_0, v^h) - \tilde{a}_s^h(u_0, v^h)| \tag{6.96}$$
$$\le C h^\ell \max_{i,j=0}^{n} \|a_{ij}\|_{W^k_\infty(\Omega)} \cdot \|u_0\|^h_{W^{k+2}_\infty(\Omega)} \|v^h\|^h_{W^{\min\{m+\tau,k\}}_\infty(\Omega)}$$
$$\le C h^{\ell-n/2-\min\{m+\tau,k\}} \max_{i,j=0}^{n} \|a_{ij}\|_{W^k_\infty(\Omega)} \cdot \|u_0\|^h_{W^{k+2}_\infty(\Omega)} \|v^h\|^h_{L^2(\Omega)}.$$

To prove (6.94), we combine (6.84), (6.89) and (6.92):

$$|\tilde{a}_s^h(u_0^h - u_0, v^h)| \le |\tilde{f}^h(v^h) - f(v^h)| + |(a_s(u_0, v^h) - \tilde{a}_s^h(u_0, v^h))|$$

hence again (6.94) with the obvious modification of (6.92). □

Theorem 6.5.6. *Let* $a(\cdot,\cdot)$ *be coercive and* $\mathcal{U}_b^h$, $\mathcal{V}_b^h$ *be conform and Condition 6.4 be satisfied. Then the approximate weak and strong solutions* u_0^h *of the discrete equations (6.81), see Theorems 6.5.2, 6.5.3 6.5.5 converge to the weak and strong solutions according to*

$$
\begin{aligned}
&\|u_0 - u_0^h\|_{H^1(\Omega)} \qquad (6.97)\\
&\quad \le C h^{m-1-n/2} \max_{i,j=0}^{n} \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|u_0\|_{W^m_\infty(\Omega)}, \ \textit{or}\\
&\quad \le C\big(h^{\ell+1-n/2-\min\{m+\tau,k+1\}} \max_{i,j=0^n} \|a_{ij}\|_{W^k_\infty(\Omega)} \cdot \|u_0\|_{W^{k+1}_\infty(\Omega)}\\
&\qquad + h^{\ell+1-n/2-\min\{m+\tau,k\}} \|f\|_{W^k_\infty(\Omega)}\big), \ \textit{for the weak and}\\
&\quad \le C h^{m-2-n/2} \max_{i,j=0}^{n} \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|u_0\|^h_{W^m_\infty(\Omega)}, \ \textit{or}\\
&\quad \le C\big(h^{\ell-n/2-\min\{m+\tau,k\}} \max_{i,j=0}^{n} \|a_{ij}\|_{W^k_\infty(\Omega)} \cdot \|u_0\|^h_{W^{k+2}_\infty(\Omega)}\\
&\qquad + h^{\ell-n/2-\min\{m+\tau,k\}} \|f\|_{W^k_\infty(\Omega)}\big) \ \textit{for the strong forms.}
\end{aligned}
$$

The two estimates for the weak and the strong errors originate by comparing with the quadrature errors in (6.76) or (6.77). Again conditions for the functions and exponents of h to guarantee convergence are shown in the formulas. If $\mathcal{V}_b^h \subset \mathcal{V}_b$ and $\mathcal{V}_b^h \subset \mathcal{V}$ should be violated, we have to add the boundary and interior boundary error terms as discussed in the last Sections.

These results, see Theorem 6.1.2, remain correct for the approximate strong bilinear forms, $\tilde{a}_s^h(\cdot,\cdot)$. This statement requires the condition that, as above, the quadrature formulas are good enough and that enough points on every edge, e, have been chosen to guarantee the differences in (6.3) to vanish with h. In the convergence result for the strong solutions (for $a_s^h(\cdot,\cdot)$) the (6.40) has to be modified as $\|u_0^h - u_0\|_{H^2(\Omega)} \le C h^{m-3-n/2} \max_{i,j=0}^{n} \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|u_0\|^h_{W^m_\infty(\Omega)}$.

6.6 Collocation Methods for FE and Spectral Methods

We have discussed in Section 4.1 the transition from an approximate variational method via quadrature formulas to collocation. This collocation is always based on the strong bilinear form and its approximations. This requires two steps: Testing the weak bilinear form $\tilde{a}^h(\cdot,\cdot)$ w.r.t. the approximate pairing $<\cdot,\cdot\tilde{>}^h$ has to be transformed into testing the strong bilinear form $\tilde{a}_s^h(\cdot,\cdot)$ w.r.t. the approximate $L^2(\Omega)$ inner product $(\cdot,\cdot)^{\tilde{}h}_{L^2(\Omega)}$. According to Remark 6.5.1 the collocation and interpolation points have to be identical. We want to generalize this approach to FEMs and spectral methods We start with

6.6.1 Collocation methods for FEs

Instead of the weak form, see (6.2), (6.2), (6.3), (6.4), we study the strong bilinear form. In contrast to Section 6.5 we extend our studies to non conforming FEs as well.

In Section 4.1 we had already indicated the equivalence of the strong quadrature and the collocation formulation (4.52) and (4.53). We refer to

Theorem 6.5.2: It shows the simultaneous stability of A^h and $\tilde{A}^h, \tilde{A}^h_s$ for $\mathcal{U}^h_b$ -coercive $a(\cdot,\cdot)$ under the Condition 6.4 and for appropriate k, ℓ, m, τ. Under the condition (6.98), the combination with the following interpolation basis $v^h_{i,T}$ immediately yields the equivalent collocation equations as in Section 4.1 and the stability.
We have to choose specific FE spaces $\mathcal{V}^h_b$: We assume the $\mathcal{N}$ in Definition 2.2.2 consists of exactly those, say d, points used in the collocation equations (6.103), below. Hence, any $v^h \in \mathcal{V}^h_b$ is uniquely determined by the $v^h(P_j),\ \forall\ P_j \in \overline{T},\ \forall\ T \in \mathcal{T}^h$. Thus, with the *Dirac delta functions* $\delta(P_i)$, the $\mathcal{N}_\mathcal{T} = \{\delta(P_i) :\ \forall\ P_i \in \overline{T}\},\ \forall\ T \in \mathcal{T}^h$. So, the following condition is appropriate for conforming and nonconforming cases, see Definition 2.2.2 and (2.31).

$$\begin{aligned} &\forall T \in \mathcal{T}^h, T = F_T(K) \text{ is affine equivalent to } K, \text{ and} \\ &(K,\ \mathcal{P},\ \mathcal{N}) \text{ is a FE with } \mathcal{N} \text{ defined by} \\ &\mathcal{N}_T = \{\delta(P_i) : \forall P_j \in \overline{K} \text{ in (6.103)}\} \text{ is a unisolvent basis for } \mathcal{P}'. \end{aligned} \tag{6.98}$$

This property has to be checked for a given set of collocation points. We have to guarantee that in fact a $\mathcal{P}, \mathcal{P}_{m-1} \subseteq \mathcal{P} \subseteq P_{m+\tau}$, see (2.34), exists, s.t. $(K, \mathcal{P}, \mathcal{N})$ unisolvently defines a FE. Then, due to the required unisolvense of $(K,\ \mathcal{P},\ \mathcal{N})$, see (2.31), and by (6.98), the interpolation basis

$$\begin{aligned} v^h_{j,T} \in \mathcal{V}^h_b \text{ with } v^h_{j,T} \in \mathcal{V}^h_b \text{ and } v^h_{j,T}(P_i) = \delta_{i,j}\delta_{T,T'} \\ \forall\ i,j = 1,\dots,d,\ \forall T, T' \in \mathcal{T}^h \end{aligned} \tag{6.99}$$

is uniquely determined.

Our next job is the definition of a quadrature formula and the corresponding error estimate of the form (6.77). We find

Proposition 6.6.1. *Under the conditions of Theorem 2.5.1, specifically (2.34) and (6.98), let I^h be the FE-interpolation operator defined for $(K,\ \mathcal{P},\ \mathcal{N}\ )$. Then*

$$q^h(w) := \sum_{T \in \mathcal{T}^h} q^h_T(w) := \int_\Omega I^h w(x)dx \tag{6.100}$$

is well defined for $w \in C(\Omega)$. The quadrature error for (6.100) is estimated as

$$\begin{aligned} |\int_\Omega w(x)dx - q^h(w)| &\le C\ (\ meas\ \Omega)^{1/q}\|w - I^h w\|^h_{L^p(\Omega)} \\ &\quad for\ 1 \le p \le \infty,\ 1/p + 1/q = 1 \\ &\le C\ (\ meas\ \Omega)^{1/q}\ h^m \|w\|_{W^m_p(\Omega)},\quad for\ w \in W^m_p(\Omega). \end{aligned} \tag{6.101}$$

A combination with Theorem 2.5.1 shows that (6.77) is satisfied as well.

In fact, the FE with all $N_{i,T} = 1$ has a bounded integral.

We recall the $a(\cdot,\cdot)$, $a_s(\cdot,\cdot)$, A, A_s in (6.2) with

$$\begin{aligned} &\textit{Dirichlet} \text{ and } \textit{natural boundary conditions}, \text{ realized as} \\ \mathcal{U}_b = \mathcal{V}_b = H_0^1(\Omega) &\text{ and } \mathcal{U}_b = \{u \in H^2(\Omega) : \ B_a u|_{\partial\Omega} = 0\}, \\ \mathcal{V}_b = \mathcal{V} = H^1(\Omega), &resp. \end{aligned} \tag{6.102}$$

The $\tilde{a}^h(\cdot,\cdot)$, $\tilde{a}_s^h(\cdot,\cdot)$, $\tilde{A}^h$, $\tilde{A}_s^h$ are explicitly formulated in (6.78) - (6.81). In short form and with the above quadrature approximation we obtain

$$\begin{aligned} a_s^h(u^h, v^h) &= \sum_{T\in\mathcal{T}^h} \int_T A_s u^h v^h dx = (f^h, v^h) \ \ \forall \ \ v^h \in \mathcal{V}_b^h \\ \tilde{a}_s^h(u^h, v^h) &= \sum_{T\in\mathcal{T}^h} \Big(\sum_{P_j = P_{j,T} \in T} w_j (A_s u^h)(P_j) v_i^h(P_j) \Big) \\ &= \sum_{T\in\mathcal{T}^h} (A_s u^h, v^h)_T^{\tilde{}h} = (f^h, v^h)^{\tilde{}h} \ \ \forall \ \ v^h \in \mathcal{V}_b^h \end{aligned} \tag{6.103}$$

Then we determine the (strong) discrete solution $u_0^h \in \mathcal{U}_b^h$ from

$$\tilde{a}_s^h(u_0^h, v^h) = (f^h, v^h)^{\tilde{}h} \text{ or } = \ (f, v^h)^{\tilde{}h} \ \ \forall \ \ v^h \in \mathcal{V}_b^h. \tag{6.104}$$

As in Section 4.1 we test the strong variational bilinear form with the above interpolation basis $v_{j,T}^h$. Under the condition (6.98), we still have the same equivalence as in (4.53), equivalent to the collocation method.

$$\begin{aligned} &\tilde{a}_s^h(u_0^h, v^h) = (f, v^h)^{\tilde{}h} = 0 \ \forall \ v^h \in \mathcal{V}_b^h \iff \\ &(A_s u_0^h - f)(P_j) = 0 \ \forall \ P_j \in T \ \forall \ T \in \mathcal{T}^h. \end{aligned} \tag{6.105}$$

We want to recall four results, essentially from Section 6.5, see Theorems 6.5.2 and 6.5.6. The coercivity of $a(\cdot,\cdot)$ implies the discrete inf–sup–condition for $a_s^h(\cdot,\cdot)$ and $\tilde{a}_s^h(\cdot,\cdot)$. This yields the stability for the $a_s^h(\cdot,\cdot)$ and $\tilde{a}_s^h(\cdot,\cdot)$ and for the equivalent collocation formulation in (6.105). The u_0^h computed from the weak and the strong discrete variational problems are identical. And the consistency and convergence of the collocation and the strong variational approach are essentially the same, except for variational crimes. So we have to discuss the implications of variational crimes, excluded in the last Section 6.5.

Now, we study the implications of variational crimes. Similarly to the transitions from (4.2) (4.3) and (4.11) to (4.13) and (4.16) to (4.29) we obtain first the analogous of (4.46) in the form

$$(f, v^h)\tilde{}^h = \tilde{a}^h(u_0^h, v^h) = \tilde{a}_s^h(u_0^h, v^h)$$
$$\forall\, v^h \in \mathcal{V}_b^h \quad \text{for conforming FEs and} \tag{6.106}$$
$$(f, v^h)\tilde{}^h = \tilde{a}^h(u_0^h, v^h) = \tilde{a}_s^h(u_0^h, v^h) + \int_{\partial\Omega} \frac{\partial u_0^h}{\partial \nu}\, v^h ds$$
$$\forall\, v^h \in \mathcal{V}_b^h \text{ for violated boundary conditions and} \tag{6.107}$$
$$(f, v^h)\tilde{}^h = \tilde{a}^h(u_0^h, v^h) = \tilde{a}_s^h(u_0^h, v^h) + \sum_{e \in \mathcal{T}^h} \int_e \frac{\partial u_0^h}{\partial \nu_e}[v^h] ds$$
$$\forall\, v^h \in \mathcal{V}_b^h \text{ for violated continuity, resp.} \tag{6.108}$$

Similarly to (4.42) and corresponding to (4.13), (4.29), (4.43), (4.44), $(f\ ,\ v^h)_{L^2(\Omega)}$, the error for the strong exact and discrete solution and for conforming methods has the form

$$\begin{aligned}\tilde{a}_s^h(u_0 - u_0^h, v^h) &= \sum_{T \in \mathcal{T}^h} \sum_{P_j \in T} w_j \big((A_s u_0) v^h\big)(P_j) - (f\ ,\ v^h)\tilde{}^h \\ &\quad - \sum_{T \in \mathcal{T}^h} \int_T \big((A_s u_0\ - f) v^h\big) dx\ + (f\ ,\ v^h)^h \\ &= \big((f\ ,\ v^h)^h - (f\ ,\ v^h)\tilde{}^h\big) \\ &\quad + \big(\tilde{a}_s^h(u_0, v^h) - a_s^h(u_0, v^h)\big).\end{aligned} \tag{6.109}$$

This requires smooth enough u_0, f to allow the evaluation of $A_s\, u_0(P_j), f(P_j)$. The estimate is again based on the quadrature errors in Ω.

Whenever a weak against a strong bilinear form has to be estimated, $\tilde{a}^h(u_0, v^h) - a_s^h(u_0, v^h)$, we have to add 0 and $\int_{\partial\Omega} \frac{\partial u_0^h}{\partial \nu}\, v^h ds$ and $\sum_{e \in \mathcal{T}^h} \int_e \frac{\partial u_0^h}{\partial \nu_e}[v^h] ds$ for conforming FEs and violated boundary conditions and violated continuity, resp.

We formulate the final result for collocation methods, in fact a Corollary of the above results, as

Theorem 6.6.2. *Let $a(\cdot,\cdot)$ be coercive, $\mathcal{U}_b^h$, $\mathcal{V}_b^h$ be conforming, (6.98) be valid and the quadrature formula be defined as in (6.76). Then the bilinear forms $a(;\cdot), a_s(;\cdot)$, $a_s^h(;\cdot)$, $\tilde{a}_s^h(;\cdot)$ satisfy e discrete inf–sup–condition. Hence unique strong exact and collocation solutions u_0 and u_0^h, resp., exist. The error between these solutions is estimated as*

$$\|u_0 - u_0^h\|_{H^2(\Omega)} \tag{6.110}$$
$$\leq C h^{m-2-n/2} \max_{i,j=0}^{n} \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|u\|_{W_\infty^m(\Omega)}^h, \quad or$$
$$\leq C\big(h^{m-n/2-\min\{m+\tau,m\}} \max_{i,j=0}^{n} \|a_{ij}\|_{W_\infty^m(\Omega)} \cdot \|u_0\|_{W_\infty^{m+2}(\Omega)}^h$$
$$+ h^{m-n/2-\min\{m+\tau,m\}} \|f\|_{W_\infty^m(\Omega)}\big) \tag{6.111}$$

Again the conditions for the functions and exponents of h to guarantee convergence are shown in the formulas.

Now, we compare the exact weak and (strong) collocation solutions u_0 and u_0^h, resp. for conforming FEs and violated boundary conditions and violated continuity, resp. For conforming FEs and violated boundary conditions we have to add 0 and $Ch^{\min\{\mu,m-2\}}\|u_0\|_{W_\infty^{\max\{m,2m'+1-\rho\}}(\Omega)}$, for violated continuity has to be modified as $Ch^{\min\{\mu',m-2\}}\|u_0\|_{H^{\max\{m,\mu'+2\}}(\Omega)}$, resp. These μ and μ' are defined in (6.34) and (??)

6.6.2 Variation Methods and Collocation for Spectral Methods

Compared to collocation for FEs the collocations for spectral methods are, theoretically, simpler for two reasons. There are no variational crimes except quadrature approximations, so there is no need for A_h, and the quadrature approximations for inner products for the terms $a(\cdot,\ \cdot)$ usually coincide with the exact integrals. We give a short review and extension of Subsection 4.2.2. Spectral-, pseudo-spectral-, collocation-methods in aliased and de-aliased forms represent different types of spectral methods. The power of these methods is observable only for smooth enough situations and simple domains. They are realized with trigonometric (Fourier) and Legendre or Chebyshev polynomial basis functions,.. The polynomials are indicated by $K = F$, $K = L$, $K = C$ below. The corresponding quadrature formulas use equidistant and Gauss-, Gauss-Radau- or Gauss-Lobatto collocation points, resp. Similar to Subsection 6.6.1 we want to relate a corresponding collocation method to the above approximations via quadrature. First we recall that spectral elements satisfy (Dirichlet or natural) boundary conditions exactly. So we apply (6.2) to $v^h \in \mathcal{V}_b^h$ starting with $v^h \in \mathcal{V}_b^h \subset \mathcal{V}_b$ satisfying Dirichlet boundary conditions. Since boundary conditions and continuity are satisfied we do have

$$\begin{aligned}(f, v^h) &= a(u_0^h, v^h) = a_s(u_0^h, v^h) \\ (f, v^h)\tilde{}^h &= \tilde{a}^h(u_0^h, v^h) = \tilde{a}_s^h(u_0^h, v^h) \\ &\forall\ v^h \in \mathcal{V}_b^h \quad \text{for our conforming spectral methods.}\end{aligned} \tag{6.112}$$

This implies that the strong and weak spectral approximation are identical. We obtain, with the weight function w and the quadrature approximation for the integral

$$\begin{aligned}a(u^h, v^h) &= \int_\Omega (A_s\ u^h)v^h dx = a_s(u^h, v^h) \approx \tilde{a}_s^h(u^h, v^h) := (A_s\ u^h, v^h)\tilde{}_w^h \\ &= (\tilde{A}_s^h\ u^h, v^h)\tilde{}_w^h = \sum_{j\in\mathbf{J}^N} (A_s u^h)(y_j)\overline{v^h(y_j)}w_j \\ &\forall\ u^h \in \mathcal{U}_b \quad \text{smooth enough, or } u^h \in \mathcal{U}_b^h \quad \forall \quad v^h \in \mathcal{V}_0^h,\end{aligned} \tag{6.113}$$

with $\overline{v^h(y_j)}$ the conjugate complex of $v^h(y_j)$. Often this $\approx$ can be replaced by $=$. In fact, for spectral methods we often have much stronger results than for Feds. Essentially Theorems 6.5.6 and 6.5.2 remain correct for spectral methods however allow a much stronger version. This is due to the fact, that compare (4.89), (4.94),

$$\begin{aligned}(u^h, v^h)_w = (u^h, v^h)^h_w \text{ for } \rho = -1, 0, 1 \ \forall \ u^h, v^h \in \mathcal{U}^h \text{ or } \in \mathcal{V}^h,\\ (u^h, v^h)_w = (u^h, v^h)^h_w (1 + \mathcal{O}(N^{-m+\iota_k(0)} \| u^h \cdot v^h \|_{H^m_w(\Omega)})) \\ \text{for } \rho = 2, \ \forall \ u^h, v^h \in \mathcal{U}^h \text{ or } \in \mathcal{V}^h \end{aligned} \tag{6.114}$$

We again define, e.g., for Dirichlet boundary conditions, an interpolation basis

$$\begin{aligned} u^h_i \in \mathcal{U}^h_b \text{ by } u^h_i(y_j) = 0 \ \forall j \in \mathbf{J}^N, y_j \in \partial\Omega, \\ u^h_i(y_j) = \delta_{ij} \ \forall \ \in \mathbf{J}^N, y_j \notin \partial\Omega, \end{aligned} \tag{6.115}$$

and similarly $v^h_i \in \mathcal{V}^h$. Then, due to (6.114)

$$\begin{aligned}(u^h_i, u^h_j)_w = (u^h_i, u^h_j)^{\tilde{}h}_w = \delta_{ij} \text{ for } \rho = -1, 0, 1, \\ = \delta_{ij} + \mathcal{O}(N^{-m+\iota_k(0)} \| u^h_i u^h_j \|_{H^m_w(\Omega)}). \text{ for } \rho = 2, \end{aligned}$$

So again the stability results in Theorem 6.6.2 remain valid. the convergence results have to be updated.
To really formulate the collocation equations, we choose as for FEs and as indicated above the interpolation basis for $\mathcal{V}^h_b$ as in (6.115). Then for (6.113) with Dirichlet conditions we have to determine

$$\begin{aligned} u^h_0 \in \mathcal{U}^h_0 \text{ s.t. } (A_s u^h_0)(y_k) = f(y_k) \ \forall \ k \in \mathbf{J}^N, y_k \notin \partial\Omega \\ \text{and } u^h_0(y_k) = 0 \ \forall \ k \in \mathbf{J}^N, y_k \in \partial\Omega. \end{aligned} \tag{6.116}$$

These represent the well known collocation methods. If instead of Dirichlet the natural or Neumann conditions are imposed, the collocation conditions in (6.116) are unchanged. However, the above $u^h_0(y_k) = 0$ have to be replaced by the corresponding, e.g., $(\partial u^h_0/\partial\nu)(y_k) = 0 \ \forall \ k \in \mathbf{J}^N, y_k \in \partial\Omega$ for Neumann conditions.

Theorem 6.6.3. *Choose $A_s, a(\cdot,\cdot), a_s(\cdot,\cdot) B_a$ and $\tilde{A}^h \approx A, \tilde{A}^h_s, a_s(\cdot,\cdot)$ $\tilde{a}^h_s(\cdot,\cdot)$ as in (6.2), and (6.113), resp., and the quadrature formulas as in (4.89), (6.113). Then for a $\mathcal{U}_b$ coercive $a(\cdot,\cdot)$ the $a(\cdot,\cdot), a_s(\cdot,\cdot), \tilde{a}^h(\cdot,\cdot), \tilde{a}^h_s(\cdot,\cdot)$ are $\mathcal{U}^h_b$ coercive again or satisfy, for $\mathcal{U}^h_b \neq \mathcal{V}^h_b$, the discrete inf–sup– conditions.*

Proof The coercivity claims are immediate consequences of (4.89) and the following remarks, implying $\tilde{a}^h(u^h, v^h) = (A \ u^h, v^h)^{\tilde{}h}_w = (\tilde{A}^h \ u^h, v^h)^{\tilde{}h}_w = a(u^h, v^h)$
$\forall \ u^h \in \mathcal{U}^h_b, \ v^h \in \mathcal{V}^h_b$ for $\rho = -1, 0, 1$ and for $\rho = 2$ if Au does not contain the term $a_{00}u$. Otherwise (6.114) or the averaged Taylor polynomials have to be used to show the discrete coercivity. □

Theorem 6.6.4. *For $u \in H^1(\Omega) a(\cdot,\cdot), a_s(\cdot,\cdot)$, and the* variational discretization errors *for $a(\cdot,\cdot)$ and $a_s(\cdot,\cdot)$ are zero . For $u \in H^m_w(\Omega)$ the* classical *and* variational consistency errors *for u and all spectral methods considered here, vanish of the order $m - \iota_K(2) - \mu_K$, with $\ell = 2$, $\iota_F(\ell) = \ell$, $\iota_C(\ell) = 2\ell$, $\iota_L(\ell) = 2\ell + n/2$, see (4.93), and $\mu_K := \mu_K(1)$, $\mu_F = 2$, $\mu_C = \mu_L = 4$, see(4.95) The* classical discretization error *can be estimated, with N independent $C = C_{(m,n,\rho)}$, by*

$$\|\tilde{A}^h I^h u - Q'^h A u\|_{\mathcal{V}_b^{h'}} \le C N^{-m+\iota_K(2)+\mu_K} \max_{i,j=0}^{n} \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|u\|_{H^m_w(\Omega)} \|v^h\|_{H^1_w(\Omega)}. \tag{6.117}$$

This indicates the exponentially good consistency for $m \to \infty$. Let u_0 and u_0^h be an exact and the approximate solution defined by

$$a(u_0, v) = f(v) \ \forall\ v \in \mathcal{V}_b \text{ and } \tilde{a}^h(u_0^h, v^h) = \tilde{f}^h(v^h) \ \ \forall\ v^h \in \mathcal{V}_b^h,$$

or the strong or collocation analogues, see (6.113), (6.116).

Finally, let $\tilde{a}^h(u_0, v^h)$ and $\tilde{a}_s^h(u_0, v^h)$ be defined for u_0 and $\forall\ v^h \in \mathcal{V}_b^h$. Then we find for the variational consistency errors *with C independent of h,*

$$\sup_{0 \ne v^h \in \mathcal{V}_b^h} \frac{|\tilde{a}^h(u_0^h - u_0, v^h)|}{\|v^h\|^h_{\mathcal{V}}} \text{ and} \tag{6.118}$$

$$\sup_{0 \ne v^h \in \mathcal{V}_b^h} \frac{|a^h(u_0, v^h) - \tilde{a}^h(u_0, v^h)|}{\|v^h\|^h_{\mathcal{V}}} + \sup_{0 \ne v^h \in \mathcal{V}_b^h} \frac{|f^h(v^h) - \tilde{f}^h(v^h)|}{\|v^h\|^h_{\mathcal{V}}} \le C N^{-m+\iota_K(2)+\mu_K} \max_{i,j=0}^{n} \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|u\|_{H^m_w(\Omega)}.$$

For a U_h coercive $a(.,.)$ we obtain the estimate

$$\|u_0^h - u_0\|^h_{H'(\Omega)} \le C N^{-m+\iota_K(2)+\mu_K} \max_{i,j=0}^{n} \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|u_0\|_{H^m_w(\Omega)}, \tag{6.119}$$

here u_0^h is any of the weak or strong or quadrature approximate or collocation spectral solutions.

For spectral methods often $(\tilde{A}^h u^h,\ v^h)^{\sim h}_w = (Au^h,\ v^h)_w$ and $f(v^h) = \tilde{f}^h(v^h)$ for simple enough f are valid. Then we obtain the improved estimate

$$\|u_0^h - u_0\|^h_{H'(\Omega)} \le C N^{-m+\iota_K(2)} \max_{i,j=0}^{n} \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|u_0\|_{H^m_w(\Omega)}. \tag{6.120}$$

Instead of the condition $f(v^h) = \tilde{f}^h(v^h)$ we can replace the $f(P_{j,T})$ in the collocation equations by sufficiently good approximations $(f, v^h_{j,T})^{\sim h}_w$.

Proof We use a modification of (6.90) based on Notation 4.1, the quadrature formula (6.113) and the Cauchy-Schwarz inequality instead of (6.76). With

the notations and results in Theorem 4.2.1 we estimate for $u \in H^m_w(\Omega)$, see (6.113)

$$\begin{aligned}
\left| < \tilde{A}^h I^h u - \tilde{Q}'^h A u,\ v^h \tilde{>}_{\mathcal{V}' \times \mathcal{V}^h_b} \right| &= \left| (\tilde{A}^h I^h u - \tilde{Q}'^h A u,\ v^h)^h_w \right| \\
&\leq C \max_{i,j=0}^{n} \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|I^h_K u - u\|_{H^2_w(\Omega)} \|v^h\|_{L^2_w(\Omega)} \qquad (6.121) \\
&\leq C N^{-m+\iota_K(2)} \max_{i,j=0}^{n} \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|u\|_{H^m_w(\Omega)} \|v^h\|_{L^2_w(\Omega)} \qquad (6.122) \\
&\leq C N^{-m+\iota_K(2)+\mu_K} \max_{i,j=0}^{n} \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|u\|_{H^m_w(\Omega)} \|v^h\|_{H^1_w(\Omega)}.
\end{aligned}$$

The exponents $\iota_K(2)$ are caused by the interpolation errors, see (4.94). The exponents μ_K are due to inverse estimates for spectral elements, see (4.94), ([23]. Since in the final error estimates this result has to be combined with $\|I^h_K u - u\|_{H^1_w(\Omega)} = \mathcal{O}(N^{-m+\iota_K(2)}) < \mathcal{O}(N^{-m+\iota_K(2)+\mu_K})$ for large enough N we have lost $\mathcal{O}(N^{\mu_K})$ compared to the optimal result. With (6.113)this implies immediately

$$\begin{aligned}
|a(u, v^h) - \tilde{a}^h(u, v^h)| &\leq C N^{-m+\iota_K(2)+\mu_K} \\
&\cdot \max_{i,j=0}^{n} \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|u\|_{H^m_w(\Omega)} \|v^h\|_{H^1_w(\Omega)} \qquad (6.123)
\end{aligned}$$

Similarly, we find

$$\begin{aligned}
|f(v^h) - \tilde{f}^h(v^h)| &\leq C \|I^h_K u - u\|_{L^2_w(\Omega)} \|v^h\|_{L^2_w(\Omega)} \qquad (6.124) \\
&\leq C N^{-m+\iota_K(0)} \|u\|_{H^m_w(\Omega)} \|v^h\|_{L^2_w(\Omega)} \\
&\leq C N^{-m+\iota_K(0)+\mu_K} \|u\|_{H^m_w(\Omega)} \|v^h\|_{H^1_w(\Omega)}
\end{aligned}$$

The last estimate follows as above:

$$|\tilde{a}^h(u^h_0 - u_0, v^h)| \leq |\tilde{f}^h(v^h) - f(v^h)| + |(a^h(u_0, v^h) - \tilde{a}^h(u_0, v^h))|$$

This is the above claim in (6.118). The estimate (6.119) is obtained by combining (6.118) and Theorem (6.6.3).

For spectral methods often $(\tilde{A}^h I^h u,\ v^h)^h_w = (A I^h u,\ v^h)_w$ and $f(v^h) = \tilde{f}^h(v^h)$ for simple enough f are valid. This implies $\mu_K(1) = 0$. □

6.7 Consistency for Nonlinear Equations

$$B_a\ u := \sum_{i,j=1}^{n} \nu_j a_{ij} \partial_i\ u + \sum_{j=1}^{n} \nu_j a_{0j}\ u. - \sum_{j=1}^{n} \partial_j\ (a_{0j} u) \qquad (6.125)$$

We do not want do burden the presentation with the full machinery. Instead we modify (4.96), and discuss the following

Example 6.1 *of a nonlinear model problem with Dirichlet boundary conditions and A_s of the form in (3.26)*

$$G_s : H^2(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega), G_s(u) := A_s u + \lambda R(u) \tag{6.126}$$
$$= A_s u + \lambda R_e(u, -\sum_{j=1}^{n} \partial_j \ (a_{0j} u), \int_{\Omega_0} u), \ \ e.g.,$$
$$= A_s u + \lambda(u^2 - \sum_{j=1}^{n} \partial_j \ (a_{0j} u)(u + \int_{\Omega_0} u) + g).$$

Similarly we can discuss natural boundary conditions.

We develop $G_s(u_0 + u)$, for a fixed u_0 into terms, which are independent of u and linear or quadratic in u. We obtain

$$G_s(u_0 + u) = G_s(u_0) + \tag{6.127}$$
$$+A_s u + \lambda(2u_0 u - (u + \int_\Omega u) \sum_{j=1}^{n} \partial_j \ (a_{0j} u_0) + (u_0 + \int_\Omega u_0) \sum_{j=1}^{n} \partial_j \ (a_{0j} u))$$
$$+\lambda(u^2 - (u + \int_\Omega u) \sum_{j=1}^{n} \partial_j \ (a_{0j} u)).$$

Now we test $G_s(u_0 + u) = 0$ in the strong $(G_s u, v)_{L^2(\Omega)}$ form and relate it to the $G(u_0 + u)$ testing in the weak form $< Gu, v >_{H^{-1}(\Omega) \times H^1(\Omega)}$. With the notations

$$A_s^{ex} u := G_s'(u_0) u := A_s u$$
$$+\lambda(2u_0 u - (u + \int_\Omega u) \sum_{j=1}^{n} \partial_j \ (a_{0j} u_0) - (u_0 + \int_\Omega u_0) \sum_{j=1}^{n} \partial_j \ (a_{0j} u)),$$
$$G_s''(u_0) \frac{u^2}{2} := \lambda(u^2 - (u + \int_\Omega u) \sum_{j=1}^{n} \partial_j \ (a_{0j} u)$$

we have to test

$$G_s(u_0 + u) = G_s(u_0) + G_s'(u_0) u + G_s''(u_0) \frac{u^2}{2} \ \text{ with}$$
$$G_s(u_0) \in L^2(\Omega), \ \text{ and } A_s^{ex} = G_s'(u_0) u : H^2(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega),$$
$$A^{ex} = G'(u_0) u : H_0^1(\Omega) \to H^{-1}(\Omega).$$

So testing in the strong or weak form yields

$$(G_s(u_0), v)_{L^2(\Omega)} = < G_s(u_0), v >_{H^{-1}(\Omega) \times H^1(\Omega)} \quad \forall v \in H^1(\Omega)$$
$$(G_s'(u_0) u, v)_{L^2(\Omega)} = (A_s^{ex} u, v)_{L^2(\Omega)} =< A^{ex} u, v >_{H^{-1}(\Omega) \times H^1(\Omega)}$$
$$- \int_{\delta\Omega} (B_a^{ex} u) v \ ds, \ \forall v \in H^1(\Omega). \tag{6.128}$$

Here A^{ex} and B_a^{ex} indicate the weak linear operator and the natural boundary conditions generated by A_s^{ex}.

Replacing in (6.128) the u, v by $u^h \in \mathcal{U}_b^h,\ v^h \in \mathcal{V}_b^h$ yields the consistency errors familiar from Sections 6.1 -6.6.

Finally, we have to test $G_s''(u_0)u^2/2$. In fact, we have changed the original form (4.96) into (6.126) to guarantee a nonstandard consistency error, which does not immediately fit to the earlier Sections 6.1 -6.6. The only interesting term is the

$$-\int_\Omega u \sum_{j=1}^n \partial_j\ (a_{0j}u)\ v^h\ dx\ =\ \int_\Omega \sum_{j=1}^n a_{0j}\ u\ \partial_j(uv^h)\ dx - \int_{\delta\Omega} uv^h B_{nl}u\ ds \quad \text{with}$$

$$-\int_{\delta\Omega} uv^h B_{nl}u\ ds := -\int_{\delta\Omega} \left(\sum_{j=1}^n uv^h\ \nu_j\ a_{0\,j}\ u\right) ds \quad \text{or}$$

$$B_{nl}u := \sum_{j=1}^n \nu_j\ a_{0\,j}\ u$$

Again we have to estimate, as in subsections 6.1-6.6 either

$$\int_{\delta\Omega} uv^h\ B_{nl}u\ ds \text{ or } \sum_{e\in\mathcal{T}^h} \int_e [uv^h] B_{nl}u\ ds \tag{6.129}$$

for smooth enough u and $v^h \in \mathcal{V}_b^h$. For the estimates in subsections 6.1 -6.3 we had studied a $B_a u$. In the proof we had only used estimates of the form

$$\|B_a u\|_{W_p^l(T)} \le C\|u\|_{W_p^{l+1}(T)}, \tag{6.130}$$

since B_a was a first order differential operator. Since our actual $B_{nl}u$ is only of the order zero we can replace (6.130) by

$$\|B_{nl}u\|_{W_p^l(T)} \le C\|u\|_{W_p^l(T)},$$

At the other side, for smooth enough u we can still use the other complementary parts of these proof : They require $uv^h = 0$ in enough points on $\partial\Omega$ or on e and smooth enough uv^h.
This strategy always works for nonlinear problems, unless $-\sum_{j=1}^n \partial_j\ (a_{0j}u)$ appears in to high powers. E.g., if we replace $u - \sum_{j=1}^n \partial_j\ (a_{0j}u)$ above by $u(-\sum_{j=1}^n \partial_j\ (a_{0j}u))^2$ and test with v^h we still find

$$\int_\Omega vu(-\sum_{j=1}^n \partial_j(a_{0j}u))^2 dx = -\int_\Omega \sum_{j=1}^n \partial_j(a_{0j}uv^h) \sum_{j=1}^n \partial_j(a_{0j}u)dx - \int_{\delta\Omega} uv\ B_{nl}u\ ds$$

with $B_{nl}u$ a differential operator of zero order. However if we consider $u(-\sum_{j=1}^n \partial_j\ (a_{0j}u))^3$ this approach fails and other techniques have to be

developed.
Whenever these techniques do apply, we have estimates for the consistency error of the same magnitude as for the linear problems considered above. We summarize:

Theorem 6.7.1. *Let $G_s : H^2(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega)$ be a nonlinear operator, obtained as an appropriate generalization of (6.126), s.t. derivatives of u do not occur in powers greater than 2. Then the consistency errors in subsections (6.1)-(6.6) remain valid with increased constants.*
If $G_s'(u_0)$ indicates a $\mathcal{U}_b$ coercive bilinear form, and u_0 and u_0^h are the exact and discrete solution, $G_s(u_0) = 0$ and $G_s^h(u_0^h) = 0$ then

$$\|u_0 - u_0^h\|_{H^1(\Omega)}^h \leq C \left\{ \inf_{u^h \in \mathcal{U}_b^h} \|u_0 - h^h\|_{H^1(\Omega)}^h + \quad \textit{consistency error} \right\}. \tag{6.133}$$

We do not want to repeat all the different formulas for the consistency errors in the last subsections. One simply has to combine, e.g. violated continuity with Theorem (6.3.1).

7. Stability for General Elliptic Operators and Variational Crimes

Based on the general results for discretization methods in Chapter 4 we have formulated the generalized Strang Lemmas in Chapter 5 and estimated the consistency errors for the different cases in Chapter 6. Additionally we have shown the coercivity or inf–sup– condition, hence the stability of the discrete operators, only for coercive original bilinear forms. So, we still have to prove stability under the influence of variational crimes and for general elliptic operators. We achieve this by the fact that general elliptic bilinear forms are obtained as compact perturbations of coercive original bilinear forms. Further extensions to Navier-Stokes equations and to bifurcation numerics are studied in [15, 16]. In Chapter 6 we have determined the consistency errors independent of the stability. Hence, the general (linear and nonlinear) stability in this Chapter allows a combination with the consistency results in Chapters 4 , particularly Section 4.4, - 6 to yield the desired convergence. We do not repeat all the detailed consistency conditions and results. Rather we formulate a final result, Theorem 7.3.1 indicating the convergence results.

We have mentioned at the beginning of Chapter 4 that the generalization to elliptic operators of order $2m$ works well in Chapters 2- 5 and it will work well in this Chapter 7 again. However, the construction of the anti crime operator in Chapter 2, the consistency estimates, the $\mathcal{U}^h-$ coercivity and inf–sup– condition w.r.t $\mathcal{U}^h,\ \mathcal{V}^h$ for variational crimes in Chapter 6 are based on specific $2m = 2-$ techniques. So, only these parts would have to be worked out for the general case $2m$ to generalize the whole results to elliptic operators of order $2m$.

We prove the uniform $\inf - \sup -$ condition, allowing variational crimes, for bounded bilinear forms $a(\cdot,\cdot)$ inducing an elliptic invertible operator A. This implies, by Theorem 2.1.7, the stability of the induced linear operators A^h. Again, we use the general notations in Section 4.3.

The $\mathcal{U}_b^h-$coercivity of $a^h(\cdot,\cdot)$ or the inf–sup– condition for $\mathcal{U}_b^h \neq \mathcal{V}_b^h$ for a $\mathcal{U}_b-$coercive $a(\cdot,\cdot)$ is proved already in Chapter 6. We generalized this in several steps to compact perturbations of the linear operator A, their approximation perturbations, e.g., by quadrature rules. This even applies to their bordered forms as needed in bifurcation numerics, [15, 16].

7.1 General Definitions and Results

We start citing well known results and definitions needed in our context.

The two following Theorems give known and well applicable criteria for stability. We discuss the relation between stability of the nonlinear and linear and a perturbed nonlinear problem. For the general case studied in [57] we need compatible approximations: The projection operators $Q'^h, \tilde{Q}'^h$ in (4.109) and the E^h introduced in Chapter 2.6, have the properties:

Definition 7.1.1. Compatible approximations: *Let for all $h \in H$ the $P^h : \mathcal{U} \to \mathcal{U}^h, Q'^h : \mathcal{V}' \to \mathcal{V}'^h$, $\tilde{Q}'^h : \mathcal{V}' \cap C(\overline{\Omega}) \to \mathcal{V}'^h$ and $E^h : \mathcal{U}^h \to \mathcal{U}$, be (uniformly) bounded linear approximation and extension operators, resp., with*

$$\begin{aligned} \sup_h \|P^h\|_{\mathcal{U}^h \leftarrow \mathcal{U}} < \infty, \quad \lim_{h\to 0} ||u - P^h u||_{\mathcal{U}}^h = 0 \ \forall\ u \in \mathcal{U}, \\ \sup_h \|Q'^h\|_{\mathcal{V}'^h \leftarrow \mathcal{V}'} < \infty, \quad \lim_{h\to 0} ||f - Q'^h f||_{\mathcal{V}'} = 0 \ \forall\ f \in \mathcal{V}', \qquad (7.1) \\ \sup_h \|\tilde{Q}'^h\|_{\mathcal{V}'^h \leftarrow \mathcal{V}' \cap C(\overline{\Omega})} < \infty, \quad \lim_{h\to 0} ||f - \tilde{Q}'^h f||_{\mathcal{V}'} = 0 \ \forall\ f \in \mathcal{V}' \cap C(\overline{\Omega}), \\ \sup_h \|E^h\|_{\mathcal{U} \leftarrow \mathcal{U}^h} < \infty, \quad \lim_{h\to 0} ||u^h - E^h u^h||_{\mathcal{U}}^h = 0 \ \ \forall\ u^h \in \mathcal{U}^h. \qquad (7.2) \end{aligned}$$

Let the properties in this and in Definition 4.3.1 be satisfied. Then we call $\mathcal{U}^h, \mathcal{V}^h$ a compatible approximation.

Remark 7.1.2. The following Theorem 7.1.3 and the compatibility in Definition 7.1.1 are satisfied for our FE and spectral approximations, E^h only for $\mathcal{U} = H^1(\Omega)$. The $\lim_{h\to 0} ||u^h - E^h u^h||_{\mathcal{U}}^h = 0 \ \ \forall\ u^h \in \mathcal{U}^h$ implies $\lim_{h\to 0} ||u - E^h P^h u||_{\mathcal{U}} = 0 \ \ \forall\ u \in \mathcal{U}$. Nevertheless, we present the essential proof for this more general case.

We only need these compatible approximations for the proof of stability for compact perturbations of monotone operators. Zeidler [67, 68] requires $\lim_{h\to 0} ||u - E^h P^h u||_{\mathcal{U}} = 0 \ \ \forall\ u \in \mathcal{U}$. The approximation property (7.1) implies, necessary for the Stetter approach,

$$\begin{aligned} \lim_{h\to 0} ||P^h u||_{\mathcal{U}}^h = ||u||_{\mathcal{U}} \ \forall u \in \mathcal{U}, \quad \lim_{h\to 0} ||Q'^h f||_{\mathcal{V}'} = ||f||_{\mathcal{V}'} \ \ \forall\ f \in \mathcal{V}', \qquad (7.3) \\ \lim_{h\to 0} ||\tilde{Q}'^h f||_{\mathcal{V}'} = ||f||_{\mathcal{V}'} \ \ \forall\ f \in \mathcal{V}' \cap C(\overline{\Omega}). \end{aligned}$$

Now, we have, see [57] and Remark 7.1.2

Theorem 7.1.3. *Let $\mathcal{U}_b^h$, $\mathcal{V}_b^h$ be admissible and compatible approximations and let the induced dicretization $G^h : \mathcal{U}_b^h \to \mathcal{V}_b^{h'}$ and $u^h \in \mathcal{U}^h$ and $r > 0$ be given. Assume G^h is continuous in $B_r(u^h)$. Furthermore let for all $v^h \in \mathcal{U}_b^h \cap B_r(u^h)$ the $((G^h)'(v^h))^{-1}$ exist and*

$$\|((G^h)'(v^h))^{-1}\| \leq S \text{ uniformly for } h \in H. \tag{7.4}$$

Then the discretization G^h is stable at u^h with stability bound S and stability threshold $r_0 = r/S$.

Theorem 7.1.4. *Let $\mathcal{U}_b^h$, $\mathcal{V}_b^h$ be admissible and compatible approximations. Let the induced discretization $\overline{G}^h := G^h + F^h : \mathcal{U}_b^h \to \mathcal{V}_b^{h'}$ for $\overline{G} = G + F : \mathcal{U}_b \to \mathcal{V}_b'$ and let a sequence $u^h \in \mathcal{U}_b^h$ be given. Assume there exists $r, L > 0$, independent of h, s.t.*

1. *G^h are continuous and stable in $B_r(u^h)$ with the stability bound S for G^h,*
2. *F^h are Lipschitz continuous in $B_r(u^h)$ with*

$$\|F^h(u_1^h) - F^h(u_2^h)\|_{\mathcal{V}_b'}^h \leq L\|u_1^h - u_2^h\|_{\mathcal{U}}^h \quad \text{and } L \cdot S < 1.$$

Then $\overline{G}^h$ is stable in u^h, with stability bound $S/(1-LS)$ and stability threshold $(1 - LS)r/S$.

By Theorem 7.1.3 we can restrict the stability proof to linear operators and the corresponding bilinear forms. We have reviewed the relation already between the discrete operators and bilinear forms many times. We repeat here for convenience the above notations:

Notation 7.1 *Let $a^h(\cdot,\cdot) : \mathcal{U}_b^h \times \mathcal{V}_b^h \to \mathbb{R}$, $f^h : \mathcal{V}_b^h \to \mathbb{R}$ denote a bilinear and a linear form. Let A^h denote the linear operators, induced by the above $a^h(\cdot,\cdot)$, hence*

$$< A^h u^h, v^h >= a^h(u^h, v^h) \quad \forall u^h \in \mathcal{U}_b^h,\ v^h \in \mathcal{V}_b^h.$$

Typically for the variational crimes we have $\mathcal{U}_b^h \not\subset \mathcal{U}_b$ or $\not\subset \mathcal{U}$, $\mathcal{V}_b^h \not\subset \mathcal{V}_b$ or $\not\subset \mathcal{V}$, or allow quadrature approximations or collocation methods. If we explicitly want to indicate the quadrature approximations, we use the notations $\tilde{A}^h, \tilde{a}^h(\cdot,\cdot)$ instead of $A^h, a^h(\cdot,\cdot)$. This implicitly defines linear operators

$$\Phi^h : \mathcal{L}(\mathcal{U}_b, \mathcal{V}_b) \to \mathcal{L}(\mathcal{U}_b^h, \mathcal{V}_b^h) \quad \text{s.t., } \Phi^h(A) = A^h \text{ with } \Phi^h(A + C) = \Phi^h(A) + \Phi^h(C). \tag{7.5}$$

Analogously, Φ^h is defined for nonlinear operators as well.

For the following discussion we have to consider the dual problem to (4.113): For $A^d : \mathcal{V}_b'' \to \mathcal{U}_b'$ we have to determine

$$v_0 \in \mathcal{V}_b'' \text{ s.t. } A^d v_0 = g' \in \mathcal{U}_b';$$

the $(A^d)^{-1}$ exists if and only if A^{-1} exists.

For the dual problem it is important, that the $\mathcal{U}^{h'}, \mathcal{V}^h = \mathcal{V}^{h''}$ define corresponding dual operators, $P^{h'} : \mathcal{U}_b' \to \mathcal{U}_b^{h'}$, defined as in (4.109) or in (4.108), e.g. $P^{h'} g' = g'|_{\mathcal{U}_b^h}$, and Q^h, E_v^h or their approximations. Therefore the situation in (4.118) implies usually a corresponding dual equation, $A^d v = g' \in \mathcal{U}_b', v \in \mathcal{V}_b''$, and a dual diagram of the form

$$\begin{array}{ccccccc} \mathcal{V}'' \subset & \mathcal{V}_b'' & \overset{A^d, \tilde{A}^d}{\longrightarrow} & \mathcal{U}_b' & \overset{exactly\ tested\ by}{\longleftrightarrow} & \mathcal{U}_b \subset \mathcal{U} \\ & Q^h \downarrow \uparrow E_v^h & \Psi^h \downarrow & \downarrow P'^h \text{or} \tilde{P}'^h & & \\ & \mathcal{V}_b^h = (\mathcal{V}_b^h)'' & \underset{(A^d)^h,\ (\tilde{A}^d)^h}{\longrightarrow} & \mathcal{U}_b^{h'} & \overset{(appr.)\ tested\ by}{\longleftrightarrow} & \mathcal{U}_b^h. \end{array} \tag{7.6}$$

Using the techniques to construct E^h, we can define an analogous E_v^h as well. Mind that $\mathcal{U}^h \subset \mathcal{U}$ implies $\mathcal{U}'^h = (\mathcal{U}^h)' \supset \mathcal{U}'$ in the sense, that any linear functional defined on $\mathcal{U}$ is defined on $\mathcal{U}^h$ as well, not vice versa. Nevertheless, $\mathcal{U}'$ is infinite-, and $\mathcal{U}'^h$ finite-dimensional.

For the case of Petrov-Galerkin methods we obtain, with $\mathcal{V}_b^h = (\mathcal{V}_b^h)''$, satisfied for our FE and spectral approximations,

$$P^{h'} g' = g'|_{\mathcal{U}_b^h} \text{ and } (A^d)^h v_0^h := P^{h'} A_h^d v_0^h = P^{h'} g', \quad v_0^h \in \mathcal{V}_b^h = (\mathcal{V}_b^h)''. \tag{7.7}$$

So we have, for Petrov-Galerkin methods, a pair of corresponding discrete equations of the form

$$A^h u_0^h = Q'^h A_h|_{\mathcal{U}_b^h} u_0^h = Q'^h f \text{ and } (A^d)^h v_0'^h = P'^h A_h^d|_{\mathcal{U}_b^h} v_0'^h = Q'^h f' \tag{7.8}$$

In case of variational crimes, we have to generalize the operators and consistency errors as we did for the original equation, see Notation 4.1. This is already included in the chosen general notation of A^h and $a^h(\cdot,\cdot)$.

Definition 7.1.5. Bi-dual approximating spaces: *Let the conditions in Definitions 4.3.1, 4.3.3, and 7.1.1 be satisfied as formulated for the dual situation, hence for $\mathcal{U}_b'^h$ and $\mathcal{V}_b^h$ and let Q^h, P'^h or $\tilde{P}'^h, E_v^h$ be the corresponding operators. We require*

$$\begin{aligned} &dist(u, \mathcal{U}^h) = \inf_{u^h \in \mathcal{U}^h} ||u - u^h||_{\mathcal{U}}^h \to 0 \quad \text{for } h \to 0 \ \forall\ u \in \mathcal{U}\ , \text{ and} \\ &dist(v, \mathcal{V}^h) = \inf_{v^h \in \mathcal{V}^h} ||v - v^h||_{\mathcal{V}}^h \to 0\ v \in \mathcal{V} \textbf{ and } \ v \in (\mathcal{V})''. \end{aligned} \tag{7.9}$$

Then $\mathcal{U}_b^h, \mathcal{V}_b^h$ and $\mathcal{U}_b'^h$, $\mathcal{V}_b^h = (\mathcal{V}_b^h)$" is called a pair of bi-dual (Petrov-Galerkin) admissible approximating spaces. *They are called* admissible *and* compatible *if* $\dim \mathcal{U}_b^h = \dim \mathcal{V}_b^h$ *and the conditions in Definition 7.1.1 are satisfied, resp.*

Remark 7.1.6. For our FE and spectral approximations, the conditions in Definitions 4.3.1, 4.3.3, 7.1.1 and 7.1.5 are satisfied for the dual situation as well.

The bi-duality condition is satisfied, whenever $\mathcal{V}$ is dense in $\mathcal{V}$". This is correct foe the cases studied here.

Under these assumptions we obtain an estimate for $||v_0 - v_0^h||_{\mathcal{V}''}$ similar to those for $\|u_0 - u_0^h\|_{\mathcal{U}}$ in (3.30), in which merely $\mathrm{dist}(u_0, \mathcal{U}^h)$ has to be replaced by $\mathrm{dist}(v_0, \mathcal{V}^h)$ in $\mathcal{V}''$. Variational crimes can be handled similarly.

7.2 Stability for Variational Crimes

Now, we turn to th problem to prove stability for compact perturbations of an operator with stable discretization.

It is well known that B^h is stable if and only if $(B^h)^d$ is stable, and for both sequences the same stability constants can be chosen, see Criterion 7.1 below. This can more easily be verified, in particular for the special case $B^h = Q'^h A|_{\mathcal{U}_b^h}$. For the cases studied in Chapter 6 we can apply those techniques to the dual operator as well. This shows, that for an important class of operators A^h and $(A^h)^d$ are simultaneously stable.

We have simultaneous consistency estimates as well for bi-dual approximating spaces, even for variational crimes. We need the relation between the stability of the above discrete A^h and the existence of A^{-1}. The next Theorem shows that stability is the stronger condition. The $(A^h)^{-1}$ can only be stable, if, here and below, the $(A^h)^{-1} \in \mathcal{L}(\mathcal{V}_b^{h'}, \mathcal{U}_b^h)$ and the $\mathcal{U}_b^h, \mathcal{V}_b^h$ are chosen as admissible approximations in the sense of Definition 4.3.1.

Theorem 7.2.1. *For $A \in \mathcal{L}(\mathcal{U}_b, \mathcal{V}_b')$ and a pair of bi-dual admissible approximations, $\mathcal{U}_b^h, \mathcal{V}_b^h$, see Definition 7.1.5, let A^h, determined by $a^h(\cdot,\cdot)$, be stable and A^h and $(A^h)^d$ be consistent. Then $A^{-1} \in \mathcal{L}(\mathcal{V}_b', \mathcal{U}_b)$ exists. The corresponding result for $\tilde{A}^h$ is formulated in Theorem 7.2.6.*

For the proof we need as in [16] the following Criterion.

Criterion 7.1 *Let Banach spaces $\mathcal{U}_b \subset \mathcal{U}, \mathcal{V}_b' \subset \mathcal{V}'$, and $A \in \mathcal{L}(\mathcal{U}_b, \mathcal{V}_b')$ and its dual $A^d \in \mathcal{L}(\mathcal{V}_b^{''}, \mathcal{U}_b')$ be given. Then the following two conditions are mutually equivalent:*

1. $A^{-1} \in \mathcal{L}(\mathcal{V}_b', \mathcal{U}_b) \Leftrightarrow ((A^d)^{-1} \in \mathcal{L}(\mathcal{U}_b', \mathcal{V}_b^{''}))$.
2. *There exist positive constants C_1, C_2 such that*
   a) $||Au||_{\mathcal{V}_b'} \geq C_1||u||_{\mathcal{U}}$ *for all* $u \in \mathcal{U}_b$ *and*
   b) $||A^d v''||_{\mathcal{U}'} \geq C_2||v''||_{\mathcal{V}_b''}$ *for all* $v'' \in \mathcal{V}_b^{''}$.

If one of the above conditions is satisfied, we obtain

$$||A^{-1}||_{\mathcal{U}_b \leftarrow \mathcal{V}_b'} = ||(A^d)^{-1}||_{\mathcal{V}_b'' \leftarrow \mathcal{U}_b'} \leq \min(1/C_1, 1/C_2).$$

Remark 7.2.2. a) Since stability requires a uniformly bounded invertability, the above conditions have to be modified for the discrete operators in an obvious way.

b) For our present state of proof we know the stability of the A^h induced

by a $\mathcal{U}_b$ coercive $a(\cdot,\cdot)$. We have to observe that Theorem 7.2.1 is applicable only to those A, for which the stability of A^h is assumed.

Proof We have to show that the existence of a bounded inverse of A follows from the stability of $(A^h)_{h\in H}$. Let $u \in \mathcal{U}_b$ and $v \in \mathcal{V}_b''$ be given. Consider $f := Au \in \mathcal{V}_b'$, $g := A^d v \in \mathcal{U}_b'$ and the corresponding discrete solutions u^h, v^h of (4.115) and (7.7), respectively. The latter are uniquely determined due to the assumed stability. One obtains the following estimates:

$$\begin{aligned}
||u^h||_{\mathcal{U}}^h &= ||(A^h)^{-1}Q'^h f||_{\mathcal{U}}^h = ||(A^h)^{-1}Q'^h Au||_{\mathcal{U}}^h \\
&\le ||(A^h)^{-1}||_{\mathcal{U}_b^h \leftarrow \mathcal{V}_b^{h'}} ||Q'^h Au||_{\mathcal{V}'}^h \\
&\le ||(A^h)^{-1}||_{\mathcal{U}_b^h \leftarrow \mathcal{V}_b^{h'}} ||Au||_{\mathcal{V}'} \le C||Au||_{\mathcal{V}'} \quad \text{and} \\
||v^h||_{\mathcal{V}}^h &= ||((A^d)^h)^{-1}P'^h g||_{\mathcal{V}}^h = ||((A^d)^h)^{-1}P'^h A^d v||_{\mathcal{V}}^h \\
&\le ||((A^d)^h)^{-1}||_{\mathcal{V}_b^h \leftarrow \mathcal{U}_b^{h'}}^h ||P'^h A^d v||_{\mathcal{U}'}^h \\
&\le ||(A^h)^{-1}||_{\mathcal{U}_b^h \leftarrow \mathcal{V}_b^{h'}} ||A^d v||_{\mathcal{U}'} \le C||A^d v||_{\mathcal{U}'},
\end{aligned}$$

The results in Chapters 5 and 6, and the pair of bi-dual, necessarily admissible, approximations, yield for the simultaneously stable and consistent A^h and $(A^h)^d$ the convergence of u^h and v^h to the fixed u and v, respectively. Consequently, we obtain with the above estimates and the continuity of the norms:

$$||Au||_{\mathcal{V}'} \ge K||u||_{\mathcal{U}} \quad \text{und} \quad ||A^d v||_{\mathcal{U}'} \ge K||v||_{\mathcal{V}}.$$

Therefore Criterion 7.1 implies that $A^{-1} \in \mathcal{L}(\mathcal{V}_b', \mathcal{U}_b)$ does exist. □

For an important class of operators, the so-called strongly monotone and coercive and elliptic operators, depending upon the author, [68], and [42], and [37], respectively, the stability has been verified already in Theorems 6.1.1, 6.4.1, 6.3.1 and 6.5.3. In general the existence of a bounded inverse of A is not sufficient to ensure that the discrete A^h or $\tilde{A}^h$ is stable, see [51]. However, the next result allows all types of elliptic equation, bordered systems, hence bifurcation numerics, and Navier-Stokes equations. We have presented this in detail in [51] and we will show, that it remains valid in the case of variational crimes as well.

Theorem 7.2.3. *Let $\mathcal{U}_b^h, \mathcal{V}_b^h$ define bi-dual admissible approximations, with $\mathcal{U}_b^h$ approximating $H^1(\Omega) = \mathcal{U}$. For $B \in \mathcal{L}(\mathcal{U}_b, \mathcal{V}_b')$, let the discrete B^h (or $\tilde{B}^h$) be stable. Furthermore, let $A = B + C$, with a compact perturbation C of A, let B^h and C^h be consistent with B and C for smooth u, resp. Then*

$$A^h \text{ (or } \tilde{A}^h) \text{ is stable } \Leftrightarrow A^{-1} \in \mathcal{L}(\mathcal{V}_b', \mathcal{U}_b). \tag{7.10}$$

Remark 7.2.4. We have indicated already by A^h (or $\tilde{A}^h$) in (7.10), that the following proof remains valid, if A^h, Q'^h are replaced by $\tilde{A}^h, \tilde{Q}'^h$. For our main examples of A, B, C see Section 3.2, this consistency of B, C (and A) has been proved in Chapter 6 under the conditions of Theorem 7.2.6. Whenever

we apply this result to nonlinear problems the approximating spaces have to be compatible as well

Proof As a consequence of Theorem 7.2.1 it suffices to show that $A^{-1} \in \mathcal{L}(\mathcal{V}_b', \mathcal{U}_b)$ implies the stability of A^h. We determine for an arbitrary $u \in \mathcal{U}_b$ and $v' := Cu$ the, by assumption, unique exact and discrete solutions, $\hat{u}$ and $\hat{u}^h$, for the equations $B\hat{u} = v'$ and $B^h\hat{u}^h = Q'^h v'$. Here B^h and Q'^h are continuous in $\mathcal{U}_b$ and $\Phi^h(B \cdot - v') = B^h \cdot - Q'^h v'$, see Chapter 6. We introduce the notations $T := B^{-1}$ and $T^h := (B^h)^{-1} Q'^h \in \mathcal{L}(\mathcal{V}_b', \mathcal{U}_b^h)$. Mind that

$$\begin{aligned} A,\ B,\ C\ &: \mathcal{U}_b \to \mathcal{V}_b',\ A^h,\ B^h,\ C^h\ : \mathcal{U}_b^h \to \mathcal{V}_b'^h \not\subset \mathcal{V}_b', \\ T &: \mathcal{V}_b' \to \mathcal{U}_b,\ T^h : \mathcal{V}_b' \to \mathcal{U}_b^h,\ Q'^h : \mathcal{V}_b' \to \mathcal{V}_b'^h. \end{aligned}$$

Furthermore, we introduce a linear bounded extension $Q_e'^h$ and the corresponding T_e^h as

$$\begin{aligned} Q_e'^h : \mathcal{V}_b' \cup \mathcal{V}_b'^h \to \mathcal{V}_b'^h,\ \text{s.t}\ Q_e'^h|_{\mathcal{V}_b'} = Q'^h,\ \ Q_e'^h|_{\mathcal{V}_b'^h} = id|_{\mathcal{V}_b'^h} \text{ and } \quad (7.11) \\ T_e^h := (B^h)^{-1} Q_e'^h \in \mathcal{L}(\mathcal{V}_b' \cup \mathcal{V}_b'^h, \mathcal{U}_b^h),\ T_e^h|_{\mathcal{V}_b'} = T^h. \end{aligned}$$

Theorem 4.4.4 implies that $||(T - T^h)Cu||_{\mathcal{U}} \to 0$ for $h \to 0$ and any $u \in \mathcal{U}_b$. C is compact, so this implies

$$||(T - T^h)C||_{\mathcal{V}_b' \leftarrow \mathcal{U}_b} \to 0 \quad \text{for } h \to 0. \tag{7.12}$$

We have to realize that $T^h \circ C^h$ is not, but $T_e^h \circ C^h : \mathcal{U}_b^h \to \mathcal{U}_b^h$ is well defined. We use Remark 6.1.3 throughout the proof, usually without mentioning it, compare the proof of Theorem 6.1.1. By (2.79) we have

$$\begin{aligned} \|I^h E^h u^h - u^h\|_{\mathcal{W}_q^1(\Omega)}^h\ ,\ \|E^h u^h - u^h\|_{\mathcal{W}_q^1(\Omega)}^h \le Ch^{(n-1)/q} \|u^h\|_{\mathcal{W}_q^1(\Omega)}^h \text{ and} \\ \|E^h u^h\|_{\mathcal{W}_\infty^1(\Omega)}^h = C\|u^h\|_{\mathcal{W}_q^1(\Omega)}^h \ \forall\ u^h \in \mathcal{U}^h. \end{aligned} \tag{7.13}$$

We combine $u^h \in \mathcal{U}_b^h$ with E^h. With the boundedly invertible A we estimate

$$\begin{aligned} \|u^h\|_{\mathcal{U}}^h &\le 2||E^h u^h||_{\mathcal{U}} \le 2||A^{-1}||_{\mathcal{U}_b \leftarrow \mathcal{V}_b'} ||AE^h u^h||_{\mathcal{V}'} \\ &= 2||A^{-1}||_{\mathcal{U}_b \leftarrow \mathcal{V}_b'} ||B(I + TC)E^h u^h||_{\mathcal{V}'} \\ &\le 2||A^{-1}||_{\mathcal{U}_b \leftarrow \mathcal{V}_b'} ||B||_{\mathcal{V}_b' \leftarrow \mathcal{U}_b} ||(I + TC)E^h u^h||_{\mathcal{U}}, \end{aligned} \tag{7.14}$$

hence

$$||(I + TC)E^h u^h||_{\mathcal{U}} \ge (||A^{-1}||_{\mathcal{U}_b \leftarrow \mathcal{V}_b'} ||B||_{\mathcal{V}_b' \leftarrow \mathcal{U}_b})^{-1} \|u^h\|_{\mathcal{U}}^h / 2. \tag{7.15}$$

We apply the stability of B^h to $w^h := B^h u^h$ to find $\|u^h\|_{\mathcal{U}}^h \le \|(B^h)^{-1}\|_{\mathcal{V}_b'^h \leftarrow \mathcal{U}_b^h} \cdot \|w^h\|_{\mathcal{V}}$. Furthermore

$$\begin{aligned}\|(I+T_e^hC^h)u^h\|_{\mathcal{U}}^h &= \|(B^h)^{-1}B^h(I+T_e^hC^h)u^h\|_{\mathcal{U}}^h \\ &\le \|(B^h)^{-1}\|_{\mathcal{V}_b'^h\leftarrow\mathcal{U}_b^h}\|B^h(I+T_e^hC^h)u^h\|_{\mathcal{V}'}^h\end{aligned}$$

implies

$$\|B^h(I+T_e^hC^h)u^h\|_{\mathcal{V}'}^h \ge \|(I+T_e^hC^h)u^h\|_{\mathcal{U}}^h \;/\; \|(B^h)^{-1}\|_{\mathcal{V}_b'^h\leftarrow\mathcal{U}_b^h}. \quad (7.16)$$

Now, we combine the consistency of C^h to C, the stability of T^h and (7.13) to estimate

$$\begin{aligned}\|T_e^h(C^h-CE^h)u^h\|_{\mathcal{V}^{h'}}^h &\le \|T_e^hC^h(u^h-I^hE^hu^h)\|_{\mathcal{V}^{h'}}^h \\ &\quad +\|T_e^h(C^hI^hE^hu^h-CE^hu^h)\|_{\mathcal{V}^{h'}}^h \\ &\le \|(B^h)^{-1}\|_{\mathcal{V}_b'^h\leftarrow\mathcal{U}_b^h}\big(h^{(n-1)/2}\|u^h\|_{\mathcal{U}}^h \\ &\quad +\|T_e^h(C^hI^hE^hu^h-CE^hu^h)\|_{\mathcal{V}^{h'}}^h\big) \qquad (7.17) \\ &\le \|(B^h)^{-1}\|_{\mathcal{V}_b'^h\leftarrow\mathcal{U}_b^h}\big(h^{(n-1)/2} \\ &\quad + \text{ consistency error of } C^h \text{ w.r.t.} E^hu^h\big)\|u^h\|_{\mathcal{U}}^h\end{aligned}$$

Now, we use

$$\begin{aligned}\Phi^h(A) &= A^h = \Phi^h(B+C) = B^h + C^h \\ &= B^h(I+(B^h)^{-1}C^h) = B^h(I+T_e^hC^h) : \mathcal{U}_b^h \to \mathcal{V}_b^{h'},\end{aligned}$$

see Notation 7.1, the identity $T_e^h|_{\mathcal{V}_b^{h'}} = (B^h)^{-1}Q_e'^h|_{\mathcal{V}^{h'}_b} = (B^h)^{-1}$ and the consistency of C^h to C to estimate

$$\begin{aligned}||A^hu^h||_{\mathcal{V}^{h'}}^h &= ||B^h(I+(B^h)^{-1}Q_e'^hC^h)u^h||_{\mathcal{V}^{h'}}^h \text{ by (7.11)} \\ &= ||B^h(I+T_e^hC^h)u^h||_{\mathcal{V}^{h'}}^h \text{ by (7.16),(7.11)} \\ &\ge \Big(||(I+T_e^hCE^h)u^h||_{\mathcal{U}}^h \\ &\quad -\|T_e^h(C^h-CE^h)u^h\|_{\mathcal{U}}^h\Big) \;/\; ||(B^h)^{-1}||_{\mathcal{U}_b^h\leftarrow\mathcal{V}_b^{h'}}^h \\ &\ge \Big(||(I+TC)E^hu^h||_{\mathcal{U}}^h - ||(T-T_e^h)CE^hu^h||_{\mathcal{U}}^h \\ &\quad -\|T_e^h(C^h-CE^h)u^h\|_{\mathcal{U}}^h\Big)/||B^{h\,-1}||_{\mathcal{U}_b^h\leftarrow\mathcal{V}_b^{h'}}^h \\ &\quad \text{by (7.15),(7.12),(7.11)} \\ &\ge \Big(||u^h||_{\mathcal{U}}^h/\big(2||A^{-1}||_{\mathcal{U}_b\leftarrow\mathcal{V}_b'}||B||_{\mathcal{V}_b'\leftarrow\mathcal{U}_b}\big) \\ &\quad -||(T-T^h)C||_{\mathcal{V}\leftarrow\mathcal{U}_b}||E^hu^h||_{\mathcal{U}} \\ &\quad -||(I-E^h)u^h||_{\mathcal{U}} - \|T_e^h(C^h-CE^h)u^h\|_{\mathcal{U}}^h\Big)/||B^{h\,-1}||_{\mathcal{U}_b^h\leftarrow\mathcal{V}^{h'}}^h \\ &\ge K'||u^h||_{\mathcal{U}}^h(1-\mathcal{O}(1)) \text{ by(7.13),(7.12),(7.17)}\end{aligned}$$

Because of (7.12) and the stability of $(B^h)_{h\in H}$ there exists a positive constant K, independent of h, such that for all $h \le h_0$ the following holds:

$$||A^h u^h||^h_{\mathcal{V}_b'} \ge K||u^h||^h_{\mathcal{U}} \quad \text{for all } u^h \in \mathcal{U}_b^h.$$

A similar estimation is valid for the dual $(A^h)^d$ as well. From $\dim \mathcal{U}_b^h = \dim \mathcal{V}_b^{h'} < \infty$ and Criterion 7.1, now with a uniform bound, we see that A^h is invertible for $h \le h_0$. Moreover, we obtain $||(A^h)^{-1}||^h_{\mathcal{U}_b^h \leftarrow \mathcal{V}^{h'}} \le 1/K$, i.e. $(A^h)_{h\in H}$ is stable. □

Theorem 7.2.3 yields a Criterion for the stability of discretizations of operators, which are compact perturbations of coercive operators. An important example for this kind of operators are $A \in \mathcal{L}(\mathcal{U}_b, \mathcal{U}_b')$ satisfying a so-called *Gårding inequality.*

Theorem 7.2.5. *Let the Banach space $\mathcal{U}_b$ be continuously, densely and compactly embedded into the Hilbert space $\mathcal{W}$. Assume that $A \in \mathcal{L}(\mathcal{U}_b, \mathcal{U}_b')$ fulfills a Gårding inequality, i.e. there exist constants $M > 0$ and m such that*

$$< Au, u >_{\mathcal{U}_b' \times \mathcal{U}_b} \ge M||u||^2_{\mathcal{U}} - m||u||^2_{\mathcal{W}} \quad \forall \quad u \in \mathcal{U}_b. \tag{7.18}$$

Then A is a compact perturbation of a coercive operator. Now, Theorem 7.2.3 is applicable and yields the stability of A^h, $\tilde{A}^h$.

Proof Identifying $\mathcal{W}$ with $\mathcal{W}'$, the above assumptions yield $\mathcal{U}_b \subset \mathcal{W} \subset \mathcal{U}_b'$. Additionally the embedding $I_{\mathcal{U}_b \to \mathcal{U}_b'}$ is compact, see Zeidler [68]. We obtain the splitting

$$A = (A + mI_{\mathcal{U}_b \to \mathcal{U}_b'}) - mI_{\mathcal{U}_b \to \mathcal{U}_b'}.$$

Moreover, $A + mI_{\mathcal{U}_b \to \mathcal{U}_b'}$ is coercive because of $||u||^2_{\mathcal{W}} =< I_{\mathcal{U}_b \to \mathcal{U}_b'} u, u >_{\mathcal{U}' \times \mathcal{U}}$ and (7.18). □

Finally, we want to approach the problem of inexact evaluation of the operators $\tilde{A}^h, \tilde{Q}'^h$. In many cases, these operators can be evaluated only approximatively, e.g., by numerical integration, inexact evaluation by divided differences, aliasing and de-aliasing, or non-exact solution of the systems by iteration methods. The influence of the first type of perturbations is analyzed in the following Theorem, an immediate generalization of the well-known first Lemma of Strang, see Section 6.5, to our situation. The iteration errors can be estimated in a similar way if they are chosen to become small, similarly to $||A^h - \tilde{A}^h||_{\mathcal{V}_b^{h'} \leftarrow \mathcal{U}_b^h} \to 0$ below.

Theorem 7.2.6. *Let $\mathcal{U}_b^h, \mathcal{V}_b^h$ define a bi-dual admissible scheme. Assume $A^h \in \mathcal{L}(\mathcal{U}_b^h, \mathcal{V}_b^{h'})$ and $(f^h) \in (\mathcal{V}_b^{h'})$ are "properly" perturbed by $\tilde{A}^h$ and $\tilde{f}^h$, respectively, hence,*

$$||A^h - \tilde{A}^h||_{\mathcal{V}_b^{h'} \leftarrow \mathcal{U}_b^h} \to 0 \text{ and } ||f^h - \tilde{f}^h||^h_{\mathcal{V}'_b} \to 0 \text{ for } h \to 0.$$

Then, for small enough h,

$$A^h \text{ is stable } \Leftrightarrow \tilde{A}^h \text{ is stable.}$$

Furthermore, if $\tilde{A}^h$ is stable and $\tilde{A}^h \cdot - \tilde{f}^h$ is consistent with $A \cdot - f$, then the uniquely determined solution $\tilde{u}_0^h$ of the perturbed equation

$$\tilde{A}^h \tilde{u}_0^h = \tilde{f}^h, \quad \tilde{u}_0^h \in \mathcal{U}_b^h$$

converges to the uniquely determined exact solution u_0 of $Au_0 = f$, more precisely, for sufficiently small h, we have the error estimates as in Theorem 6.5.3, see (5.16),

$$\|u_0 - u_0^h\|^h \leq C \left(\inf_{u^h \in \mathcal{U}_b^h} \|u_0 - u^h\|^h + ||A_h u_0 - \tilde{A}_h u_0||_{\mathcal{V}'_b}^h + ||f - \tilde{f}^h||_{\mathcal{V}'_b}^h \right).$$

We have assumed the usual situation, that u_0^h is smooth enough to permit $A^h u_0$.

Proof The equivalence of the stability of $\tilde{A}^h$ and A^h is a direct consequence of the well-known Theorem of Neumann. With $||A^h - \tilde{A}^h||\,||(\tilde{A}^h)^{-1}|| < 1$ this yields the estimate

$$||(A^h)^{-1}|| \leq \frac{||(\tilde{A}^h)^{-1}||}{1 - ||A^h - \tilde{A}^h||\,||(\tilde{A}^h)^{-1}||}.$$

Finally, we obtain the desired error estimate by Theorem 6.5.3. □

Remark 7.2.7. We combine the stability results of this Chapter and consistency results of the last Chapter 6. Then we obtain the final convergence results for all cases of conforming and non conforming FE and spectral methods treated in this Booklet. The results concerning stability and convergence of bordered systems, including Navier-Stokes equations directly carry over from [16]. In fact, the bordered systems in bifurcation numerics represent a compact perturbation of A. For Navier-Stokes problems we distinguish two cases of a moderate coefficient of kinematic viscosity (and hence only moderately interesting phenomena) or very small coefficient (causing all the turbulence problems). For the first case, compact perturbations allow to prove convergence directly by compact perturbation results from the Stokes operator. In both cases, whenever FEMs with or without variational crimes have been shown to yield convergent methods for the Stokes or the Navier-Stokes problems, the corresponding bordered systems again represent compact perturbations of the linearized Navier-Stokes operator. In all these cases, bifurcation numerics based on bordered systems for variational crimes in FEs and collocation and De-aliasing in spectral methods, yield converging bifurcation scenarios, see [9, 15, 16, 6, 7].

7.3 Convergence for General FE and Spectral Methods for Linear and Nonlinear Problems

After proving stability for a large class of operators and their discretization, we finally formulate the convergence results. They are obtained by combining this stability with the consistency estimates in Chapter 6. We do not repeat all the formulas for the conditions imposed in the Theorems in Chapter 6, but rather refer to the old numbers. The smoothness requirements for the solutions, u_0, are different for the different approaches and are documented by the index of the norms $\|u_0\|$ listed below. In Theorem 7.2.3 we have required: For $a, B \in \mathcal{L}(\mathcal{U}_b, \mathcal{V}_b')$, let the discrete B^h (or $\tilde{B}^h$) be stable and let $A = B + C \in \mathcal{L}(\mathcal{U}_b, \mathcal{V}_b')$ with a compact perturbation C of A. Let B^h and C^h be consistent with B and C, and $A^{-1} \in \mathcal{L}(\mathcal{V}_b', \mathcal{U}_b)$, hence, is boundedly invertible. For Theorem 7.2.6 we assume $A^h \in \mathcal{L}(\mathcal{U}_b^h, \mathcal{V}_b^{h\prime})$ and $(f^h) \in (\mathcal{V}_b^{h\prime})$ to be "properly" perturbed by $\tilde{A}^h$ and $\tilde{f}^h$, respectively, hence,

$$||A^h - \tilde{A}^h||_{\mathcal{V}_b^{h\prime} \leftarrow \mathcal{U}_b^h} \to 0 \text{ and } ||f^h - \tilde{f}^h||^h_{\mathcal{V}'_b} \to 0 \text{ for } h \to 0. \tag{7.19}$$

These conditions are satisfied, as we have seen, for FEMs and spectral methods applied to elliptic differential operators as 3.2 and for Navier-Stokes operators, whenever the $A^{-1} \in \mathcal{L}(\mathcal{V}_b', \mathcal{U}_b)$.

Theorem 7.3.1. *Let $\mathcal{U}_b^h, \mathcal{V}_b^h$ define a FE or a spectral method with or without variational crimes. Let A be an elliptic differential operator as in Section 3.2 or a Navier-Stokes operators, with appropriate discretization, with $A^{-1} \in \mathcal{L}(\mathcal{V}_b', \mathcal{U}_b)$, hence, A is boundedly invertible. For a quadrature approximation or for collocation methods let (7.19) be satisfied. We do not repeat all the detailed conditions, but refer to the corresponding Theorems. We find for*

FEMs without variational crimes and continuous FEMs with natural boundary conditions, *see Theorems 3.3.4 and 6.2.1:*

$$\|u_0^h - u_0\|^h_{H^1(\Omega)} \le C\|I^h u - u\|_{H^1(\Omega)} \le Ch^{m-1}\|u_0\|_{H^m(\Omega)}. \tag{7.20}$$

FEMs with violated Dirichlet boundary conditions, *see Theorem 6.2.5:*

$$\begin{aligned} \|u_0^h - u_0\|^h_{H^1(\Omega)} &\le C\|I^h u - u\|_{H^1(\Omega)} + C_\rho h^\mu \|u\|_{W_\infty^{2m'+1-\rho}(\Omega)} \text{ with} \\ &\le Ch^{\min\{\mu, m-2\}} \|u_0\|_{W_\infty^{\max\{m, 2m'+1-\rho\}}(\Omega)} \\ \mu &:= 2m' - \rho - \min\{2m' - \rho, m + \tau\} + 1/2 \\ \text{and } \rho &= 0, 1, 2 \text{ for Gauss-, Gauss-Radau, Gauss-Lobatto points} \end{aligned} \tag{7.21}$$

FEMs with violated continuity conditions, *see Theorem 6.3.1:*

$$\begin{aligned} \|u_0^h - u_0\|^h_{H^1(\Omega)} &\le C\|I^h u - u\|_{H^1(\Omega)} + C_\rho h^{\mu'} \|u\|_{W_\infty^{\mu'+2}(\Omega)} \\ &\le Ch^{\min\{\mu', m-2\}} \|u_0\|_{H^{\max\{m, \mu'+2\}}(\Omega)}. \\ \text{with } \mu' &= 2m' - 1 - \rho - (m + \tau). \end{aligned}$$

FEMs in the isoparametric case, *see Theorem 6.4.2:*

$$\|u_0^h - u_0\|_{H^1(\Omega)}^h \le Ch^{m-1}\left(\|u\|_{H^m(\Omega)} + \|u_0\|_{W_\infty^1(\Omega)}\right). \tag{7.22}$$

FEMs with quadrature approximations methods, *see Theorem 6.5.6:*

$$\begin{aligned}
&\|u_0 - u_0^h\|_{H^1(\Omega)} \qquad (7.23)\\
&\quad \le Ch^{m-1-n/2}\max_{i,j=0}^n \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|u\|_{W_\infty^m(\Omega)},\ \textit{or}\\
&\quad \le C\big(h^{\ell+1-n/2-\min\{m+\tau,k+1\}}\max_{i,j=0^n} \|a_{ij}\|_{W_\infty^k(\Omega)} \cdot \|u_0\|_{W_\infty^{k+1}(\Omega)}\\
&\qquad +h^{\ell+1-n/2-\min\{m+\tau,k\}}\|f\|_{W_\infty^k(\Omega)}\big),\ \textit{for the weak and}\\
&\quad \le Ch^{m-2-n/2}\max_{i,j=0}^n \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|u\|_{W_\infty^m(\Omega)}^h,\ \textit{or}\\
&\quad \le C\big(h^{\ell-n/2-\min\{m+\tau,k\}}\max_{i,j=0}^n \|a_{ij}\|_{W_\infty^k(\Omega)} \cdot \|u_0\|_{W_\infty^{k+2}(\Omega)}^h\\
&\qquad +h^{\ell-n/2-\min\{m+\tau,k\}}\|f\|_{W_\infty^k(\Omega)}\big)\ \textit{for the strong forms.}
\end{aligned}$$

Collocation methods on non degenerate subdivisions, *see Theorem 6.6.2:*

$$\begin{aligned}
&\|u_0 - u_0^h\|_{H^2(\Omega)} \qquad (7.24)\\
&\quad \le Ch^{m-2-n/2}\max_{i,j=0}^n \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|u\|_{W_\infty^m(\Omega)}^h,\ \textit{or}\\
&\quad \le C\big(h^{m-n/2-\min\{m+\tau,m\}}\max_{i,j=0}^n \|a_{ij}\|_{W_\infty^m(\Omega)} \cdot \|u_0\|_{W_\infty^{m+2}(\Omega)}^h\\
&\qquad +h^{m-n/2-\min\{m+\tau,m\}}\|f\|_{W_\infty^m(\Omega)}\big) \qquad (7.25)
\end{aligned}$$

Spectral methods including quadrature approximations methods, *see Theorem 6.6.4:*

$$\|u_0^h - u_0\|_{H^1(\Omega)}^h \le CN^{-m+\iota_K(2)+\mu_K}\max_{i,j=0}^n \|a_{ij}\|_{L_\infty(\Omega)} \cdot \|u\|_{H_w^m(\Omega)}\ \textit{with} \tag{7.26}$$

$\ell = 2,\ \iota_F(\ell) = \ell,\ \iota_C(\ell) = 2\ell,\ \iota_L(\ell) = 2\ell + n/2,\ \mu_F = 2,\ \mu_C = \mu_L = 4$. *This shows the extremely good (exponential) convergence for large m and spectral methods.*

Nonlinear problems: *All the above discretization methods for nonlinear problems: Under the conditions of Theorem 6.7.1 the combination of the estimate in (6.133),*

$$\|u_0 - u_0^h\|_{H^1(\Omega)}^h \le C\left\{\inf_{u^h \in \mathcal{U}_b^h} \|u_0 - h^h\|_{H^1(\Omega)}^h + \quad \textit{consistency error}\right\} \tag{7.27}$$

with the above estimates yields the convergence for nonlinear problems in one of the above realizations.

8. Petrov-Galerkin Methods for Bordered Systems

To numerically compute bifurcation scenarios, extended and, in particular, bordered systems have been introduced by Keller and used by many authors, see Chapter 4.2. In the mean time, the concept of *bordered systems,* obtained by few new parameters and equations, see (8.7), is the method of choice. We combine this case and use our generalized Petrov-Galerkin methods to solve these types of linear operator equations. Again we can, for stability arguments, restrict the discussion to linear problems, see Chapter 4.2. We interpret this combination of bordering with general Petrov-Galerkin discretization methods as compact perturbation of an invertible operator B with stable $(B^h)_{h\in H}$. This will yield the desired stability results for bordered systems.

8.1 Petrov-Galerkin Methods for Bordered Systems

Now suppose, we have the following splitting of $\mathcal{U}_b, \mathcal{V}_b$ (and $\mathcal{U}_b', \mathcal{V}_b'$, we indicate necessary dual assumptions this way), see Jepson/Spence [40],

$$\begin{aligned} \mathcal{U}_b = \mathcal{N}_{\mathcal{U}_b} \oplus \mathcal{M}_{\mathcal{U}_b}, \mathcal{V}_b = \mathcal{N}_{\mathcal{V}_b} \oplus \mathcal{M}_{\mathcal{V}_b} \\ (\text{ and } \mathcal{U}_b' = \mathcal{N}_{\mathcal{U}_b'} \oplus \mathcal{M}_{\mathcal{U}_b'}, \mathcal{V}_b' = \mathcal{N}_{\mathcal{V}_b'} \oplus \mathcal{M}_{\mathcal{V}_b'}) \end{aligned} \tag{8.1}$$

with m-dimensional subspaces $\mathcal{N}_{\mathcal{U}_b}, \mathcal{N}_{\mathcal{V}_b}$ (and $\mathcal{N}_{\mathcal{U}_b'}, \mathcal{N}_{\mathcal{V}_b'}$) and (closed) complements $\mathcal{M}_{\mathcal{U}_b}, \mathcal{M}_{\mathcal{V}_b}$ (and $\mathcal{M}_{\mathcal{U}_b'}, \mathcal{M}_{\mathcal{V}_b'}$). Let

$$Q \in \mathcal{L}(\mathcal{U}_b, \mathcal{N}_{\mathcal{U}_b}), \hat{Q} \in \mathcal{L}(\mathcal{V}_b, \mathcal{N}_{\mathcal{V}_b}) \text{ (and } Q' \in \mathcal{L}(\mathcal{U}_b', \mathcal{N}_{\mathcal{U}_b}'), \hat{Q}' \in \mathcal{L}(\mathcal{V}_b', \mathcal{N}_{\mathcal{V}_b}') \text{)}$$

and the complementary $I - Q, I - \hat{Q}$ (and $I - Q'$, $I - \hat{Q}'$) be the bounded projections which are induced by the above splittings. For the practical computations, we assume the m-dimensional dual subspaces $\mathcal{N}_{\mathcal{U}_b'} \subset \mathcal{U}_b'$,$\mathcal{N}_{\mathcal{V}_b}' \subset \mathcal{V}_b'$ to be chosen such that

$$\mathcal{M}_{\mathcal{U}_b} = (\mathcal{N}_{\mathcal{U}_b'})^\perp, \mathcal{M}_{\mathcal{V}_b} = (\mathcal{N}_{\mathcal{V}_b}')^\perp \text{(and } \mathcal{M}_{\mathcal{U}_b}' = (\mathcal{N}_{\mathcal{U}_b})^\perp, \mathcal{M}_{\mathcal{V}_b'} = (\mathcal{N}_{\mathcal{V}_b})^\perp)$$

w.r.t. $< \cdot, \cdot >$, and the bi-orthogonal bases satisfy

$$\mathcal{N}_{\mathcal{U}_b} = [\phi_1, \ldots, \phi_m] \subset \mathcal{U}_b, \quad \mathcal{N}_{\mathcal{U}_b'} = [\phi_1', \ldots, \phi_m'] \subset \mathcal{U}_b' \text{ and } (8.2)$$
$$\mathcal{N}_{\mathcal{V}_b} = [\psi_1, \ldots, \psi_m] \subset \mathcal{V}_b, \quad \mathcal{N}_{\mathcal{V}_b'} = [\psi_1', \ldots, \psi_m'] \subset \mathcal{V}_b' \text{ with}$$
$$< \phi_i, \phi_j' >_{\mathcal{U}_b \times \mathcal{U}_b'} = \delta_{i,j} \text{ and } < \psi_i, \psi_j' >_{\mathcal{V}_b \times \mathcal{V}_b'} = \delta_{i,j}, i, j = 1, \ldots, m.$$

Then the above $Q, \hat{Q}$ are defined as

$$Qu := \sum_{i=1}^{m} < u, \phi_i' >_{\mathcal{U}_b \times \mathcal{U}_b'} \phi_i \in \mathcal{N}_{\mathcal{U}_b} \text{ for } u \in \mathcal{U}_b \text{ and}$$
$$\hat{Q}f := \sum_{i=1}^{m} < f, \psi_i' >_{\mathcal{V}_b \times \mathcal{V}_b'} \psi_i \in \mathcal{N}_{\mathcal{V}_b} \text{ for } f \in \mathcal{V}_b. \quad (8.3)$$

We project the operator equation (4.113) to obtain

$$(I - \hat{Q})Au = (I - \hat{Q})f, \quad u \in \mathcal{M}_{\mathcal{U}_b} \quad (8.4)$$

and the complementary equation

$$(\hat{Q})Au = (\hat{Q})f.$$

This splitting is important for discretizing generalized inverses, in particular for the Liapunov-Schmidt methods in nonlinear problems. (8.4) has an unique solution for every $f \in \mathcal{V}_b$, if and only if

$$((I - \hat{Q})A|_{\mathcal{M}_{\mathcal{U}_b}})^{-1} \in \mathcal{L}(\mathcal{M}_{\mathcal{V}_b}, \mathcal{M}_{\mathcal{U}_b}). \quad (8.5)$$

For a Fredholm operator A with index 0 we have, see [40]

$$(8.5) \Leftrightarrow \mathcal{N}(Q) \cap \mathcal{N}((I - \hat{Q})A) = \{0\}. \quad (8.6)$$

To treat equation (8.4) with the methods from the preceeding Chapters we transform (8.4). We define $L \in \mathcal{L}(\mathcal{U}_b \times \mathbb{R}^m, \mathcal{V}_b' \times \mathbb{R}^m)$, its application to $(u, \alpha)^T$, and (8.4) as in Linear Algebra, as

$$L = \begin{pmatrix} A & \psi_1, \ldots, \psi_m \\ < \cdot, \phi_1' >_{\mathcal{U}_b \times \mathcal{U}_b'} & 0, \ldots, 0 \\ \vdots & \vdots \quad \vdots \\ < \cdot, \phi_m' >_{\mathcal{U}_b \times \mathcal{U}_b'} & 0, \ldots, 0 \end{pmatrix}, \; L \begin{pmatrix} u \\ \alpha \end{pmatrix} := \begin{pmatrix} Au + \sum_{i=1}^{m} \alpha_i \psi_i \\ < u, \phi_1' >_{\mathcal{U}_b \times \mathcal{U}_b'} \\ \vdots \\ < u, \phi_m' >_{\mathcal{U}_b \times \mathcal{U}_b'} \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}, \quad (8.7)$$

which is equivalent to

$$(I - \hat{Q})Au = (I - \hat{Q})f, \quad (8.8)$$
$$\sum_{i=1}^{m} \alpha_i \psi_i = \hat{Q}(f - Au) = \sum_{i=1}^{m} < \psi_i (f - Au) >_{\mathcal{V}_b \times \mathcal{V}_b'} \psi_i.$$

The solution $u \in \mathcal{U}_b$ of (8.8) automatically is $u \in \mathcal{M}_{\mathcal{U}_b}$ since $< u, \phi_1' >_{\mathcal{U}_b \times \mathcal{U}_b'} = \cdots = < u, \phi_m' >_{\mathcal{U}_b \times \mathcal{U}_b'} = 0$. Hence, we have derived the connection between

(8.4) and (8.7). Now we can treat equation (8.7) with the methods from Chapter 7. We obtain an analogous result to Theorems 7.2.1 and 7.2.3. This is the main Theorem of this Chapter and the goal of the Booklet:

Theorem 8.1.1. *Let* $(\mathcal{U}_b^h,\ \mathcal{V}_b^h)_{h\in H}$ *be approximating spaces with* $\dim\mathcal{U}_b^h = \dim\mathcal{V}_b^h$ *and let* $Q_{ext}^{'h} := (Q^{'h}, I_{\mathbb{R}^m})$. *Furthermore, let* $A = B + C$ *with* $A, B, C \in \mathcal{L}(\mathcal{U}_b, \mathcal{V}_b')$, C *be compact and* B *be boundedly invertible with stable* $Q^{'h}B|_{\mathcal{U}_b^h}$, *see Theorem7.2.3. Then the following conditions 1. - 3. are mutually equivalent and each implies 4:*

$$\begin{aligned} &1.\ ((I-\hat{Q})A|_{\mathcal{M}_{\mathcal{U}_b}})^{-1} \in \mathcal{L}(\mathcal{M}_{\mathcal{V}_b}, \mathcal{M}_{\mathcal{U}_b}),\\ &2.\ \forall\ \ f \in \mathcal{V}_b (8.4), (8.8) \textit{are uniquely solvable,}\\ &3.\ L^{-1} \in \mathcal{L}(\mathcal{V}_b \times \mathbb{R}^m, \mathcal{U}_b \times \mathbb{R}^m),\\ &4.\ (Q_{ext}^{'h} L|_{\mathcal{U}_b^h \times \mathbb{R}^m})_{h\in H} \textit{ is stable.} \end{aligned}$$

Here, the first condition states the existence of the generalized inverse, see (8.5). Under the conditions of Theorem 7.2.1 the preceding 1. - 4. are equivalent.

Proof: *1.* $\Leftrightarrow$ *2.* $\Leftrightarrow$ *3.* follows immediately from the above discussion concerning the unique solvability of (8.4),(8.8) and (8.7) .
Since $dist(\binom{u}{\alpha}, \mathcal{U}_b^h \times \mathbb{R}^m) = dist(u, \mathcal{U}_b^h)$ and $dist(\binom{v}{\alpha}, \mathcal{V}_b^h \times \mathbb{R}^m) = dist(v, \mathcal{V}_b^h)$ for all $\binom{u}{\alpha} \in \mathcal{U}_b \times \mathbb{R}^m$ and $\binom{v}{\alpha} \in \mathcal{V}_b \times \mathbb{R}^m$, respectively., the $(\mathcal{U}_b^h \times \mathbb{R}^m,\ \mathcal{V}_b^h \times \mathbb{R}^m)_{h\in H}$ are approximating spaces with $\dim(\mathcal{U}_b^h \times \mathbb{R}^m) = \dim(\mathcal{V}_b^h \times \mathbb{R}^m)$. The projectors $P^h, Q^{'h}$ in (4.118) are extended to $\mathcal{U}_b \times \mathbb{R}^m$ and $\mathcal{V}_b \times \mathbb{R}^m$ by, e.g., $Q_{ext}^{'h}(v,\alpha)^T := (Q^{'h}v, \alpha)^T$ implying $\|Q_{ext}^{'h}(v,\alpha)^T - (v,\alpha)^T\|_{\mathcal{V}_b\times\mathbb{R}^m} = \|Q^{'h}v - v\|_{\mathcal{V}}$ a.s.o.
In order to prove the implication *3.* $\Rightarrow$ *4.* for L, we have to check the conditions in Theorem 7.2.3, i.e. we have to show that $L \in \mathcal{L}(\mathcal{U}_b \times \mathbb{R}^m, \mathcal{V}_b \times \mathbb{R}^m)$ is a compact perturbation of an operator $B_{ext} \in \mathcal{L}(\mathcal{U}_b \times \mathbb{R}^m, \mathcal{V}_b \times \mathbb{R}^m)$ with stable discretization $Q_{ext}^{'h} B_{ext}|_{\mathcal{U}_b^h \times \mathbb{R}^m}$. We define the operators $\Phi \in \mathcal{L}(\mathbb{R}^m, \mathcal{U}_b')$, $\Phi^d \in \mathcal{L}(\mathcal{U}_b^{''}, \mathbb{R}^m)$, and $\Psi \in \mathcal{L}(\mathbb{R}^m, \mathcal{V}_b)$ by

$$\Phi\alpha := \textstyle\sum_{i=1}^m \alpha_i \phi_i',\ \text{and}\ \Psi\alpha := \sum_{i=1}^m \alpha_i \psi_i',\ \text{with}$$
$$\Phi^d \in \mathcal{L}(\mathcal{U}_b^{''}, \mathbb{R}^m),\ \text{and}\ \Phi^d u := (< u, \phi_i' >_{\mathcal{U}_b^{''}\times\mathcal{U}_b'})_{i=1}^m \ \text{for}\ \ u \in \mathcal{U}_b\ \subset \mathcal{U}_b^{''}$$

and obtain $L = \binom{A\ \ \Psi}{\Phi^d\ \ 0}$. We write L in the form

$$L = \begin{pmatrix} B & 0 \\ 0 & I_{\mathbb{R}^m} \end{pmatrix} + \begin{pmatrix} C & \Psi \\ \Phi^d & -I_{\mathbb{R}^m} \end{pmatrix} =: B_{ext} + C_{ext}.$$

The bounded invertability of B and the stability of $Q^{'h}B|_{\mathcal{U}_b^h}$ (for $h \le h_0$), respectively, imply immediately that B_{ext} and $Q_{ext}^{'h} B_{ext}|_{\mathcal{U}_b^h \times \mathbb{R}^m}$ for $h \le h_0$, respectively., have an (equi-) bounded inverse: In fact,

$$||(Q_{ext}^{'h} B_{ext}|_{\mathcal{U}_b^h \times \mathbb{R}^m})^{-1}||_{\mathcal{U}_b^h \times \mathbb{R}^m \leftarrow \mathcal{V}_b^h \times \mathbb{R}^m} \leq ||(Q^{'h} B|_{\mathcal{U}_b^h})^{-1}||_{\mathcal{U}_b^h \leftarrow \mathcal{V}_b^h} + 1$$

and the stability of $(Q^{'h} B|_{\mathcal{U}_b^h})_{h \in H}$ shows that $(Q_{ext}^{'h} B_{ext}|_{\mathcal{U}_b^h \times \mathbb{R}^m})^{-1}$ is uniformly bounded (with respect to h), and thus stable.
As operators with finite-dimensional domains, the Φ, Ψ and $I_{\mathbb{R}^m}$ are compact. Because $C \in C(\mathcal{U}_b, \mathcal{V}_b)$ the same holds for C_{ext} . Thus, we have checked all assumptions of Theorem 7.2.3 and this new Theorem is proved.
We may directly employ numerical schemes to yield approximations $\mathcal{N}_{\mathcal{U}_b}^h \subset \mathcal{U}_b^h, \mathcal{N}_{\mathcal{V}_b}^h \subset \mathcal{V}_b^h$ (and $\mathcal{N}_{\mathcal{U}_b'}^h \subset \mathcal{U}_b^{h'}, \mathcal{N}_{\mathcal{V}_b'}^h \subset \mathcal{V}_b^{h'}$) in the following sense: There exist bi-orthogonal systems $\phi_i^h, \phi_i^{'h}, \psi_i^h, \psi_i^{h'}$, $i = 1, \ldots, m$, see (8.2) with

$$\mathcal{N}_{\mathcal{U}_b}^h = [\phi_1^h, \ldots, \phi_m^h] \subset \mathcal{U}_b^h, \ \mathcal{N}_{\mathcal{U}_b'}^h = [\phi_1^{h'}, \ldots, \phi_m^{h'}] \subset \mathcal{U}_b^{h'},$$
$$\mathcal{N}_{\mathcal{V}_b}^h = [\psi_1^h, \ldots, \psi_m^h] \subset \mathcal{V}_b^h, \ \mathcal{N}_{\mathcal{V}_b^{h'}} = [\psi_1^{h'}, \ldots, \psi_m^{h'}] \subset \mathcal{V}_b^{h'},$$

often with $m = \dim N(G_0')$. These approximations usually satisfy even for $\dim N(G_0') > 1$

$$||P^h \phi_i - \phi_i^h||_{\mathcal{U}_b^h} \xrightarrow{h \to 0} 0, \ ||P^{h'} \phi_i' - \phi_i^{'h}||_{\mathcal{U}_b^{h'}} \xrightarrow{h \to 0} 0 \text{ and} \tag{8.9}$$
$$||Q^{'h} \psi_i - \psi_i^h||_{\mathcal{V}_b^h} \xrightarrow{h \to 0} 0, \ ||\hat{Q}^{h'} \psi_i' - \psi_i^{h'}||_{\mathcal{V}_b^{h'}} \xrightarrow{h \to 0} 0. \tag{8.10}$$

The choice (8.9) automatically satisfies (8.5) or (8.6) for small h. With a possibly perturbed $\tilde{A}^h \approx A^h = Q^{'h} A|_{\mathcal{U}_b^h}$, $\tilde{A}^h \in \mathcal{L}(\mathcal{U}_b^h, \mathcal{V}_b^h)$ we define the operator $\tilde{L}^h \in \mathcal{L}(\mathcal{U}_b^h \times \mathbb{R}^m, \mathcal{V}_b^h \times \mathbb{R}^m)$ by

$$\tilde{L}^h \begin{pmatrix} u^h \\ \alpha^h \end{pmatrix} := \begin{pmatrix} \tilde{A}^h u^h + \sum_{i=1}^m \alpha_i^h \psi_i^h \\ < u^h, \phi_1^{h'} >_{\mathcal{U}_b^h \times \mathcal{U}_b^{h'}} \\ \vdots \\ < u^h, \phi_m^{h'} >_{\mathcal{U}_b^h \times \mathcal{U}_b^{h'}} \end{pmatrix}. \tag{8.11}$$

We interpret $\tilde{L}^h$ as a perturbation of the operator $Q_{ext}^{'h} L|_{\mathcal{U}_b^h \times \mathbb{R}^m}$ with

$$||Q_{ext}^{'h} L|_{\mathcal{U}_b^h \times \mathbb{R}^m} - \tilde{L}^h||_{\mathcal{V}_b^h \times \mathbb{R}^m \leftarrow \mathcal{U}_b^h \times \mathbb{R}^m} \leq ||\tilde{A}^h - Q^{'h} A|_{\mathcal{U}_b^h}||_{\mathcal{V}_b^{h'} \leftarrow \mathcal{U}_b^h} \tag{8.12}$$
$$+ \sum_{i=1}^m \left\{ ||P^{h'} \phi_i - \phi_i^{h'}||_{\mathcal{U}_b^{h'}} + ||Q^{'h} \psi_i - \psi_i^h||_{\mathcal{V}_b^h} \right\}. \tag{8.13}$$

Now we can state, compare Theorem 7.2.6

Theorem 8.1.2. *Let $(\mathcal{U}_b^h, \mathcal{V}_b^h)_{h \in H}$ be a pair of approximating spaces with $\dim \mathcal{U}_b^h = \dim \mathcal{V}_b^h$ and $f \in \mathcal{V}_b$. Assume $(\tilde{A}^h)_{h \in H} \in (\mathcal{L}(\mathcal{U}_b^h, \mathcal{V}_b^h))_{h \in H}$ and $(\tilde{f}^h)_{h \in H} \in (\mathcal{V}_b^h)_{h \in H}$, are "proper" perturbations of $(Q^{'h} A|_{\mathcal{U}_b^h})_{h \in H}$ and $(Q^{'h} f)_{h \in H}$, respectively., i.e.*

$$||\tilde{A}^h - Q'^h A|_{\mathcal{U}_b^h}||_{\mathcal{V}_b^h \leftarrow \mathcal{U}_b^h} \to 0 \text{ and } ||f^h - Q'^h f||_{\mathcal{V}_b} \to 0 \text{ for } h \to 0.$$

We define $\tilde{L}^h$ as in (8.11) and assume the appropriate regularity, hence, let (8.5) or (8.6) be correct for $\mathcal{N}_{\mathcal{U}_b}$ and $\mathcal{N}_{\mathcal{V}_b}$ spanned by $\phi_i{}'$ and $\psi_i^{h'}$, respectively. Then, for sufficiently small h,

$$(\tilde{L}^h)_{h\in H} \text{ is stable} \Leftrightarrow (Q'^h_{ext} L|_{\mathcal{U}_b \times \mathbb{R}^m})_{h\in H} \text{ is stable.}$$

the solution $(\tilde{u}_0^h,\ \tilde{\alpha}_0^h)^T$ of the perturbed Petrov-Galerkin equation

$$\tilde{L}^h(\tilde{u}_0^h,\ \tilde{\alpha}_0^h)^T = (f_0^h,\ 0)^T,\ (\tilde{u}_0^h,\ \tilde{\alpha}_0^h) \in \mathcal{U}_b^h \times \mathbb{R}^m,$$

exists uniquely and converges to the unique solution $(u_0, \alpha_0)^T$ of (8.7), more precisely, for sufficiently small h, we have the error estimate

$$\|\tilde{u}_0^h - u_0\|_{\mathcal{U}} + ||\tilde{\alpha}_0 - \alpha_0|| \le C\big(\|I - P^h\|_{\mathcal{U}_b \leftarrow \mathcal{U}_b}\ \|f\|_{\mathcal{V}} \tag{8.14}$$

$$+\|\tilde{f}^h - Q'^h f\|_{\mathcal{V}} + \|\tilde{A}^h - Q'^h A^h|_{\mathcal{U}_b^h}\|_{\mathcal{V}_b \leftarrow \mathcal{U}_b} + \sum_{i=1}^{m} \big\{||P^h\phi_i - \phi_i^h||_{\mathcal{U}_b^h}$$

$$+||Q'^h\psi_i - \psi_i^h||_{\mathcal{V}_b^h} + ||P^{h'}\phi_i' - \phi_i^{h'}||_{\mathcal{U}_b^{h'}} + ||\hat{P^{h'}}\psi_i' - \psi_i^{h'}||_{\mathcal{V}_b^{h'}}\big\}\big)$$

The constant C is mainly the stability constant $\|(\tilde{L}^h)^{-1}\|_{\mathcal{U}_b \times \mathbb{R}^m \leftarrow \mathcal{V}_b \times \mathbb{R}^m}$.

Proof: Under the above assumptions and with (8.12) we obtain

$$||Q'^h_{ext} L|_{\mathcal{U}_b^h \times \mathbb{R}^m} - L^h||_{\mathcal{V}_b^{h'} \times \mathbb{R}^m \leftarrow \mathcal{U}_b^h \times \mathbb{R}^m} \to 0, h \to 0.$$

Thus, the rest of the proof follows from Theorem 7.2.6.

9. Application to the Navier-Stokes operator

The stationary Navier-Stokes equation has the form

$$G(u,p) := \begin{pmatrix} -\nu\Delta u + \sum_{i=1}^{n} u_i \partial_i u + \text{grad } p \\ div\ u \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix} \text{ in } \Omega$$
$$u = 0 \text{ on } \Gamma = \partial\Omega, \int_\Omega p dx = 0, \text{ where} \tag{9.1}$$

$u = (u_1, \ldots, u_n)^T, v = (v_1, \ldots, v_n)^T, f = (f_1, \ldots, f_n)^T : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n, p : \Omega \to \mathbb{R}, n \leq 3;$

here u, p and f denote velocity, pressure and forcing term of an incompressible medium and v, q test functions, respectively. The condition $\int_\Omega p dx = 0$ is imposed to guarantee a unique p. The linearization of $G(u,p)$ for $(u_0 \equiv 0, p_0)$ applied to an increment $(u,p), u \neq 0$ is the Stokes operator, S. It has the form

$$S(u,p) := \begin{pmatrix} -\Delta u + \nabla p \\ -div\ u \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix} \text{ in } \Omega, \tag{9.2}$$
$$u = 0 \text{ on } \Gamma = \partial\Omega, \quad \int_\Omega p dx = 0.$$

We can use either directly all the stability results available for (9.1), see e.g. [63, 30], and combine it with the results in Chapter 8. Or we interpret S and its generalization, a saddle point problem see (9.10), as bordered system. This is shown to satisfy the stability conditions in Theorem 7.2.1 in $'$ 7, if the well-known Brezzi-Babuska conditions are satisfied for the Stokes operators, see (9.13) and Theorem 9.1.2. We show that, for moderate (and not so important) ν, the linearized Navier-Stokes operator represents a compact perturbation of S, hence satisfies the conditions of Theorem 7.2.3. As bi-dual approximating spaces we use three different types of finite elements, see Examples 9.1 and 9.2.
For this second approach, see above, we only study moderate Reynolds numbers and therefor normalize $\nu = 1$ and require

$$\Omega \subset \mathbb{R}^n, \text{ bounded and } \Gamma = \partial\Omega \text{ Lipschitz continuous.} \tag{9.3}$$

9.1 The Stokes operator

For the Stokes operator, essentially following the presentation in Hackbusch [37, 38], we introduce $\mathcal{V}_b, W$, as

$$\begin{aligned} \mathcal{V}_b &:= \underbrace{H_0^1(\Omega) \times \cdots \times H_0^1(\Omega)}_{n \quad times}, \\ W &:= L^2_*(\Omega) := \{p \in L^2(\Omega) : \int_\Omega p(x)dx = 0\} \text{ and} \\ \mathcal{U} &:= \mathcal{V}_b \times W. \end{aligned} \tag{9.4}$$

To avoid too many technical details, we restrict the discussion to Petrov-Galerkin methods. With the inner product $< \cdot, \cdot >_{\mathbb{R}^n}$ we apply the usual Green formula to $(-\Delta u + \nabla p, v)_{\mathcal{V}_b' \times \mathcal{V}_b} := (< -\Delta u + \nabla p, v >_{\mathbb{R}^n})_2 := \int_\Omega < -\Delta u + \nabla p, v >_{\mathbb{R}^n} dx$ to obtain

$$\begin{aligned} &(-\Delta u + \nabla p, v)_{\mathcal{V}_b' \times \mathcal{V}_b} = a(u,v) + b(p,v) = f(v), \quad \forall \ \ u, v \in \mathcal{V}_b, p \in W \\ &-\int_\Omega q(x) div\ u(x)dx = b(q,u) = 0 \ \ \forall \ \ q \in W, u \in \mathcal{V}_b \text{ with } div\ u = 0; \end{aligned}$$

here the bilinear forms $a(\cdot,\cdot), b(\cdot,\cdot)$ are defined as

$$\begin{aligned} a(u,v) &:= \int_\Omega < \nabla u(x), \nabla v(x) >_{\mathbb{R}^n} dx \text{ for } u, v \in \mathcal{V}_b, \\ b(p,v) &:= -\int_\Omega p(x) div\ v(x)dx \text{ for } p \in L^2_*(\Omega), v \in \mathcal{V}_b. \end{aligned} \tag{9.5}$$

For bounded Ω, see (9.3), the $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are continuous, [37, 38], Lemma 12.2.12,

$$|a(u,v)| \le C_a \|u\|_{\mathcal{V}} \cdot \|v\|_{\mathcal{V}}, |b(p,v)| \le C_b \|v\|_{\mathcal{V}} \cdot \|p\|_W. \tag{9.6}$$

This yields the weak formulation of (9.2), where we have replaced f and 0 by f^1 and f^2, respectively.

$$\begin{aligned} &\text{For given } f = (f^1, f^2) \in \mathcal{V}_b' \times W' \text{ determine } u_0 \in \mathcal{V}_b, p_0 \in W \text{ such that} \\ &\quad a(u_0, v) + b(p_0, v) = f^1(v) \quad \forall \ \ v \in \mathcal{V}_b, \\ &\quad b(q, u_0) = f^2(q) \ \forall \ \ q \in W. \end{aligned} \tag{9.7}$$

To obtain the situation in Chapter 7, we replace the $\mathcal{V}_b, W, a(\cdot,\cdot), b(\cdot,\cdot), f(\cdot), 0$ in (9.4) - (9.7) by general Banach spaces $\mathcal{V}_b, W$, continuous bilinear forms a, b and linear forms $f^1 \in \mathcal{V}_b', f^2 \in W'$, respectively. Furthermore, we introduce $\mathcal{U} := \mathcal{V}_b \times W$ and $x := (u,p), y := (v,q) \in \mathcal{U}$ to obtain

$$\begin{aligned} c(x,y) &:= c\left(\begin{pmatrix} u \\ p \end{pmatrix}, \begin{pmatrix} v \\ q \end{pmatrix}\right) := a(u,v) + b(p,v) + b(q,u), \\ f(y) &:= f^1(v) + f^2(q), \quad f \in \mathcal{U}. \end{aligned} \tag{9.8}$$

Then $c(\cdot,\cdot)$ is a continuous bilinear form on $\mathcal{U}\times\mathcal{U}$. Now, let $A \in \mathcal{L}(\mathcal{V}_b, \mathcal{V}_b')$, $B \in \mathcal{L}(W, \mathcal{V}_b')$ and $C \in \mathcal{L}(\mathcal{U},\mathcal{U}')$ be the linear operators induced by $a(\cdot,\cdot), b(\cdot,\cdot)$ and $c(\cdot,\cdot)$, respectively. Then

$$C := \begin{pmatrix} A & B \\ B^d & 0 \end{pmatrix} \in \mathcal{L}(\mathcal{U},\mathcal{U}'), \quad Cx = \begin{pmatrix} A & B \\ B^d & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} Au + Bp \\ B^d u \end{pmatrix}, \tag{9.9}$$

$$\begin{aligned} c(x,y) = \langle Cx, y\rangle_{\mathcal{U}'\times\mathcal{U}} &= \left\langle \begin{pmatrix} Au + Bp \\ B^d u \end{pmatrix}, \begin{pmatrix} v \\ q \end{pmatrix} \right\rangle_{\mathcal{U}'\times\mathcal{U}} \\ &= \langle Au, v\rangle_{\mathcal{V}_b'\times\mathcal{V}_b} + \langle Bp, v\rangle_{\mathcal{V}_b'\times\mathcal{V}_b} + \langle B^d u, q\rangle_{W'\times W} \\ &= a(u,v) + b(p,v) + b(q,u) = c(x,y). \end{aligned}$$

Certainly, $c(\cdot,\cdot)$ is neither elliptic nor coercive, since $c(x,y) = 0$ for all $x = (0,p)$.
In this generalized context the Stokes problem, see(9.5), (9.7), can be formulated as in (4.113) or (4.114), (4.115), as a so called *saddle point problem*

$$\begin{aligned} &\text{for given } f \in \mathcal{U}' \text{ determine } x_0 = (u_0, p_0) \in \mathcal{U} \text{ such that} \\ &\quad c(x_0, y) = \langle f, y\rangle_{\mathcal{U}'\times\mathcal{U}} \text{ for all } y \in \mathcal{U} \text{ or equivalently} \\ &\quad Cx_0 \quad = f \in \mathcal{U}'. \end{aligned} \tag{9.10}$$

Nečas shows, see Hackbusch Satz 12.2.14, [37]

Theorem 9.1.1. *Let Ω satisfy (9.3) and let C be defined by (9.8) (9.9) with the operators A, B as induced by (9.5). Then C is boundedly invertible, $C^{-1} \in \mathcal{L}(\mathcal{U}',\mathcal{U})$, and the solution $(u_0, p_0) \in \mathcal{V}_b \times W = (H_0^1(\Omega))^3 \times L_*^2(\Omega)$ satisfies*

$$\|u_0\|_{\mathcal{V}_b)} + \|p_0\|_{L^2(\Omega)} \le C_\Omega(\|f^1\|_{\mathcal{V}_b} + \|f^2\|_W). \tag{9.11}$$

Smoother f^1, f^2 and Ω imply smoother u, p and refine (9.11).
For example, let $\mathcal{V}_b$, W be Banach spaces, $A \in \mathcal{L}(\mathcal{V}_b, \mathcal{V}_b')$, $B \in \mathcal{L}(W, \mathcal{V}_b')$ and $(\mathcal{V}_b^h \times W^h)_{h\in H}$ and $(\mathcal{V}_b^{h'} \times W^{h'})_{h\in H}$ be approximating spaces for $\mathcal{V}_b \times W$ and $\mathcal{V}_b' \times W' = (\mathcal{V}_b \times W)'$, resp. Consider the above operator $C := \left(\begin{smallmatrix} A & B \\ B^d & 0 \end{smallmatrix}\right)$ with bounded inverse $C^{-1} \in \mathcal{L}(\mathcal{V}_b' \times W', \mathcal{V}_b \times W)$. One can show that $\dim \mathcal{V}_b^h \ge \dim W^h$ is a necessary criterion for the stability of the discrete operator, cf. Hackbusch [37]. I.e. if one chooses inappropriate approximating spaces, the stability of a discrete operator, C^h, does not necessarily follow from the existence of the bounded inverse, C^{-1}, of C. To introduce appropriate way, approximating spaces for $\mathcal{V}_b$ and W : Approximating spaces $(\mathcal{V}_b^h)_{h\in H}, (W^h)_{h\in H}$ yield $(\mathcal{V}_b^h \times W^h)_{h\in H}$, obviously again approximating spaces for $\mathcal{V}_b \times W$. They are bi-dual Petrov-Galerkin approximations for $\mathcal{U}$ simultaneously with $\mathcal{V}_b^h, W^h$ for $\mathcal{V}_b, W$. Then we replace (9.7) and (9.10) by

$$\begin{aligned} &\text{for given } f = (f^1, f^2) \in \mathcal{V}_b' \times W' \text{ determine } (u_0^h, p_0^h) \in \mathcal{V}_b^h \times W^h \text{ by} \\ &\quad a(u_0^h, v^h) + b(p_0^h, v^h) = f^1(v^h) \quad \forall \quad v^h \in \mathcal{V}_b^h, \\ &\quad\quad b(q^h, u_0^h) = f^2(q^h) \quad \forall \quad q^h \in W^h \end{aligned} \tag{9.12}$$

and

for given $f \in \mathcal{U}'$ determine $x_0^h = (u_0^h, p_0^h) \in \mathcal{U}^h = \mathcal{V}_b^h \times W^h$ by

$$c(x_0^h, y^h) = f(y^h) \quad \forall \quad y^h = (v^h, q^h) \in \mathcal{U}^h, \text{ equivalently } C^h x_0^h = f^h.$$

To use the results in Chapters 7 and 8 we have to combine $C^{-1} \in \mathcal{L}(\mathcal{U}', \mathcal{U})$, see Theorem 7.2.3, and the stability of C^h in (9.13). The famous Brezzi-Babuska-conditions guarantees stability. It has, for the Stokes operator, the form, [37],

$$\begin{aligned} \text{Let } \mathcal{V}_{b,0}^h &:= \{v \in \mathcal{V}_b^h : b(y,v) = 0 \quad \forall \quad y \in W^h\} \text{ and} \\ &\inf \quad \{\sup\{|a(x,y)| : y \in \mathcal{V}_{b,0}^h, \|y\|_{\mathcal{V}} = 1\} : x \in \mathcal{V}_{b,0}^h, \|x\|_{\mathcal{V}} = 1\} = \alpha_h > 0, \\ &\inf \quad \{\sup\{|b(w,y)| : y \in \mathcal{V}_b^h, \|y\|_{\mathcal{V}} = 1\} : w \in W, \|w\|_{\mathcal{V}} = 1\} = \beta_h > 0 \\ &\text{and} \quad \alpha_h \geq \alpha > 0, \beta_h \geq \beta > 0. \end{aligned} \tag{9.13}$$

A combination with Hackbusch Satz 12.3.11, [37] yields.

Theorem 9.1.2. *Let bi-dual approximating spaces $(\mathcal{U}^h)_{h \in H}$ be given with $\dim W^h \leq \dim \mathcal{V}_b^h < \infty$, and the f_1, f_2, and $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ in (9.7) be continuous linear and bilinear forms on $\mathcal{V}_b, W$ and $\mathcal{V}_b \times \mathcal{V}_b, W \times \mathcal{V}_b$, respectively. Then the discrete problem (9.12) is uniquely solvable, if the Brezzi-Babuska inf-sup-condition (9.13) is satisfied. If Ω is chosen according to (9.3), (9.13). Then C^h in (9.13) is stable and, under the conditions of Theorem 7.2.1, C is invertible, $C^{-1} \in \mathcal{L}(\mathcal{U}', \mathcal{U})$ in (9.10) . The conditions of Theorems ?? and 7.2.6 are satisfied and*

$$\|u_0^h\|_{\mathcal{V}}^2 + \|p_0^h\|_W^2 \leq C_h(\|f_1\|_{\mathcal{V}_b'}^2 + \|f_2\|_{W'}^2),$$

with $C_h = C_h(\alpha, \beta, C_a, C_b)$ and C_a, C_b in (9.6).

$$\begin{aligned} \|u_0 - u_0^h\|_{\mathcal{V}}^2 + \|p_0 - p_0^h\|_W^2 \leq (1 + C \ \|(P_{ext}^h C|_{\mathcal{V}_b^h \times W^h})^{-1}\|_{\mathcal{V}_b^h \times W^h \leftarrow \mathcal{V}_b^h \times W^h} \\ \times \|I - P_{ext}^h\|_{\mathcal{V}_b \times W \leftarrow \mathcal{V}_b \times W} \|(C)^{-1}\|_{\mathcal{V}_b \times W \leftarrow \mathcal{V}_b \times W} (\|f_1\|_{\mathcal{V}_b'}^2 + \|f_2\|_{W'}^2). \end{aligned} \tag{9.14}$$

For the specific case of the Stokes operator we only present three examples for $\mathcal{U}^h = \mathcal{V}_b^h \times W^h$, see [37]. For other examples see [63, 63, 30]. Let $\Omega \subset \mathbb{R}^2$ be a polygon, hence $\mathcal{V}_b = H_0^1(\Omega) \times H_0^1(\Omega)$ and τ^h a quasi-uniform triangulation for Ω. That is, there exists a fixed $\sigma_0 > 0$ such that for every triangle $T \in \tau^h$ the quotient $\dim T /$ inradius $T \leq \sigma_0$, where inradius T is defined as maximal radius of any circle in T.

Example 9.1 Piecewise linear elements and bubble functions: *We define, for linear functions $u \in \mathcal{P}_1$*

$$W^h := \{q^h : \quad \forall \quad T \in \tau : q^h|_T \in \mathcal{P}_1, \ \int_\Omega q^h dx = 0\} \subset L_*^2(\Omega). \tag{9.15}$$

To introduce the bubble functions u on τ let $\tilde{T}$ represent the reference triangle $\tilde{T} = \{(\xi\eta) : \xi, \eta > 0, \xi + \eta < 1\}$ and Φ the bijective affine mapping $\Phi : \tilde{T} \to T$. Define, for every $T \in \tau^h$,

$$\begin{aligned}
\tilde{u}(\xi,\eta) &:= \xi \cdot \eta(1-\xi-\eta) \text{ for } (\xi,\eta) \in \tilde{T}, \quad \text{and } := 0 \text{ otherwise}, \\
u_T(x,y) &:= \tilde{u}(\Phi^{-1}(x,y)) \text{ for } (x,y) \in T \text{ and } := 0 \text{ otherwise, } u_T \in H_0^1(\Omega) \text{ and} \\
\mathcal{V}_{b1}^h &:= \{v^h \in H_0^1(\Omega) : \ \forall \ T \in \tau : v^h|_\tau \text{ linear combinations of } \mathcal{P}_1 \\
&\quad \text{and bubble functions on } \tau\} \\
\mathcal{V}_b^h &: = \mathcal{V}_{b1}^h \times \mathcal{V}_{b1}^h, \quad \mathcal{U}^h = \mathcal{V}_b^h \times W^h.
\end{aligned}$$

■

Example 9.2 *1)* Piecewise linear elements on $\tau_{h/2}$ and τ^h : *Let $\tau_{h/2}$ be defined by replacing each $T \in \tau^h$ by four congruent sub-triangles. Then*

$$\begin{aligned}
&\mathcal{V}_{b1}^h := \{v_1^h \in H_0^1(\Omega) : \text{ linear elements on } \tau_{h/2}\}, \quad \mathcal{V}_b^h := \mathcal{V}_{b1}^h \times \mathcal{V}_{b1}^h, \\
&W^h := \{q \in L_*^2(\Omega) : \text{ linear elements on } \tau^h\}, \quad \mathcal{U}^h = \mathcal{V}_b^h \times W^h.
\end{aligned}$$

2) Piecewise quadratic and linear elements on τ^h : *Let*

$$\begin{aligned}
&\mathcal{V}_{b,2}^h := \{v_1^h \in H_0^1(\Omega) : \text{quadratic elements on } \tau^h\}, \quad \mathcal{V}_b^h := \mathcal{V}_{b,2}^h \times \mathcal{V}_{b,2}^h, \\
&W^h := \{q \in L_*^2(\Omega :) \text{ linear elements on } \tau^h\}, \quad \mathcal{U}^h = \mathcal{V}_{b,2}^h \times W^h.
\end{aligned} \tag{9.16}$$

■

The FEs in Examples 9.1 and 9.2 approximate $\mathcal{U}$ as required in Chapter 7.

9.2 The Linearized Navier-Stokes operator

Now we discuss the linearized Navier-Stokes operator again for $\Omega \subset \mathbb{R}^n$ and $\mathcal{V}_b = (H_0^1(\Omega))^n$ and follow [63, 63, 30]. Again we multiply with the test function $v = (v_1, \ldots, v_n)$ and use the Green formula to obtain, with $a(\cdot,\cdot), b(\cdot,\cdot)$ in (9.5)

For given $(f^1, 0) \in \mathcal{V}_b' \times W'$ determine $u_0 \in \mathcal{V}_b, p_0 \in W$ such that
$\int_\Omega \langle -\nu \Delta u_0 + \sum_{i=1}^n (u_0)_i \partial_i u_0 + \text{grad } p_0, v\rangle_{\mathbb{R}^n} dx$
$= \nu a(u_0, v) + d(u_0, u_0, v) + b(p_0, v) = f^1(v) \ \ \forall \ \ v \in \mathcal{V}_b$ and
$b(p_0, v) = 0 \quad \forall \ \ v \in \mathcal{V}_b$; here
$d(u,v,w) := \sum_{i,j=1}^n \int_\Omega u_i (\partial_i v_j) w_j dx$

For bounded Ω, see (9.3) and $n \leq 4$, see Temam,[63] Lemma 1.2, Ch. II, Chapter 1.

$$d(u,v,w) \text{ is a bounded tri-linear form on } \mathcal{V}_b \times \mathcal{V}_b \times \mathcal{V}_b.$$

To linearize we consider for fixed u, v and small w

$$d(u+w, u+w, v) - d(u,u,v) = d(u,w,v) + d(w,u,v) + o(w).$$

An analogous results holds for the nonlinear (9.1): We obtain

$$G'(u,p)(w,r) = \begin{pmatrix} -\nu\Delta w + \sum_{i=1}^{n}(u_i\partial_i w + w_i\partial_i u) + grad\ r \\ div\ w \end{pmatrix} \tag{9.17}$$

and, with the $a(\cdot,\cdot), b(\cdot,\cdot), d(\cdot,\cdot)$ in (9.5) and (9.17).

$$\left(\langle (G'(u,p)(w,r)),(v,q)\rangle_{\mathbb{R}^{n+1}}\right)_{(L^2(\Omega))^2} = \begin{pmatrix} \nu a(w,v) + d(u,w,v) + d(w,u,v) + b(r,v) \\ b(q,w) \end{pmatrix}$$
$$\forall (v,q) \in \mathcal{V}_b \times W = \mathcal{U}.$$

Now we consider the $d(u,w,v), d(w,u,v)$

$$d(u,w,v) = \sum_{i,j=1}^{n} \int_\Omega u_i(\partial_i w_j)v_j dx, \quad \text{for } v \in \mathcal{V}_b \tag{9.18}$$

$$d(w,u,v) = \sum_{i,j=1}^{n} \int_\Omega w_i(\partial_i u_j)v_j dx$$
$$= -\sum_{i,j=1}^{n} \int_\Omega w_i u_j \partial_i v_j dx \quad \forall \quad v \in \mathcal{V}_b.$$

For fixed u, the $d(u,w,v) + d(w,u,v)$ are continuous bilinear forms, corresponding to the (sum of) operator(s) $\sum_{i,j=1}^{n}(u_i\partial_i w + w_i\partial_i u)$ in (9.17). This is, for fixed u, linear and bounded in $w \in \mathcal{V}_b$. Then the continuous bilinear forms $d(u,w,v)$ and $d(w,u,v) \in \mathbb{R}$ define elements

$$d(u,w,\cdot), d(w,u,\cdot) \in \mathcal{V}_b'$$

hence define linear continuous operators

$$D_1, D_2 \in \mathcal{L}(\mathcal{V}_b, \mathcal{V}_b') \text{ as } D_1 w := d(u,w,\cdot), \quad D_2 w := d(w,u,\cdot), \text{ for } u \text{ fixed} \tag{9.19}$$

By (9.17), mind that u is fixed, see (9.18)

$$\langle D_1 w, v\rangle_{\mathcal{V}_b' \times \mathcal{V}_b} = d(u,w,v), \quad \langle D_2 w, v\rangle_{\mathcal{V}_b' \times \mathcal{V}_b} = d(w,u,v). \tag{9.20}$$

Now the embedding $I : H_0^1(\Omega) \to L^2(\Omega)$ is continuous and compact and (9.20) shows that

$$D_1 v = D_1 I v \ \forall \ \ v \in H_0^1(\Omega).$$

Hence, as a product of a compact and a continuous operator, $D_1 = D_1 I$ is a compact operator. The same is correct for D_2 as well.
Similarly to the transformation from (9.7) to (9.10) we use here, with the factor ν in (9.1),

$$C := \begin{pmatrix} \nu A & B \\ B^d & 0 \end{pmatrix}, D := \begin{pmatrix} D_1 + D_2 & 0 \\ 0 & 0 \end{pmatrix}. \tag{9.21}$$

and, with slightly different $x = (w, r), y = (v, q)$,

$$\begin{aligned} (C + D)x &= \begin{pmatrix} \nu Aw + Br + D_1 w + D_2 w \\ B^d w \end{pmatrix}, \\ \langle (C + D)x, y \rangle_{\mathcal{U}' \times \mathcal{U}} &= \nu \langle Aw, v \rangle_{\mathcal{V}_b' \times \mathcal{V}_b} + \langle Br, v \rangle_{\mathcal{V}_b' \times \mathcal{V}_b} \\ &\quad + \langle D_1 w + D_2 w, v \rangle_{\mathcal{V}_b' \times \mathcal{V}_b} + \langle B^d w, q \rangle_{W' \times W} \end{aligned}$$

and formulate

$$\begin{aligned} &\text{for given } f \in \mathcal{U}' \text{ determine } x_0 = (w_0, r_0) \text{ such that} \\ &\qquad \langle (C + D)x_0, y \rangle_{\mathcal{U}' \times \mathcal{U}} = \langle f, y \rangle_{\mathcal{U}' \times \mathcal{U}} \quad \forall \quad y \in \mathcal{U} \\ &\qquad\quad \text{or equivalently } (C + D)x_0 = f \in \mathcal{U}'. \end{aligned}$$

Again Remark **??**) applies. So we can guarantee the existence of $C + D$ if -1 is not an eigenvalue of the compact operator $C^{-1}D$. Resuming these results we have the following

Theorem 9.2.1. . *For bi-dual approximating spaces $(\mathcal{U}^h)_{h \in H}$ let the Stokes operator S or its generalization C for S and C see (9.2) and (9.9), respectively, yield stable discretizations, see Theorem 9.1.2. Hence $(\mathcal{U}^h)_{h \in H}$ applied to a bijective $G'(u, p)$ and to $G(u, p)$ have the desirable convergence properties of Theorems* **??** *and 7.2.6. Furthermore for non-bijective $G'(u, p)$, the bifurcation numerics, based on bordered systems, yield converging bifurcations scenarios, satisfying the estimate in Theorems 8.1.1 and 8.1.2.*

References

1. E. L. Allgower, P. Ashwin, K. Böhmer, and Z. Mei. Liapunov-Schmidt reduction for a bifurcation problem with periodic boundary conditions on a square domain. In E. L. Allgower, K. Georg, and R. Miranda, editors, *Exploiting Symmetry in Applied and Numerical Analysis*, volume 29 of *Lectures in Applied Mathematics*, pages 11–22, Providence, RI, 1993. American Mathematical Society.
2. P. Ashwin, K. Böhmer, and Z. Mei. A numerical Liapunov-Schmidt method for finitely determined problems. In E.L. Allgower, K. Georg, and R. Miranda, editors, *Exploiting Symmetry in Applied and Numerical Analysis*, Lectures in Applied Mathematics, pages 49–69, Providence, 1993. AMS.
3. P. Ashwin, K. Böhmer, and Z. Mei. A numerical Liapunov-Schmidt method with applications to Hopf bifurcation on a square. *Math. Comp.*, 64:649–670 and S29–S22, 1995.
4. R. E. Bank. *PLTMG: Asoftware Package for Solving elliptoc Partial Differential Equations. User's Guide 6.0.* SIAM, Philadelphia, 1990.
5. Ch. Bernard and Y. Maday. Spectral methods. In P.G. Ciarlet and J.L. Lions, editors, *Handbook of Numerical Analysis*, volume V of *Techniques of Scientific Computing (Part 2)*. Elsevier Science B.V., 1997.
6. K. Böhmer. On hybrid methods for bifurcation studies for general operator equations. In B. Fiedler, editor, *Ergodic theory, Analysiss, and Efficient Simulation of Dynamical Systems*, pages 73–107,2001, Berlin, Heidelberg, New York. Springer.
7. K. Böhmer. On numerical bifurcation studies for general operator equations. In J. Sprekels B. Fiedler, K. Gröger, editor, *International Conference on Differential Equations, 1999*, volume 2, pages 877–883,2000, Singepore. World Scientific.
8. K. Böhmer. On a numerical Liapunov-Schmidt method for operator equations. Technical Report 2, Philipps-Universität Marburg, Fachbereich Mathematik, 1989.
9. K. Böhmer. On a numerical Liapunov-Schmidt method for operator equations. *Computing 51*, pages 237–269, 1993.
10. K. Böhmer and S. Dahlke. Stability and convergence for wavelets applied to general elliptic problems. in preparation, University of Marburg, 2002.
11. K. Böhmer, C. Geiger, and J. Rodriguez. On a numerical Liapunov-Schmidt spectral method, part I: Review of spectral methods. Technical report, SP-Report, 1996.
12. K. Böhmer, C. Geiger, and J. Rodriguez. On a numerical Liapunov-Schmidt spectral method, part II: The reduction method and its applications. Technical report, SP-Report, 1997.
13. K. Böhmer, C. Geiger, and J. Rodriguez. On a numerical Liapunov-Schmidt spectral method and applications to biological pattern formation. *SIAM J. Numer. Anal.*, ?:?, 2002.

14. K. Böhmer and Z. Mei. On a numerical Liapunov-Schmidt method. In E.L. Allgower and K. Georg, editors, *Computational Solutions of Nonlinear Systems of Equations*, volume 26 of *Lectures in Applied Mathematics*, pages 79–98, Providence, 1990. AMS.
15. K. Böhmer and N. Sassmannshausen. Numerical Liapunov-Schmidt spectral method for k-determined problems. In T.Healey, editor, *Computational Methods and Bifurcation Theory with Applications*, volume 170 of *Computer Methods in Applied Mechanics and Engineering*, pages 277–312, 1999.
16. K. Böhmer and N. Sassmannshausen. Stability for generalized petrov-galerkin methods applied to bifurcation, 2001. submitted to ZAMM.
17. D. Braess. *Finite Elemente*. Springer, Berlin, 1997.
18. S.C. Brenner and L.R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer Verlag, New York, 1996.
19. F. Brezzi, J. Rappaz, and P.A. Raviart. Finite-dimensional approximation of nonlinear problems, I. Branches of nonsingular solutions. *Numer. Math.*, 36:1–25, 1980.
20. F. Brezzi, J. Rappaz, and P.A. Raviart. Finite-dimensional approximation of nonlinear problems, II. Limit points. *Numer. Math.*, 37:1–28, 1981.
21. F. Brezzi, J. Rappaz, and P.A. Raviart. Finite-dimensional approximation of nonlinear problems, III. Simple bifurcation points. *Numer. Math.*, 38:1–30, 1981.
22. G. Caloz and J. Rappaz. Numerical analysis for nonlinear and bifurcation problems. In P.G. Ciarlet and J.L. Lions, editors, *Handbook of Numerical Analysis*, volume V of *Techniques of Scientific Computing (Part 2)*. Elsevier Science B.V., 1998.
23. C. Canuto, M. Y. Hussaini, A. Quarteroni, and T.A. Zang. *Spectral methods in fluid dynamics*. 3. Auflage. Springer Verlag, Berlin, 1997.
24. P.-G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North-Holland, Amsterdam, New York, 1978.
25. P.-G. Ciarlet and P.A. Raviart. Conforming and nonconforming finite element methods for solving the stationary stokes problem. *R.A.I.R.O.*, 1973.
26. M. Crouzeix and J. Rappaz. *On numerical approximation in bifurcation theory*. Masson, Paris and Springer, Berlin, 1990.
27. E.J. Doedel. On the construction of discretizations of elliptic partial differential equations. *Journal of Diff. Equations and Applications*, 3:389–416, 1998.
28. E.J. Doedel and H.Sharifi. Collocation methods for continuation problems in nonlinear pdes. Technical report, Department of Computer Science, Concordia University Montreal, Canada, 1999. Applied Mathematics.
29. D. Gilbarg and Trudinger N.S. *Elliptic Partial Differential Equations of Second Order*. Springer Verlag, 1983.
30. V. Girault and P. A. Raviart. *Finite Element Methods for Navier-Stokes Equations*. Springer Series in Computational Mathematics, 5. Springer-Verlag, Berlin New York, 1986.
31. D. Gottlieb and S. A. Orszag. Numerical analysis of spectral methods, theory and applications. Technical report, Society for Industrial and Applied Mathematics,, 1977.
32. A. Griewank and G. W. Reddien. Characterization and computation of generalized turning points. *SIAM J. Numer. Anal.*, 21:176–185, 1984.
33. A. Griewank and G. W. Reddien. The approximate solution of defining equations for generalized turning points. *SIAM J. Numer. Anal.*, 33:1912–1920, 1996.

34. A. Griewank and G.W. Reddien. The approximation of generalized turning points by projection methods with superconvergence to the critical parameter. *Numer. Math.*, 48:591–606, 1984.
35. A. Griewank and G.W. Reddien. Characterization and computation of generalized turning points. *SIAM J.Numer. Anal.*, 21:176–185, 1984.
36. A. Griewank and G.W. Reddien. Computation of cusp singularities for operator equations and their discretization. *Journal of Computational and Applied Mathematics, North Holland*, 26:133–153, 1989.
37. W. Hackbusch. *Theorie und Numerik elliptischer Differentialgleichungen.* Teubner Verlag, Stuttgart, 1986.
38. W. Hackbusch. *Elliptic Differential Equations: theory and numerical treatment.* Springer Verlag, Berlin, 1992.
39. M.F. Wheeler H.H.Rachford. An H^{-1} galerkin procedure for the two-point boundary value problem. In C. de Boor, editor, *Mathematical Aspects of Finite Element Methods in Partial Differential Equations*, pages 353–382. Academic Press, New York, 1975.
40. A.D. Jepson and A. Spence. On a reduction process for nonlinear equations. *SIAM J. Math. Anal.*, 20:39–56, 1989.
41. M. Lenoir. Optimal isoparametric finite elements and error estimates for domains involving curved boundaries. *SIAM J. Numer. Anal.*, 23:562–580, 1986.
42. J.T. Oden and J.N. Reddy. *An introduction to the mathematical theory of finite elements.* Wiley, New York, 1976.
43. W. Petryshyn. On the aproximation-solvability of nonlinear equations. *Math. Ann.*, 177:156–164, 1968.
44. W. Petryshyn. On the aproximation-solvability of equations involving a-proper and pseudo-a-proper mappings. *Bull. Amer. Math. Soc.*, 81:223–312, 1975.
45. W. Petryshyn. Solvability of linear and quasilinear elliptic boundary value problems via the a-proper mapping theory. *Numer. Funct. Anal. Optim.*, 2:591–635, 1980.
46. W. Petryshyn. Approximation-solvability of periodic boundary value problems via the a-proper mapping theory. In *Proc. Sympos. Pure Math.*, volume 45, pages 261–282. Amer. Math. Soc., Providence, RI, 1986.
47. J. Rappaz and G. Raugel. Finite-dimensional approximation of bifurcation problems at a multiple eigenvalue. Rapport n°71, Centre de Mathématiques appliquées, Ecole Polytechnique Palaiseau, 1981.
48. J. Rappaz and G. Raugel. Approximations of double bifurcation points for nonlinear eigenvalue problems. In J. Whiteman, editor, *Mafelap 1981*, pages 453–461. Academic Press, New York, 1982.
49. G. Raugel. Approximation numérique de problemes nonlinéaires, 1985.
50. H.J. Reinhardt. *Analysis of approximation methods for differential and integral equations.* Springer Verlag, Berlin, Heidelberg, New York, 1985.
51. N. Sassmannshausen and K. Böhmer. Petrov-Galerkin methods for projected linear operator equation, stability and convergence. *DFG-Schwerpunktprogramm Danse*, 1998. Preprint.
52. H. M. Schultz. Spline analysis. Prentice Hall, Englewood Cliffs, 1973.
53. R. Schumann. The convergence of rothe's method for parabolic differential equations. *Z. Anal. Anwendungen*, 6:559–574, 1987.
54. R. Schumann. The convergence ot rothe's method for parabolic differential equations. *Z. Anal. Anwendungen*, 6:559–574, 1987.
55. R. Schumann. *Eine neue Methode zur Gewinnung starker Regularitätsaussagen für das Signoriniproblem in der linearen Elastizitätstheorie.* PhD thesis, Karl-Marx-Universität Leipzig, 1987. Dissertation.

56. R. Schumann and E. Zeidler. The finite difference method for quasilinear elliptic equations of order 2m. *Numer. Funct. Anal. Optimiz.*, 1:161–194, 1979.
57. H. Stetter. *Analysis of discretization methods for ordinary differential equations.* Springer Verlag, Berlin-Heidelberg-New York, 1973.
58. F. Stummel. Diskrete Konvergenz linearer Operatoren I. *Math. Ann.*, 190:45–92, 1970.
59. F. Stummel. Diskrete Konvergenz linearer Operatoren II. *Math. Z.*, 120:231–264, 1971.
60. F. Stummel. Diskrete Konvergenz linearer Operatoren III. *Math. Z.*, 190:196–216, 1972. Proc. Oberwolfach, ISNM 20.
61. F. Stummel. Stability and discrete convergence of differentiable mappings. *Rev. Roum. Math. Pures e. Appl.*, 21:63–96, 1976.
62. M.E. Taylor. *Partial Differential equations III, Nonlinear equations.* Springer, New York, Berlin, Heidekberg, 1 edition, 1996.
63. R. Temam. Navier-stokes equations and nonlinear functional analysis. CBMS-NSF-Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, 1983.
64. R. Temam. *Navier-Stokes equations, theory and numerical analysis*, volume 2 of *Studies in mathematics and its applications.* North-Holland, Amsterdam, 1984.
65. L. N. Trefethen. *Spectral Methods in MATLAB.* SIAM, Philadelphia, USA, 2000.
66. G. Vainikko. *Funktionsanalysis der Diskretisierungsmethoden.* Teubner Texte zur Mathematik. Teubner, Leipzig, 1976.
67. E. Zeidler. *Nonlinear functional analysis and its applications I, fixed-point theorems.* Springer Verlag, New York, Berlin, Heidelberg, London, Paris, Tokyo, 1990.
68. E. Zeidler. *Nonlinear functional analysis and its applications II, monotone operators.* Springer Verlag, New York, Berlin, Heidelberg, London, Paris, Tokyo, 1990.

Index