Universität Karlsruhe
Rechenzentrum
D-76128 Karlsruhe
Germany

Interner Bericht Nr. 70/97

A-Posteriori Error Estimates for the Finite Element Solution of Non-Linear Variational Problems

Author: Dipl.-Math. Lutz Grosz
Numerikforschung für Supercomputer
Rechenzentrum der Universität Karlsruhe
Karlsruhe, July 17, 1997

A-Posteriori Error Estimates for the Finite Element Solution of Non-Linear Variational Problems

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der Fakultät für Mathematik der
Universität Karlsruhe
genehmigte

DISSERTATION

von
Dipl.–Math. Lutz Grosz
aus Celle

Tag der mündlichen Prüfung: 25. Juni 1997
Referent: Prof. Dr. Willi Schönauer
Korreferent: Priv. Doz. Dr. Rüdiger Weiss

Abstract

For finite element methods (FEMs) a–posteriori error estimates that base on the evaluation of the variational equation regarding higher order approximations are a very successful concept proposed by various authors. This thesis presents a very general framework for this kind of a–posteriori error estimates for non–linear variational problems on Banach spaces. The error estimates consider the errors arising from the FEM approximation, numerical integration and termination of the iterative solver. By balancing the discretization and termination error an optimal stopping criterion for the non–linear solver of the discrete variational equation is constituted. The new projecting a–posteriori error estimate is derived from the general framework. It reuses the stiffness matrix assembled during the iteration procedure. Therefore the projecting error estimate is cost–effectively computable and can be easily implemented into an existing code. Moreover it can be used for most FEM applications without any adapting to the treated variational problem. The abstract formulation is applied to a model problem to illustrate the application in the scope of the FEM. It turns out that the quality of the discussed type of error estimates is mainly influenced by the smoothness of the sought solution. Various practical examples demonstrate that the projecting error estimate works successfully for a wide range of FEM applications.

Contents

Chapter 1

Introduction

1.1 Background

The finite element method (FEM) is the most popular method to calculate approximative solutions of partial differential equations (PDEs). FEMs are successfully used in a lot of practical applications, e.g. in heat transfer analysis, see Bathe [17], in structural analysis, see Zienkiewicz [68], and in fluid dynamics, see Chung [23]. Moreover a well-developed mathematical analysis is known for a lot of FEM applications, e.g. see Brezzi [21], Ciarlet [24, 25], Girault [37], Fluegge [36].

To get an impression of what we are speaking, let us look at a simple model problem on a bounded domain $\Omega \subset \mathbb{R}^n$, see Quarteroni [52]. It is the linear Neumann boundary value problem

$$\begin{aligned} -\nabla(a\nabla u) + bu &= f \quad \text{in } \Omega \\ \frac{\partial u}{\partial n} &= 0 \quad \text{on } \partial\Omega \ . \end{aligned} \tag{1.1}$$

The material functions a and b have an upper bound C and positive lower bound c.

The *weak solution* u of this boundary value problem is given by the corresponding *variational problem* on the Hilbert space $V := H^1(\Omega)$: find a solution $u \in V$ with

$$< v, F(u) >= 0 \text{ for all } v \in V \tag{1.2}$$

where the *residual operator* F is defined by

$$< v, F(u) >:= \int_\Omega \Big(a(\nabla v)(\nabla u) + (bu - f)v \Big) dx \text{ for all } u, v \in V \ . \tag{1.3}$$

The *strong* formulation (1.1) is transformed to the weak formulation (1.2) by multiplying the strong formulation with a so–called *test function* v, integrating the resulting equation over the domain and using partial integration to move one ∇–operator in the term $\nabla(a\nabla u)$ to the test function v. The arising boundary integrals are removed by inserting the boundary condition $\frac{\partial u}{\partial n} = 0$. It is assumed that the involved functions are smooth enough to execute this procedure. However, in the final formulation (1.2) less restrictive requirements on the involved functions regarding their smoothness are needed to formulate the problem correctly as well as to prove the existency of a solution. This is the reason why in some applications the weak formulation is preferred to the strong formulation (e.g. if the material functions have jumps).

Since the material functions a and b have an upper and a positive lower bound the operator F defined by equation (1.3) is V-*elliptic*. Therefore the variational problem (1.2) has exactly one solution $u \in V$.

The variational problem (1.2) cannot be solved by a computer since V has not a finite dimension. It has to be discretized by reducing the variational problem to a finite dimensional subspace V_h of the space V:

$\mathcal{T}_h$ denotes a triangulation of the domain Ω with mesh size h, which is a subdivision of the domain Ω into so-called *elements* $T \in \mathcal{T}_h$ (e.g. triangles) of maximal diameter h. The triangulation $\mathcal{T}_h$ has to fulfil specific properties. For a fixed order k the set V_h is the vector space of all piecewise polynomials of maximal order k:

$$V_h := \{v_h \in V | \text{ for all } T \in \mathcal{T}_h : v_h|_T \in P_k\} \ . \tag{1.4}$$

P_k denotes the space of all polynomials on $\mathbb{R}^n$ of maximal order k. The finite element discretization of the variational problem (1.2) is to find the *discrete solution* $u_h \in V_h$ with

$$< v_h, F(u_h) >= 0 \text{ for all } v_h \in V_h \ . \tag{1.5}$$

By using a suitable basis of V_h this *discrete variational problem* is equivalent to a system of linear equations. The coefficient matrix, called *stiffness matrix*, is extremely sparse, see Schwarz [57]. It can be shown that the linear system has an unique solution. For mesh size $h \to 0$ the calculated discrete solutions u_h converge to the unknown solution u. The convergence order depends on the smoothness of the solution u and the used polynomial degree k.

1.2 A-Posteriori Error Estimate

'Has the returned discrete solution u_h a sufficient accuracy ?' That is just the point for the user who has to solve the variational problem (1.2). He wishes to get an estimate of the approximation error in addition to the returned discrete solution u_h. By a so-called *a-posteriori error estimate* an approximative distribution of the true error

$$e_h := u - u_h \tag{1.6}$$

is calculated after the discrete solution u_h has been determined (here it is assumed that the discretized variational problem (1.5) is solved exactly). Naturally the additional costs to get this error estimate should be as low as possible compared to the costs for the calculation of the discrete solution u_h. The error estimate allows the user to assess the quality of the result and, if it is necessary, to start an adaptive refinement procedure to improve the FEM mesh $\mathcal{T}_h$ until a desired accuracy is obtained, see Zienkiewicz [71].

Following the assessments of the participants at the FEM'50-conference 'Fifty Years Anniversary of the Courant Element' at Jyväskyla, Finland, 1993, the a-posteriori error estimate and adaptive approaches for non–linear and non–elliptic problems are the most outstanding problems in the finite elements today, see Babuška [5].

A full a-posteriori error estimate has to consider all sources of errors in the FEM algorithms. Namely these are the following five error sources:

- *Interpolation error:* The admissible solution as well as the test functions are only selected from the space of piecewise polynomials V_h.
- *Integration error:* On a computer the residual functional F can only be approximatively evaluated by a numerical quadrature scheme.
- *Stopping error:* If the discrete variational problem (1.5) is solved iteratively (e.g. by conjugate gradient methods or in the case of a non–linear problem by the Newton-Raphson method) a stopping criterion terminates the iteration. Therefore the returned approximation is not exactly equal to discrete solution u_h.
- *Domain representation error:* In general the triangulation $\mathcal{T}_h$ is not an exact representation of the domain Ω, especially if its boundary is curved.

- *Dirichlet condition interpolation error:* Instead of a Neumann boundary condition a Dirichlet boundary condition '$u|_{\Gamma_D} = u_D$' may be prescribed on a boundary portion $?_D \subset \partial\Omega$. The Dirichlet boundary condition is replaced by an interpolation condition for piecewise polynomials.

In practice the triangulation $\mathcal{T}_h$ and the data for the interpolation of the Dirichlet condition are produced by a mesh generator, e.g. by I–DEAS [44]. Therefore the actual error of the domain representation and the error of the interpolation of the Dirichlet conditions are unknown for a finite element solver. As there is this lack in the input data their influence cannot be covered in the a–posteriori error estimate. This is the reason why these errors are not considered explicitly in the following discussions although their influence on the quality of the solution approximation can be significant.

An a–posteriori error estimate $\eta_h \in V$ is called *equivalent to the true error* e_h defined by equation (1.6), if there is a value $Q > 0$, called an *effectivity index*, with

$$\frac{1}{Q} \leq \liminf_{h \to 0+} \theta_h \leq \limsup_{h \to 0+} \theta_h \leq Q \tag{1.7}$$

where it is

$$\theta_h := \frac{\|\eta_h\|}{\|e_h\|} \,. \tag{1.8}$$

This notation was introduced by Babuška [6]. Actually the condition (1.7) expresses that the true error e_h and the error estimate η_h have exactly the same convergence order if the mesh size h decreases to zero. The quality of the error estimate depends on the value of the effectivity index Q.

Naturally the used norm has a fundamental influence on the effectivity index. By using the problem depending *energy norm* instead of the H^1-norm some authors prove that inequality (1.7) holds for their error estimate even with $Q = 1$. Such a–posteriori error estimates are called *asymptotically exact* as they represent the exact error level for $h \to 0$. However, the use of an energy norm is questionable since a conclusion from the energy norm to the H^1-norm can be very risky even if the condition number of the problem is very large. The situation becomes much more complicated if non–linear problems are considered as there is no canonical energy norm. Some concepts regarding energy norms for non–linear problems are presented by Bank [14, 15].

The a–posteriori error estimates currently used can be subdivided into three classes, see Babuška [13], Verfuerth [62, 63, 64] and Zhu [67]:

- The methods in the first class are called *averaging methods*, see Zienkiewcz [70, 72, 73], Ainsworth [3], Duran [33]. They are probably the most popular error estimates for FEMs in the engineering sciences. The basic idea is the construction of a higher order approximation Gu_h of the gradient ∇u_h. Essentially the error is estimated by $Gu_h - \nabla u_h$. In general the reliability and robustness of the estimate depends on the FEM mesh, see Babuška [11].

- The second class contains the *interpolation error estimates*, see Demkowicz [30] and Johnson [45]. In general these estimates do not work very reliably and give poor results.

- The estimates in the third class are called *residual error estimates* since they essentially base on the evaluation of the residual operator F for the calculated discrete solution u_h.

Since the techniques of the residual estimate are those, which are the most flexible and stable, more details are presented.

The most famous residual error estimate has been introduced by Babuška-Miller in 1978, see Babuška [10]. It gives an estimate of the discretization error on every element. It bases on the weighted sum of the residual in the strong formulation of the PDE (1.1) and the jumps of the derivatives of the discrete solution u_h over the element boundaries. The estimate is very easy and inexpensive. Its crucial point is setting of the values for the weights adapted to the problem. Moreover the evaluation of the strong formulation for the discrete solution u_h can be very complicated if only the weak formulation is given, the problem is non–linear or higher order polynomials are used. There are a lot of publications on the Babuška-Miller error estimate, see e.g. Bornemann [18], Verfuerth [62, 63], Kunert [46].

A more flexible approach than the Babuška-Miller error estimate is the solution of the *error equation*

$$\int_\Omega \Big(a(\nabla v)(\nabla e_h) + bve_h \Big) dx = - < v, F(u_h) > \quad \text{for all } v \in V \tag{1.9}$$

for the sought error e_h. The error equation is set up by inserting the exact solution $u = e_h + u_h$ obtained from equation (1.6) into the variational problem (1.2). By solving the error equation (1.9) the error on the level of

equation expressed by the residual $F(u_h)$ is shifted to the level of the solution represented by the error $e_h \in V$. Unfortunately the error equation (1.9) is a variational problem in the space V like the original problem (1.2). Therefore the error equation (1.9) has to be solved approximatively, too.

Regarding the evaluation of the residual function $F(u_h)$ two types can be distinguished:

- The first type is called *strong residual estimate*, since it goes back to the strong formulation (1.1) of the underlying boundary value problem when building up the right hand side of the error equation. On every element two residuals occur in the right hand side: One is the residual from the evaluation of the PDE at the interior of the element. The other residual is the jump of the normal derivative of the discrete solution u_h on the contact faces to the neighboring elements, see Babuška [9, 10], Bank [16], Verfuerth [61].
- The second type is called *weak residual estimate* since these estimates evaluate the residual for the calculated discrete solution u_h in the weak formulation (1.2), see Zienkiewicz [69], Liu [47], Bank [15], Deufelhard [31].

The approximative solution of the error equation (1.9) has to be as inexpensive as possible. Since the mounting and solution of many small, independent variational problems (namely one small system for every element or node) seems to be more inexpensive than the solution of one large problem, the use of *domain decomposition methods* is appropriate. A very popular method is the *localization*, i.e. the approximative solution of the error equation in the neighborhood of elements or nodes, e.g. by using a suitable subspace of the space V. Typically special polynomials of higher order than for the discrete solution u_h are used for the construction of such subspaces. Babuška [10] suggested to solve a local Dirichlet problem on the neighboring elements of every node. Bank [16] inspected local Neumann problems on every element (for the Stokes equations see Verfuerth [61]).

Another concept is the application of the *hierarchical FEM*, see Yserentant [66]. Here the error equation is solved in a larger space

$$V_{h+} := V_h \oplus V_h^c \tag{1.10}$$

where V_h^c is spanned by refining elements or by higher order polynomials. If the solution is smooth enough a better approximation than the discrete solution u_h can be calculated from the larger space V_{h+}. Typically hierarchical

bases are used to construct the space V_h^c, see Zienkiewicz [69]. Expecting that there are only little changes in the components in the space V_h the error equation (1.9) is only solved in the space V_h^c instead of the total space V_{h+}, see Bank [15]. In some cases the computational costs can be reduced by approximating the stiffness matrix by a diagonal matrix, see Deufelhard [31], or by localization which can be done by using bubble-shaped basis functions over the elements, see Liu [47]. Actually there is a relationship between the localization and hierarchical methods, see Bornemann [18], Verfuerth [63].

All these error estimates are designed for special variational problems or PDEs and mostly the analysis is only made for the special model problem, e.g. for the Neumann boundary value problem (1.1). The application to a specific problem requires additional development effort by the user especially as it has to be ensured that the error estimate is well-defined by the discrete error equation. By way of contrast a program package like VECFEM [38] which is designed to be applied to a large class of variational problems needs a more general a–posteriori error estimate concept. It must suit to a wide range of applications, even if there is another, better error estimate for a specific application. Especially such an a–posteriori error estimate concept has to consider non–linear variational problems which are typical for non–standard FEM applications. In addition it should be embedded into the solution procedure of the non–linear, discrete variational problem, see Schoenauer [56, 55] for finite difference methods.

1.3 Outline

In the following a new a-posteriori error estimate is presented. This estimate can be applied to a large class of non–linear variational problems without any problem specific modifications. It meets the essential requirements of an a–posteriori error estimate for a general purpose program package like VECFEM. The class of applications includes the variational problems in the heat transfer analysis (e.g. the model problem (1.1)), structural analysis and fluid dynamics.

The new a–posteriori error estimate uses the idea of the hierarchical error estimate concept, though the error equation is solved in the original approximation space V_h. Therefore this new error estimate is called *projecting error estimate*. Since the error estimate is computed in the space V_h it is ensured that the error estimate is always well-defined. Moreover the stiffness matrix of the calculation for the discrete solution u_h can be reused, which saves the

mounting of a new stiffness matrix and, if a direct solution method is used, the calculating of a new LU-decomposition. Only a new right hand side has to be mounted.

This thesis has three parts: The second chapter reflects on so–called *well–posed* non–linear variational problems on a Banach space, see Definition 1. The discussion considers that on a computer the linear functional F can only be evaluated approximatively (e.g. by numerical integration) and that the discrete variational problem has to be solved iteratively (e.g. by the Newton-Raphson method). In Theorem 2 the non–linear version of the famous Lemma of Strang [59] gives an estimate of the error e_h arising from the interpolation, integration and stopping error. Basing on the extension V_{h+} of the original approximation space V_h (see equation (1.10)) a class of a–posteriori error estimates is introduced. They consider the relevant error sources mentioned above. In Theorem 4 a criterion is established when the investigated error estimates are equivalent to the true error in the sense of inequality (1.7). Bounds for the effectivity index are given. From this very general framework the new a–posteriori error estimate technique, called *projecting error estimate*, is derived and its relationship to hierarchical error estimates is discussed. The second important result of the second chapter is the introduction of an optimal stopping criterion for the iterative solver of the discrete variational problem. Theorem 3 shows that the criterion is optimal in the sense that the solution approximations calculated with this stopping criterion converge to the unknown solution u with the same convergence order as the exact discrete solution u_h of the discrete variational problem (1.5).

The third chapter demonstrates how to apply the projecting error estimate to the FEM for the non–linear version of the introduced model problem (1.1). The space V_h^c in extension (1.10) is constructed by higher order polynomials. The propositions of the general framework are verified. The well–known analysis of Ciarlet [24] for the FEM on linear problems is quoted and modified for non–linear problems and the projecting error estimate. In Theorem 14 it is shown that the projecting error estimate for the FEM is equivalent to the true error in the sense of inequality (1.7) if the sought solution is smooth enough. The validity of the results for other variational problems than the model problem is discussed.

In the fourth chapter the practical behavior of the projecting error estimate is investigated for some applications. For the tests a modification of the VECFEM program package [38] is used. At first a special case of the model problem is presented to confirm the estimate for the effectivity index

which has been given in the third chapter of the thesis. A second example demonstrates the behavior if the sought solution has a singularity. The third example is an application from the structural analysis. The last example shows the use of the projecting error estimate for mixed FEM problems by solving the Navier-Stokes equations.

Chapter 2

The Abstract Variational Problem

For a function $f : V \to W$ and any $K \subset V$ it is set

$$f[K] := \{f(v) | v \in K\}. \tag{2.1}$$

The function $f|_K : K \to W$ defined by

$$f|_K(v) := f(v) \text{ for all } v \in K \tag{2.2}$$

denotes the *restriction of f to K*. For a second function $g : X \to Y$ with $f[V] \subset X$ the function $g \circ f : V \to Y$ defined by

$$g \circ f(v) := g(f(v)) \text{ for all } v \in V \tag{2.3}$$

denotes the *chain of f and g*. The mapping $I_V : V \to V$ defined by

$$I_V(v) := v \text{ for all } v \in V \tag{2.4}$$

denotes the *identity operator on V*. Mostly the index V will be omitted.

For functions $f, g : V \to \mathbb{R}$ the following convention is used when suprema and infima of ratios are computed:

$$\begin{aligned} \sup_{u \in V} \frac{f(u)}{g(u)} &:= \sup_{u \in V; g(u) \neq 0} \frac{f(u)}{g(u)} \\ \inf_{u \in V} \frac{f(u)}{g(u)} &:= \inf_{u \in V; g(u) \neq 0} \frac{f(u)}{g(u)} . \end{aligned} \tag{2.5}$$

Let be $(V, \|.\|_V)$ and $(W, \|.\|_W)$ Banach spaces. The vector space of all linear and continuous operators $L : V \to W$ defined by $u \to Lu$ is denoted by $\mathcal{L}(V, W)$. With the norm

$$\|L\|_{\mathcal{L}(V,W)} := \sup_{v \in V} \frac{\|Lv\|_W}{\|v\|_V} \tag{2.6}$$

the vector space $\mathcal{L}(V, W)$ is a Banach space. If $L \in \mathcal{L}(V, W)$ fulfills the following three conditions

$$\begin{array}{ll} 1. & Lv \neq 0 \text{ for all } 0 \neq v \in V \\ 2. & W = L[V] \\ 3. & L^{-1} \in \mathcal{L}(W, V) \end{array} \tag{2.7}$$

where $L^{-1} : W \to V$ is defined by $L^{-1} \circ L = I_V$, then the operator L is called an *isomorphism*. It is

$$\frac{1}{\|L^{-1}\|_{\mathcal{L}(W,V)}} = \inf_{v \in V} \frac{\|Lv\|_W}{\|v\|_V} \, . \tag{2.8}$$

The operator L^{-1} is called the *inverse operator* of L. The dual space V^* of V defined by

$$V^* := \mathcal{L}(V, \mathbb{R}) \tag{2.9}$$

denotes the vector space of all continuous, linear functionals on V. It is a Banach space. The duality mapping $< ., . >: V \times V^* \to \mathbb{R}$ defined by

$$< v, F > := Fv \text{ for all } v \in V \text{ and all } F \in V^* \tag{2.10}$$

gives the value of the linear functional $F \in V^*$ for the element $v \in V$. By equation (2.6) the norm of $F \in V^*$ is given by

$$\|F\|_{V^*} = \sup_{v \in V} \frac{< v, F >}{\|v\|} \, . \tag{2.11}$$

2.1 The Well–Posed Variational Problem

Let be $(V, \|.\|)$ a Banach space and $F : K \to V^*$ a fixed operator on $K \subset V$ with values in V^* defined by $u \to F(u)$. F may be non–linear. The following problem, called a *variational problem*, is investigated: find a solution $u \in K$ with

$$F(u) = 0 \, . \tag{2.12}$$

Since $F(u)$ is in the dual space of V this equation actually means to find an $u \in K$ with

$$< v, F(u) >= 0 \text{ for all } v \in V \ . \tag{2.13}$$

Problems of this type arise from the weak formulation of boundary value problems (e.g. see the introduced model problem (1.2), Chapter 3 of this thesis, Quarteroni [52]), minimizing problems and saddle–point problems (e.g. see Brezzi [21]). As pointed out in the introduced model problem (1.2) the evaluation of the term $< v, F(u) >$ can require the calculation of integrals, see equation (1.3).

If F is an affine operator on $K := V$, i.e. there is an linear operator $L \in \mathcal{L}(V, V^*)$ and $f \in V^*$ with

$$F(u) = Lu - f \ , \tag{2.14}$$

the variational problem (2.12) is a *linear problem*. The equation (2.13) can be written as

$$< v, Lu >=< v, f > \text{ for all } v \in V \ . \tag{2.15}$$

The functional f is called the *right hand side* of the linear variational problem (2.12). If the linear operator L is an isomorphism the variational problem (2.15) has the unique solution $u = L^{-1}f$. From the definition (2.6) of $\|L\|_{\mathcal{L}(V,V^*)}$ and equation (2.8) for the calculation of $\|L^{-1}\|_{\mathcal{L}(V^*,V)}$ the operator F fulfills the following condition for all $u_1, u_2 \in V$:

$$\begin{aligned} \frac{1}{\|L^{-1}\|_{\mathcal{L}(V^*,V)}}\|u_1 - u_2\| &\overset{(2.8)}{\leq} \|F(u_1) - F(u_2)\|_{V^*} \\ &\overset{(2.14)}{=} \|L[u_1 - u_2]\|_{V^*} \\ &\overset{(2.6)}{\leq} \|L\|_{\mathcal{L}(V,V^*)}\|u_1 - u_2\| \ . \end{aligned} \tag{2.16}$$

Therefore an estimate of the following type holds for all $u_1, u_2 \in K$:

$$D_{min}\|u_1 - u_2\| \leq \|F(u_1) - F(u_2)\|_{V^*} \leq D_{max}\|u_1 - u_2\| \ , \tag{2.17}$$

where D_{min} and D_{max} are real positive constants with

$$\begin{aligned} D_{min} &:= \frac{1}{\|L^{-1}\|_{\mathcal{L}(V^*,V)}} \\ D_{max} &:= \|L\|_{\mathcal{L}(V,V^*)} \ . \end{aligned} \tag{2.18}$$

The right estimate in inequality (2.17) ensures that a small perturbation of u_1 by $u_1 - u_2$ effects a small change on the image $F(u_1)$. Moreover the left

estimate in inequality (2.17) expresses that $F(u_1)$ and $F(u_2)$ with a small distance are produced by u_1 and u_2 with a small distance. This type of problems are called *well–posed*. As this property is essential in the discussions of this thesis it is noticed in the following definition:

Definition 1 *The operator $F : K \to V^*$ is called well–posed with condition number D^2 if for all $u_1, u_2 \in K$:*

$$\frac{1}{D}\|u_1 - u_2\| \leq \|F(u_1) - F(u_2)\|_{V^*} \leq D\|u_1 - u_2\| \ . \tag{2.19}$$

Remark 1: The condition number in the sense of Definition 1 is not unique.

Remark 2: If $F : K \to V^*$ is well–posed with condition number D^2 and $f \in V^*$ then $F_f : K \to V^*$ defined by

$$< v, F_f(u) > := < v, F(u) > + < v, f > \ \text{ for all } u \in K, v \in V \tag{2.20}$$

is well–posed with condition number D^2. f is called an *additional load*. In this sense the linear functional f occurring in an affine operator defined by equation (2.14) is an additional load.

If condition (2.17) holds the operator F is well–posed with condition number $\max(D_{min}^{-2}, D_{max}^2)$. Especially the affine operator F defined by equation (2.14) is well–posed with condition number

$$\max(\|L^{-1}\|_{\mathcal{L}(V^*,V)}, \|L\|_{\mathcal{L}(V,V^*)})^2 \ . \tag{2.21}$$

The feature 'well–posed' ensures that the solution of the variational problem (2.12) is unique in K. However, it is not guaranteed that a solution exists. The following theorem gives an easy criterion for the existence of a solution, if V is a Hilbert space. It is quoted from the theory of monotone operators, see Brezis [20]:

Theorem 1 (Brezis, 1973) *Let V be a Hilbert space, $F : V \to V^*$ and $D > 0$ a constant with*

$$\|F(u_1) - F(u_2)\|_{V^*} \leq D\|u_1 - u_2\| \tag{2.22}$$

and

$$\frac{1}{D}\|u_1 - u_2\|^2 \leq < u_1 - u_2, F(u_1) - F(u_2) > \tag{2.23}$$

for all $u_1, u_2 \in V$. Then the operator F is well–posed with condition number D^2 and the variational problem (2.12) has exactly one solution $u \in V$.

Proof : see Brezis [20]:

By the Riesz representation theorem it is $V = V^*$ and

$$< v, v >= \|v\|^2 \text{ for all } v \in V \ , \tag{2.24}$$

see Heuser [43]. From the conditions (2.22) and (2.23) it is obvious, that F is well–posed with condition number D^2. To show the existence of a solution the operator $\Phi : V \to V$ is defined by

$$\Phi(u) = u - \frac{1}{D^3} F(u) \tag{2.25}$$

for all $u \in V$. It is shown that Φ is a contracting operator on V:

For all $u_1, u_2 \in V$ it is

$$\begin{aligned}
& \|\Phi(u_1) - \Phi(u_2)\|^2 \\
\overset{(2.25)+(2.24)}{=} \quad & < u_1 - u_2 - \frac{1}{D^3}(F(u_1) - F(u_2)), \\
& \qquad\qquad u_1 - u_2 - \frac{1}{D^3}(F(u_1) - F(u_2)) > \\
= \quad & \|u_1 - u_2\|^2 - \frac{2}{D^3} < u_1 - u_2, F(u_1) - F(u_2) > + \\
& \qquad\qquad \frac{1}{D^6}\|F(u_1) - F(u_2)\|^2 \\
\overset{(2.22)+(2.23)}{\leq} \quad & (1 - \frac{2}{D^3}\frac{1}{D} + \frac{D^2}{D^6})\|u_1 - u_2\|^2 \\
= \quad & (1 - \frac{1}{D^4})\|u_1 - u_2\|^2 \ .
\end{aligned} \tag{2.26}$$

By Banach's fixed point theorem, see Heuser [43], the contracting operator Φ has a fixed point u:

$$u = \Phi(u) = u - \frac{1}{D^3} F(u) \ . \tag{2.27}$$

Therefore it is $F(u) = 0$, thus u is a solution of the variational problem (2.12). This proves the theorem •

Remark 1: In the framework of monotone operators the operator F is called *strictly monotone* if condition (2.23) holds.

Remark 2: If the operator F is an affine operator defined by equation (2.14) the condition (2.23) is equivalent to the condition

$$\frac{1}{D}\|v\|^2 \leq < v, Lv > \ \text{ for all } v \in V \ . \tag{2.28}$$

If $L \in \mathcal{L}(V, V^*)$ fulfills this condition the linear operator L is called *V-elliptic* (or *coercive*). From the well–known Lax–Milgram–Lemma which is the linear version of Theorem 1 a V-elliptic linear operator $L \in \mathcal{L}(V, V^*)$ is an isomorphism, see Brezzi [21]. Especially the problem (2.15) has an unique solution for all right hand sides $f \in V^*$.

2.2 The Discretization

The variational problem (2.12) cannot be solved with a computer but an approximation of the exact solution can be calculated by discretization:

Let V_h be a finite dimensional subspace of the space V and K_h a subset of the set $V_h \cap K$. The index h is interpreted as a real number which refers to a mesh size, see Section 3.4. The space V_h is spanned by a suitable basis $\varphi^h = \{\varphi_i^h\}_{i=1,d^h} \subset V_h$ e.g. in the finite element method by using a nodal basis, see Zienkiewicz [68].

The original problem (2.12) is now solved in in the finite dimensional space V_h instead of the total space V: find a discrete solution $u_h \in K_h$ with

$$< v_h, F(u_h) >= 0 \text{ for all } v_h \in V_h \ . \tag{2.29}$$

In general a computer cannot exactly evaluate the real value $< v_h, F(u_h) >$ as numerical integration has to be used to calculate the involved integrals, see Section 3.5. Therefore it has to be assumed that only an approximation $F_h : K_h \to V_h^*$ of the operator F is known. Keep in mind that the discrete operator F_h does not have to be defined on the space V and and the set K. Actually the following *discrete variational problem* is solved for the sought discrete solution $u_h \in K_h$:

$$F_h(u_h) = 0 \ . \tag{2.30}$$

As $F_h(u_h) \in V_h^*$ the discrete variational problem means to find $u_h \in K_h$ with

$$< v_h, F_h(u_h) >= 0 \text{ for all } v_h \in V_h \ . \tag{2.31}$$

Theorem 1 applied to the discrete operator F_h in the space V_h gives a criterion for the existence of the discrete solution u_h.

To solve the discrete variational problem (2.30) on a computer the discrete solution u_h is represented in the basis φ^h by

$$u_h = \sum_{i=1}^{d^h} u_{h,i} \varphi_i^h \tag{2.32}$$

where $(u_{h,i})_{i=1,d^h} \in \mathbb{R}^{d^h}$. As every element in V_h can be represented by the basis φ^h problem (2.31) is equivalent to find a vector $(u_{h,i})_{i=1,d^h} \in \mathbb{R}^{d^h}$ with

$$< \varphi_j^h, F_h(\sum_{i=1}^{d^h} u_{h,i}\varphi_i^h) >= 0 \text{ for all } j = 1, \ldots, d^h \ . \tag{2.33}$$

This is a system of d_h non–linear equations for the d_h coefficients $(u_{h,i})_{i=1,d^h}$ in the representation (2.32) of the sought discrete solution u_h.

Starting from an initial guess $u_h^{(0)} \in K_h$ a sequence of approximations $(u_h^{(k)})_{k\in\mathbb{N}}$ is calculated by using their representations by the selected basis φ^h:

$$u_h^{(k)} = \sum_{i=1}^{d^h} u_{h,i}^{(k)}\varphi_i^h \tag{2.34}$$

with $(u_{h,i}^{(k)})_{i=1,d^h} \in \mathbb{R}^{d^h}$ for all $k \in \mathbb{N}_0$. The difference $u_h^{(k)} - u_h^{(k-1)}$ of two sequential approximations, called the k-th *correction*, is given by

$$u_h^{(k)} - u_h^{(k-1)} = \sum_{i=1}^{d^h}(u_{h,i}^{(k)} - u_{h,i}^{(k-1)})\varphi_i^h = \sum_{i=1}^{d^h} \Delta u_i^{(k)}\varphi_i^h \tag{2.35}$$

where $(\Delta u_i^{(k)})_{i=1,d^h} \in \mathbb{R}^{d^h}$ for all $k \in \mathbb{N}$. The correction is calculated from the following system of d_h linear equation:

$$\begin{aligned} \sum_{i=1}^{d^h} < \varphi_j^h, L_h^{(k-1)}\varphi_i^h > \Delta u_i^{(k)} = - < \varphi_j^h, F_h(u_h^{(k-1)}) > \\ \text{for all } j = 1, \ldots, d^h \end{aligned} \tag{2.36}$$

where $L_h^{(k-1)} \in \mathcal{L}(V_h, V_h^*)$ is a suitable isomorphism. If the discrete operator F_h is smooth, the isomorphism $L_h^{(k-1)}$ can be set to the derivative DF_h of the discrete operator F_h at the $(k-1)$–th approximation $u_h^{(k-1)}$. Then the iteration (2.36) corresponds with the *Newton–Raphson method* which is a very efficient method for solving non–linear equations, see Stoer [58]. If V is a Hilbert space, one can set $L_h^{(k-1)} = \gamma I_{V_h}$ with a suitable real value $\gamma \in \mathbb{R}$. That is the *method of successive approximation*. The proof of Theorem 1 shows that $\gamma := \frac{1}{D_h^3}$ ensures convergence if D_h^2 denotes the condition number of F_h.

However, the $d^h \times d^h$ coefficient matrix

$$< \varphi^h, L_h^{(k-1)}\varphi^h >:= (< \varphi_j^h, L_h^{(k-1)}\varphi_i^h >)_{j,i=1,d^h} \tag{2.37}$$

which is called the *stiffness matrix* and the right hand side vector

$$< \varphi^h, F_h(u_h^{(k-1)}) >:= (< \varphi_j^h, F_h(u_h^{(k-1)}) >)_{j=1,d^h} \ , \tag{2.38}$$

called the *iteration defect*, have to be assembled. The iteration defect has to be evaluated for the current $(k-1)$–th approximation $u_h^{(k-1)}$ in every iteration step using the basis representation (2.34) for the approximation $u_h^{(k-1)}$. If the isomorphism $L_h^{(k-1)}$ (e.g. when using the modified Newton–Raphson method or the method of successive approximation) is not changed during the iteration the stiffness matrix has to be assembled only at the beginning but in general a new stiffness matrix has to be assembled in every iteration step.

If the representation (2.35) of the correction $u_h^{(k)} - u_h^{(k-1)}$ is used and it is considered that every element in the space V_h can be represented by the basis φ^h it turns out that the linear system (2.36) is equivalent to the equation

$$< v_h, L_h^{(k-1)}[u_h^{(k)} - u_h^{(k-1)}] >= - < v_h, F_h(u_h^{(k-1)}) > \ \text{ for all } v_h \in V_h \ . \tag{2.39}$$

Since $L_h^{(k-1)}[u_h^{(k)} - u_h^{(k-1)}]$ and $F_h(u_h^{(k-1)})$ are in the dual space V_h^* that means

$$L_h^{(k-1)}[u_h^{(k)} - u_h^{(k-1)}] = -F_h(u_h^{(k-1)}) \ . \tag{2.40}$$

Therefore the iteration procedure (2.36) to solve the discrete variational problem (2.30) can be written down as

$$u_h^{(k)} := u_h^{(k-1)} - (L_h^{(k-1)})^{-1} F_h(u_h^{(k-1)}) \tag{2.41}$$

for all $k \in \mathbb{N}$ where $u_h^{(0)} \in K_h$ is an initial guess and for all $k \in \mathbb{N}$ the linear operators $L_h^{(k-1)}$ are suitable isomorphisms.

The iteration procedure (2.41) is terminated by a suitable stopping criterion. Therefore the discrete variational problem (2.30) is not exactly solved, but an approximation $\hat{u}_h \in K_h$ of discrete solution u_h is computed. Especially the iteration defect $F_h(\hat{u}_h)$ is not equal to zero. What is the quality of the calculated approximation $\hat{u}_h$ compared to the sought solution u ? The well–known Lemma of Strang [59] gives an estimate for a linear variational problem. The following theorem is the non–linear version of this lemma:

Theorem 2 (Grosz) *Let be $F : K \to V^*$ well–posed with condition number D^2, $u \in K$ with $F(u) = 0$, $V_h \subset V$, $K_h \subset K \cap V_h$ and $F_h : K_h \to V_h^*$ well–posed with condition number D_h^2. Then for all $\hat{u}_h \in K_h$ and $v_h \in K_h$ the following inequality holds:*

$$\begin{aligned} \|u - \hat{u}_h\| &\leq D_h \|F_h(\hat{u}_h)\|_{V_h^*} \\ &+ (1 + DD_h)\|v_h - u\| \\ &+ D_h \|F_h(v_h) - F(v_h)\|_{V_h^*} . \end{aligned} \tag{2.42}$$

Proof: Let be $\hat{u}_h, v_h \in K_h \subset K$. Using the fact that the discrete operator F_h is well–posed it is:

$$\begin{aligned} \|\hat{u}_h - u\| &\leq \|\hat{u}_h - v_h\| + \|v_h - u\| \\ &\leq D_h \|F_h(\hat{u}_h) - F_h(v_h)\|_{V_h^*} + \|v_h - u\| \\ &\leq D_h \|F_h(\hat{u}_h)\|_{V_h^*} + D_h \|F_h(v_h)\|_{V_h^*} + \|v_h - u\| . \end{aligned} \tag{2.43}$$

Further estimates for the value $\|F_h(v_h)\|_{V_h^*}$ are obtained by the fact that $F(u) = 0$:

$$\begin{aligned} \|F_h(v_h)\|_{V_h^*} &= \|F_h(v_h) - F(u)\|_{V_h^*} \\ &\leq \|F_h(v_h) - F(v_h)\|_{V_h^*} + \|F(v_h) - F(u)\|_{V_h^*} \\ &\leq \|F_h(v_h) - F(v_h)\|_{V_h^*} + \|F(v_h) - F(u)\|_{V^*} . \end{aligned} \tag{2.44}$$

In the last estimate the fact is used that for all $G \in V^*$ it is $\|G\|_{V_h^*} \leq \|G\|_{V^*}$ as $V_h \subset V$. Since the operator F is well–posed it turns out from inequality (2.44) that

$$\|F_h(v_h)\|_{V_h^*} \leq \|F_h(v_h) - F(v_h)\|_{V_h^*} + D\|v_h - u\| . \tag{2.45}$$

After this estimate was inserted into inequality (2.43) the inequality of the theorem was proved •

The first term on the right hand side of inequality (2.42) is called the *stopping error* or synonymously the *termination error* since it considers that the discrete variational problem (2.30) is solved by an iterative method. The second term considers the error which is produced by the reduction of the space V to the finite dimensional subspace V_h. It is called the *interpolation error*. The third term considers the error from the approximation of the operator F by the discrete operator F_h by numerical integration. Therefore this term is called the *integration error*.

In the following the behavior of the discretization is analyzed if the 'mesh size h goes to zero'. This means that a family of finite dimensional subspaces $(V_h)_{h \in \mathcal{H}}$ of the space V is given where the set $\mathcal{H} \subset \mathbb{R}_+$ has the unique

accumulation point 0. To simplify the following formulations '$(V_h)_{h>0}$' is written instead of $(V_h)_{h\in\mathcal{H}}$. Moreover it is written

$$h \to 0 \tag{2.46}$$

to express that a condition holds for every sequence of mesh sizes in the index set $\mathcal{H}$ which converge to zero.

In this notation the following corollary is a direct consequence of Theorem 2 when it is assumed that the discrete problem is solved exactly. The problem of a suitable stopping criterion will be discussed below in Theorem 3.

Corollary 1 *Let $F : K \to V^*$ be well–posed with condition number D^2, u an element of the set K with $F(u) = 0$ and $(V_h)_{h>0}$ be a family of finite dimensional subspaces of the space V. For every $h > 0$ let be*

$$\mathcal{I}_h u \in K_h \subset V_h \cap K \tag{2.47}$$

with $\mathcal{I}_h u \to u$ for $h \to 0$ at least of order $p_1 > 0$, i.e. there is a constant $C_1 > 0$ with

$$\|u - \mathcal{I}_h u\| \leq C_1 h^{p_1} \text{ for all } h > 0 \,. \tag{2.48}$$

For all $h > 0$ let $F_h : K_h \to V_h^*$ be a well–posed operator on K_h with condition number D^2 and $F_h \to F$ at $\mathcal{I}_h u$ for $h \to 0$ at least of order p_2, i.e. there is a constant $C_2 > 0$ with

$$\|F_h(\mathcal{I}_h u) - F(\mathcal{I}_h u)\|_{V_h^*} \leq C_2 h^{p_2} \text{ for all } h > 0 \,. \tag{2.49}$$

If $u_h \in K_h$ with $F_h(u_h) = 0$ then $u_h \to u$ for $h \to 0$ at least of order $\min(p_1, p_2)$, i.e. there is a constant $C_3 > 0$ with

$$\|u - u_h\| \leq C_3 h^{\min(p_1,p_2)} \text{ for all } h > 0 \,. \tag{2.50}$$

Proof: The corollary is a direct conclusion from Theorem 2. Essential is the fact that the condition number $D_h = D$ does not depend on h and no termination error (i.e. $F_h(u_h) = 0$) occurs •

The assumptions of Corollary 1 state a consistency condition for the approximation space V_h. If any element $\mathcal{I}_h u$ with properties (2.48) and (2.49) is found then by Corollary 1 the solutions of the discrete variational problems in the spaces V_h converge to the sought solution. Estimation (2.48) describes the approximation properties of the sets $(K_h)_{h>0}$ for the elements in the set K. The property (2.49) shows the approximation properties of the discrete

operators $(F_h)_{h>0}$ for the operator F. In the finite element application the operator $\mathcal{I}_h$ is an interpolation operator in the space V_h.

Remark : In practice Theorem 2 as well as Corollary 1 are not suitable to give an estimate of the error $u - \hat{u}_h$. The reason is that the condition numbers of the involved operators as well as the constants C_1 and C_2 in the inequalities (2.48) and (2.49) are unknown. In the finite element application there are some ideas to estimate the values for C_1 by higher order interpolation, see Demkowicz [30] and Johnson [45], but the computed bounds are not reliable and overestimate the true error dramatically.

In the next section a very general technique is presented how the discretization error $u - \hat{u}_h$ can be estimated reliablely. Since the estimate is calculated *after* the discrete variational problem (2.30) has been solved the technique is called *a-posteriori* error estimate.

2.3 A–Posteriori Error Estimate

If an approximative solution $\hat{u}_h \in K_h$ of the discrete variational problem (2.30) is computed the true error

$$e_h := u - \hat{u}_h \tag{2.51}$$

cannot be determined since naturally the sought solution u is unknown. Yet an approximation of the true error, called an *a–posteriori error estimate*, can be computed. In Theorem 4 an error estimate (more exactly a family of error estimates) will be introduced and a criterion is proposed to check when the error estimate represents the true error well.

The main handicap to get the true e_h is that a computer cannot represent the space V. It seems to be a good idea to expand the space V_h to a space $V_{h+} \subset V$ in a suitable way and to calculate a second better discrete solution u_{h+} from this greater vector space V_{h+}, see Zienkiewicz [69], Deufelhard [31], Bank [15], Bornemann [18]. If the approximation u_{h+} is actually a better approximation of the exact solution u then one can expect that

$$\eta_{h+} := u_{h+} - \hat{u}_h \tag{2.52}$$

is a good a–posteriori error estimate. More exactly this works as described in the following:

Let $V_{h+} \subset V$ be a finite dimensional subspace of the space V with $V_h \subset V_{h+}$ and let $F_{h+} : K_{h+} \to V_{h+}^*$ be an operator on K_{h+} with $K_h \subset K_{h+} \subset V_{h+} \cap K$.

$u_{h+} \in K_{h+}$ denotes the solution of the discrete variational problem

$$F_{h+}(u_{h+}) = 0 \ . \tag{2.53}$$

The situation that the discrete solution u_{h+} is actually a better approximation of the solution u than the discrete solution u_h is characterized by the following definition where the elements u_h, u_{h+} and u are not necessaryly the solutions of variational problems, see Bank [15].

Definition 2 (Bank 1993) *Let be $u \in V$ and for all $h > 0$ $u_h, u_{h+} \in V_h$ and $r_h \geq 0$ with*

$$\|u - u_{h+}\| \leq r_h \|u - u_h\| \ . \tag{2.54}$$

Then $(u_h, u_{h+})_{h>0}$ is called saturated for the element u if there is a constant $r_0 \in \mathbb{R}$ with

$$0 \leq \limsup_{h \to 0} r_h \leq r_0 < 1 \ . \tag{2.55}$$

The value r_0 is called a saturation bound.

Essential in this definition is the condition $r_0 < 1$ which ensures that at least for a small mesh size h the discrete solution u_{h+} gives a better approximation of the solution u than the discrete solution u_h. By using the notations of Corollary 1 the set of pairs $(u_h, u_{h+})_{h>0}$ is saturated for the solution u with saturation bound $r_0 = 0$ if for $h \to 0$ the discrete solutions u_h converge to the solution u with maximal order p and the approximations u_{h+} converge to the solution u at least of order q with $q > p$.

Before the a–posteriori error estimate is investigated the question of an optimal stopping criterion for the iterative solver of the discrete variational problem (2.30) is answered: For a given approximation $\hat{u}_h$ of the discrete solution u_h the norm of the discrete operator $F_{h+}(\hat{u}_h)$ can be evaluated to involve it into a stopping criterion. Using heuristical arguments the stopping criterion given in the following theorem has been introduced by Schoenauer [56, 55] for the finite difference method and by Grosz [39] for finite element methods. The idea is to stop the iteration if the stopping error is in the order of the (estimated) discretization error. Actually the stopping criterion produces approximations $\hat{u}_h$ that have the same convergence order to the sought solution u like the exact discrete solutions u_h.

Theorem 3 (Grosz) *Let be $V_h \subset V_{h+} \subset V$, $K_h \subset K_{h+} \cap V_h$, $K_{h+} \subset V_{h+} \cap K$, $F_{h+} : K_{h+} \to V_{h+}^*$ and $F_h : K_h \to V_h^*$ well–posed operators with condition number D^2, $u_h \in K_h$ with $F_h(u_h) = 0$ and $u_{h+} \in K_{h+}$ with $F_{h+}(u_{h+}) = 0$. If*

$(u_h, u_{h+})_{h>0}$ is saturated for the element $u \in V$ with saturation bound r_0 and for all $h > 0$ $\hat{u}_h \in K_h$ fulfills the stopping criterion:

$$\|F_h(\hat{u}_h)\|_{V_h^*} \leq \lambda \|F_{h+}(\hat{u}_h)\|_{V_{h+}^*} \tag{2.56}$$

with fixed $0 \leq \lambda < \lambda_0 := \frac{1}{D^2} \min(\frac{1-r_0}{2r_0}, 1)$, then it is

$$\limsup_{h \to 0} \frac{\|u - \hat{u}_h\|}{\|u - u_h\|} \leq \frac{1 + r_0 D^2 \lambda}{1 - D^2 \lambda} \; . \tag{2.57}$$

Especially the approximations u_h and $\hat{u}_h$ have the same convergence order to the element u for $h \to 0$. In addition $(\hat{u}_h, u_{h+})_{h>0}$ is also saturated for the element u with saturation bound

$$\hat{r}_0 := r_0 \frac{1 + D^2 \lambda}{1 - r_0 D^2 \lambda} < 1 \; . \tag{2.58}$$

Proof: First estimation (2.57) is proved: Since the discrete operator F_h is well–posed with condition number D^2 and $F_h(u_h) = 0$ it is

$$\begin{array}{rcl} \|\hat{u}_h - u_h\| & \leq & D \|F_h(\hat{u}_h) - F_h(u_h)\|_{V_h^*} \\ & = & D \|F_h(\hat{u}_h)\|_{V_h^*} \; . \end{array} \tag{2.59}$$

By inserting the stopping criterion (2.56) and using $F_{h+}(u_{h+}) = 0$ one gets

$$\begin{array}{rcl} \|\hat{u}_h - u_h\| & \underset{(2.56)}{\leq} & D \|F_h(\hat{u}_h)\|_{V_h^*} \\ & \leq & D\lambda \|F_{h+}(\hat{u}_h)\|_{V_{h+}^*} \\ & = & D\lambda \|F_{h+}(\hat{u}_h) - F_{h+}(u_{h+})\|_{V_{h+}^*} \\ & \leq & D^2 \lambda \|\hat{u}_h - u_{h+}\| \; . \end{array} \tag{2.60}$$

In the last estimation the fact is used that the discrete operator F_{h+} is well–posed with condition number D^2. The approximation u_h is introduced into the right hand side by using the triangle inequality:

$$\|\hat{u}_h - u_h\| \leq D^2 \lambda (\|\hat{u}_h - u_h\| + \|u_h - u_{h+}\|) \; . \tag{2.61}$$

As it is $D^2 \lambda < 1$ this inequality is solved for $\|\hat{u}_h - u_h\|$:

$$\|\hat{u}_h - u_h\| \leq \frac{D^2 \lambda}{1 - D^2 \lambda} \|u_h - u_{h+}\| \; . \tag{2.62}$$

After the element u was put into the right hand side the definition of the factor r_h by inequality (2.54) in Definition 2 is inserted:

$$\begin{array}{rcl} \|\hat{u}_h - u_h\| & \leq & \frac{D^2 \lambda}{1 - D^2 \lambda} (\|u_h - u\| + \|u - u_{h+}\|) \\ & \underset{(2.54)}{\leq} & D^2 \lambda \frac{1 + r_h}{1 - D^2 \lambda} \|u_h - u\| \; . \end{array} \tag{2.63}$$

Since it is $\limsup_{h\to 0} r_h \leq r_0$ inequality (2.57) is verified.

To show that the set $(\hat{u}_h, u_{h+})_{h>0}$ is saturated for the element u an estimation of type (2.54) with the approximation $\hat{u}_h$ instead of the approximation u_h and an appropriate factor r_h (denoted by $\hat{r}_h$) has to be established: Starting from $\|u_h - u\|$ inequality (2.60) is used to obtain:

$$\begin{array}{rcl} \|u_h - u\| & \underset{(2.60)}{\leq} & \|u_h - \hat{u}_h\| + \|\hat{u}_h - u\| \\ & \leq & D^2\lambda\|\hat{u}_h - u_{h+}\| + \|\hat{u}_h - u\| \ . \end{array} \tag{2.64}$$

By inserting the element u into the first term of the right hand side it is

$$\begin{array}{rcl} \|u_h - u\| & \underset{(2.54)}{\leq} & D^2\lambda(\|\hat{u}_h - u\| + \|u - u_{h+}\|) + \|\hat{u}_h - u\| \\ & \leq & (1 + D^2\lambda)\|\hat{u}_h - u\| + r_h D^2\lambda\|u_h - u\| \ . \end{array} \tag{2.65}$$

As it is $r_h D^2\lambda < 1$ at least for a small mesh size h this can be solved for $\|u_h - u\|$ to get the estimation:

$$\|u_h - u\| \leq \frac{1 + D^2\lambda}{1 - r_h D^2\lambda}\|\hat{u}_h - u\| \ . \tag{2.66}$$

This inequality is inserted into the definition (2.54) of the factor r_h and it turns out that

$$\begin{array}{rcl} \|u - u_{h+}\| & \overset{(2.54)}{\leq} & r_h\|u - u_h\| \\ & \underset{(2.66)}{\leq} & r_h\dfrac{1 + D^2\lambda}{1 - r_h D^2\lambda}\|\hat{u}_h - u\| \ . \end{array} \tag{2.67}$$

Therefore it is for all small mesh sizes $h > 0$

$$\|u - u_{h+}\| \leq \hat{r}_h\|u - \hat{u}_h\| \tag{2.68}$$

with

$$\hat{r}_h := r_h\frac{1 + D^2\lambda}{1 - r_h D^2\lambda} \ . \tag{2.69}$$

If $\limsup_{h\to 0}$ of $(\hat{r}_h)_{h>0}$ is calculated it turns out that

$$\limsup_{h\to 0}\hat{r}_h \leq \hat{r}_0 := r_0\frac{1 + D^2\lambda}{1 - r_0 D^2\lambda} \tag{2.70}$$

as it is $r_0 D^2\lambda < 1$. It remains to show that $\hat{r}_0 < 1$:

As it has been assumed that $\lambda < \lambda_0$ it is

$$D^2\lambda < D^2\lambda_0 \leq \frac{1 - r_0}{2r_0} \ . \tag{2.71}$$

Therefore the following estimations hold:

$$\begin{aligned} \hat{r}_0 &= r_0 \frac{1 + D^2\lambda}{1 - r_0 D^2 \lambda} \\ &< r_0 \frac{1 + \frac{1-r_0}{2r_0}}{1 - r_0 \frac{1-r_0}{2r_0}} \\ &= \frac{2r_0 + 1 - r_0}{2 - (1 - r_0)} \\ &= 1\,. \end{aligned} \tag{2.72}$$

This proves the theorem •

The iteration procedure (2.41) for the solution of the variational problem $F_h(u_h) = 0$ should be terminated after the k-th iteration step if the condition (2.56) given in Theorem 3 holds for $\hat{u}_h := u_h^{(k)}$. To check this criterion the value $\|F_{h+}(u_h^{(k)})\|_{V_{\hat{h}+}^*}$ has to be calculated or estimated in every iteration step. In spite of the additional effort the use of the stopping criterion saves much computing time, see Example 2 and Example 4 in Chapter 4. The stopping criterion (2.56) is optimal in the sense that the returned approximations $(\hat{u}_h)_{h>0}$ have the same convergence order to the sought solution u like the exact calculated discrete solutions $(u_h)_{h>0}$. The reason is that if the stopping criterion is fulfilled for the first time during the iteration procedure the discretization error starts to dominate the termination error. It is emphasized that the factor λ can be very small as the condition number D of the discrete operators F_h and F_{h+} can be very large. However a lot of tests have shown that $\lambda = 0.075$ is a suitable selection for a large class of applications although the factor λ should be smaller when following Theorem 3.

Remark 1: The stopping criterion (2.56) can be replaced by

$$\|F_h(\hat{u}_h)\|_{V_h^*} \leq \lambda \|F_{h+}(\hat{u}_h)\|_{V_{\hat{h}+}^*} \tag{2.73}$$

where $V_{\hat{h}+}$ can be any subspace of V_{h+}. Actually this condition is stronger than the original criterion (2.56) but sometimes it is simpler and cheaper to be checked.

Remark 2: In the propositions of Theorem 3 it has been assumed that F_h and F_{h+} are well–posed with same condition number D. This is not a restriction as the condition number D can be set to $\max(D_1, D_2)$ if the discrete operator F_h is well–posed with condition number D_1 and the discrete operator F_{h+} is well–posed with condition number D_2. However, the discrete operators F_h and F_{h+} should have a condition number which is very close

to the condition number of the operator F as they are its approximation. Therefore it can be assumed that the operators F_h, F_{h+} and F are well–posed with a common condition number D (that is independent of the mesh size h!).

To assess the quality of an a–posteriori error estimate the following criterion was introduced by Babuška [6]:

Definition 3 (Babuška 1992) *Let be $(\hat{u}_h)_{h>0} \subset V$ a set of approximations of the element $u \in V$. The subset $(\eta_h)_{h>0}$ of V is called equivalent to the true error if there is a constant $Q > 0$ with*

$$\frac{1}{Q} \leq \liminf_{h \to 0+} \theta_h \leq \limsup_{h \to 0+} \theta_h \leq Q \tag{2.74}$$

where it is

$$\theta_h := \frac{\|\eta_h\|}{\|u - \hat{u}_h\|} \, . \tag{2.75}$$

The constant Q is called an effectivity index.

Remark: If the set of error estimates $(\eta_h)_{h>0}$ is equivalent to the true error this means that they have exactly the same asymptotic behavior for $h \to 0$ like the exact error $e_h = u - \hat{u}_h$. If the levels of the error estimates are correct depends on the value of the effectivity index Q. The error estimates become more fuzzy if the value of the effectivity index Q increases. In the case that one can set $Q = 1$ the error estimates $(\eta_h)_{h>0}$ represent the correct level of the true error for $h \to 0$. Such estimates are called *asymptotically exact*.

The following lemma confirms that the expansion of the space V_h is a successful approach to estimate the error of the calculated approximation. It is essential that the expansion V_{h+} is large enough which is covered by $(u_h, u_{h+})_{h>0}$ being saturated. The following lemma is important for the further discussions:

Lemma 1 (Grosz) *Let be $V_h \subset V_{h+} \subset V$, $K_h \subset K_{h+} \cap V_h$, $F_{h+} : K_{h+} \to V_{h+}^*$ and $F_h : K_h \to V_h^*$ well–posed operators with condition number D^2, $u_h \in K_h$ with $F_h(u_h) = 0$, $u_{h+} \in K_{h+}$ with $F_{h+}(u_{h+}) = 0$ for all $h > 0$ and $(u_h, u_{h+})_{h>0}$ saturated for $u \in V$. If for all $h > 0$ $\hat{u}_h \in K_h$ fulfills the stopping criterion (2.56) in Theorem 3 the set $(\eta_{h+})_{h>0}$ defined by*

$$\eta_{h+} := u_{h+} - \hat{u}_h \tag{2.76}$$

for all $h > 0$ is equivalent to the true error in the sense of Definition 3. More precisely it is

$$1 - \hat{r}_0 \leq \liminf_{h \to 0} \frac{\|\eta_{h+}\|}{\|e_h\|} \leq \limsup_{h \to 0} \frac{\|\eta_{h+}\|}{\|e_h\|} \leq 1 + \hat{r}_0 \tag{2.77}$$

where for all $h > 0$ $e_h := u - \hat{u}_h$ denotes the exact error. The constant $0 \leq \hat{r}_0 < 1$ is defined by equation (2.58) in Theorem 3.

Proof: Because of the triangle inequality it is

$$\begin{array}{rcl} \|e_h\| & \overset{(2.51)}{=} & \|u - \hat{u}_h\| \\ & \leq & \|u - u_{h+}\| + \|u_{h+} - \hat{u}_h\| \\ & \leq & \hat{r}_h \|u - \hat{u}_h\| + \|\eta_{h+}\| \\ & = & \hat{r}_h \|e_h\| + \|\eta_{h+}\| \end{array} \tag{2.78}$$

where the factor $\hat{r}_h$ is defined by equation (2.69) in the proof of Theorem 3. Moreover the following estimation holds:

$$\begin{array}{rcl} \|\eta_{h+}\| & \overset{(2.76)}{=} & \|\hat{u}_h - u_{h+}\| \\ & \leq & \|\hat{u}_h - u\| + \|u - u_{h+}\| \\ & \leq & (1 + \hat{r}_h)\|e_h\| \; . \end{array} \tag{2.79}$$

By combining both estimates (2.78) and (2.79) one gets

$$(1 - \hat{r}_h)\|e_h\| \leq \|\eta_{h+}\| \leq (1 + \hat{r}_h)\|e_h\| \tag{2.80}$$

which proves the lemma •

The lemma states that the error estimate η_{h+} defined by equation (2.76) is a reliable a–posteriori error estimate. Yet the calculation of the error estimate η_{h+} requires the solution of the non–linear, discrete variational equation (2.53) in the expansion V_{h+} to get the better discrete solution u_{h+}. Similar to the solution of the discrete variational problem (2.30) for the discrete solution u_h this has to be done by using an iterative method which is analogously to iteration procedure (2.41). Certainly $\hat{u}_h \in K_h \subset K_{h+}$ is a good initial guess for this iteration procedure. Then one iteration step will be enough to calculate an approximation $\hat{u}_{h+} \in V_{h+}$ of the better discrete solution u_{h+} with a sufficient accuracy. The approximation

$$\eta_h^I := \hat{u}_{h+} - \hat{u}_h \tag{2.81}$$

of the error estimate η_{h+} will be equivalent to the true error in the sense of Definition 2. The equation determining the error estimate η_h^I is obtained readily from the formula of the iteration procedure (2.40):

$$L_{h+}\eta_h^I = L_{h+}[\hat{u}_{h+} - \hat{u}_h] = -F_{h+}(\hat{u}_h) \; . \tag{2.82}$$

$L_{h+} \in \mathcal{L}(V_{h+}, V_{h+}^*)$ is a suitable isomorphism. This *error equation* is a linear, discrete variational problem in the expansion V_{h+}: find error estimate $\eta_h^I \in V_{h+}$ with

$$< v_{h+}, L_{h+}\eta_h^I >= - < v_{h+}, F_{h+}(\hat{u}_h) > \text{ for all } v_{h+} \in V_{h+} . \tag{2.83}$$

The calculation of the error estimate η_h^I requires the mounting of a new stiffness matrix. As in practice the dimension of the expansion V_{h+} is twice the dimension of the space V_h the dimension of this stiffness matrix is twice the dimension of the stiffness matrix used in an iteration procedure (2.41) to calculate the discrete solution u_h. Therefore the mounting of the coefficient matrix for the error equation (2.83) requires at least the fourfold computational effort. To face the question how these costs can be reduced a more general concept is introduced to calculate a–posteriori error estimates for the approximation $\hat{u}_h$ basing on an error equation of type (2.83).

Assume there is a space $V_{\bar{h}} \subset V$ where an isomorphism $L_{\bar{h}} \in \mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)$ is known. The inverse of $L_{\bar{h}}$ should be easily computable. Moreover it is assumed that there is an operator $\mathcal{J}_{h+} \in \mathcal{L}(V_{\bar{h}}, V_{h+})$ which joins every element in the space $V_{\bar{h}}$ with an element in the expansion V_{h+}. An a–posteriori error estimate $\eta_{\bar{h}} \in V_{\bar{h}}$ is defined by

$$< v_{\bar{h}}, L_{\bar{h}}\eta_{\bar{h}} >= - < \mathcal{J}_{h+}v_{\bar{h}}, F_{h+}(\hat{u}_h) > \text{ for all } v_{\bar{h}} \in V_{\bar{h}} . \tag{2.84}$$

Depending on the selection of the space $V_{\bar{h}}$ and the joining operator $\mathcal{J}_{h+}$ various error estimates are defined, see below.

The new error equation (2.84) is deduced from the error equation (2.83): When it is set $\eta_h^I := \mathcal{J}_{h+}\eta_{\bar{h}}$ and $v_{h+} := \mathcal{J}_{h+}v_{\bar{h}}$ with $\eta_{\bar{h}}, v_{\bar{h}} \in V_{\bar{h}}$ the error equation (2.83) is transformed to

$$< \mathcal{J}_{h+}v_{\bar{h}}, L_{h+}\mathcal{J}_{h+}\eta_{\bar{h}} >= - < \mathcal{J}_{h+}v_{\bar{h}}, F_{h+}(\hat{u}_h) > \text{ for all } v_{\bar{h}} \in V_{\bar{h}} . \tag{2.85}$$

A new linear operator $L_{\bar{h}} \in \mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)$ defined by

$$< v_{\bar{h}}, L_{\bar{h}}w_{\bar{h}} >=< \mathcal{J}_{h+}v_{\bar{h}}, L_{h+}\mathcal{J}_{h+}w_{\bar{h}} > \text{ for all } v_{\bar{h}}, w_{\bar{h}} \in V_{\bar{h}} \tag{2.86}$$

is introduced. After using the definition of the linear operator $L_{\bar{h}}$ the equation (2.85) was moved to the error equation (2.84). Keep in mind that the linear operator $L_{\bar{h}}$ is not necessarily an isomorphism if the linear operator L_{h+} is one. This depends strictly on the used joining operator $\mathcal{J}_{h+}$.

In general, the error equation (2.85) is not equivalent to the starting error equation (2.83) as the dimension of the space $V_{\bar{h}}$ can be lower than the

dimension of the expansion V_{h+}. Therefore a loss of information takes place when going from the error equation (2.83) to the equation (2.85) defining the error estimate $\eta_{\bar{h}}$. However, it has to be assumed that $L_{\bar{h}}$ is an isomorphism. Moreover it will turn out that under certain circumstances one gets over the loss of information, i.e. the error estimate $\eta_{\bar{h}}$ is still equivalent to the true error in the sense of Definition 3.

There are three interesting selections for the space $V_{\bar{h}}$ basing on the splitting

$$V_{h+} = V_h \oplus V_h^c \tag{2.87}$$

with $V_h \cap V_h^c = \{0\}$:

- At first one can set $V_{\bar{h}} := V_{h+}$, $\mathcal{J}_{h+} := I_{V_{h+}}$ and $L_{\bar{h}} := L_{h+}$. Then the error estimate (2.84) is equal to the error estimate η_h^I defined by equation (2.83). This is called the *inflating a-posteriori error estimate*. But still the target to reduce the computational costs for the error estimate is not reached. But if it is assumed that the components of the better discrete solution u_{h+} belonging to the space V_h are close to the discrete solution u_h so it is sufficient to look only to the components in the space V_h^c.

- This is the idea for the *hierarchical error estimate* (denoted by η_h^H). Here $V_{\bar{h}} := V_h^c$, $\mathcal{J}_{h+} := I_{V_{\bar{h}}}$ and $L_{\bar{h}} := L_h^c \in \mathcal{L}(V_h^c, (V_h^c)^*)$ are set. The base φ^h is extended by additional basis elements $(\varphi_j^{h+})_{j=d^h+1,d^{h+}}$ spanning the space V_h^c. For the calculation of the error estimate the stiffness matrix

  $$(< \varphi_j^{h+}, L_h^c \varphi_i^{h+} >)_{i,j=d^h+1,d^{h+}} \tag{2.88}$$

  and the defect

  $$(< \varphi_j^{h+}, F_{h+}(\hat{u}_h) >)_{j=d^h+1,d^{h+}} \tag{2.89}$$

  have to be mounted. The operator L_h^c has to be a suitable isomorphism and should be selected in a way that the inverse of its stiffness matrix can be easily calculated. Common selections use a lumped matrix, the reduction of the matrix (2.88) to its main diagonal elements in combination with hierarchical bases, e.g. see Zienkiewicz [69], Bank [15], Deufelhard [31], or the solution of element-by-element problems, see Liu [47]. Although this error estimate works very well for a wide range of applications, there is no general method for the selection of the linear operator L_h^c. The essential problem is that it has to be an isomorphism.

- The new error estimate is called the *projecting error estimate* (denoted by η_h^P). It bases on the idea of projecting the error equation (2.83) back

to the space V_h where the solution approximation $\hat{u}_h$ is calculated. This is achieved by setting $V_{\bar{h}} := V_h$ and $\mathcal{J}_{h+}$ to an interpolation operator into V_{h+}, for more details see Section 3.6. Then one can set $L_{\bar{h}} := L_h^{(k-1)}$ which has been used to calculate the returned approximation $\hat{u}_h = u_h^{(k)}$, see iteration procedure (2.41). The profit is that the stiffness matrix and, if a direct solution method for the solution of the systems of linear equations is used, its LU-decomposition or other manipulations of the stiffness matrix (e.g. reordering, ILU-factorization for preconditioning) are reused for the a-posteriori error estimate. Only the new defect

$$< \mathcal{J}_{h+}\varphi^h, F_{h+}(\hat{u}_h) >:= (< \mathcal{J}_{h+}\varphi_j^h, F_{h+}(\hat{u}_h) >)_{j=1,d^h} \tag{2.90}$$

has to be mounted.

Returning to the general point of view it is obvious that the error estimate defined by error equation (2.84) is not a good estimate if the range of $\mathcal{J}_{h+}$ is a subset of the space V_h, i.e. $\mathcal{J}_{h+}[V_{\bar{h}}] \subset V_h$. As no contribution out of the space V_h is involved only the error from the termination of the iteration procedure and the integration error is considered. To insert the interpolation error the range of $\mathcal{J}_{h+}$ has to be large enough. Here it is assumed that the range of $\mathcal{J}_{h+}$ contains *all* components that are added to the space V_h to construct the expansion V_{h+}, i.e. it holds

$$V_h^c \subset \mathcal{J}_{h+}[V_{\bar{h}}] \; . \tag{2.91}$$

In the following a more handy formulation of this condition is used which says that a right hand side inverse $\mathcal{J}_{\bar{h}} \in \mathcal{L}(V_h^c, V_{\bar{h}})$ of $\mathcal{J}_{h+}$ on the space V_h^c exists:

$$\mathcal{J}_{h+}\mathcal{J}_{\bar{h}} v_h^c = v_h^c \text{ for all } v_h^c \in V_h^c \; . \tag{2.92}$$

The conditions (2.91) and (2.92) are equivalent. For the inflating and the hierarchical error estimate the involved joining operator has the required property (2.92) because of the definition of the method. For the construction of the projecting error estimate this property has to be considered when selecting $\mathcal{J}_{h+}$.

The following Lemma 2 and Lemma 4 intend to prove an estimation of the type

$$q_h \|\eta_{h+}\| \leq \|\eta_{\bar{h}}\| \leq Q_h \|\eta_{h+}\| \tag{2.93}$$

for the a–posteriori error estimate $\eta_{\bar{h}}$ defined by equation (2.84) where the element η_{h+} is defined by equation (2.76). The positive values q_h and Q_h depend on the mesh size h. By combining this estimation with the results of Lemma 1 it is proved in Theorem 4 that an a–posteriori error estimate $\eta_{\bar{h}}$ is equivalent to the true error in the sense of Definition 3.

Lemma 2 (Grosz) *Let $F_{h+} : K_{h+} \to V_{h+}^*$ be well–posed with condition number D_{h+}^2, $\hat{u}_h, u_{h+} \in K_{h+}$ with $F_{h+}(u_{h+}) = 0$, $L_{\bar{h}} \in \mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)$ and $\mathcal{J}_{h+} \in \mathcal{L}(V_{\bar{h}}, V_{h+})$. Then for the error estimate $\eta_{\bar{h}}$ defined by equation (2.84) the following estimate holds with $\eta_{h+} = u_{h+} - \hat{u}_h$:*

$$\|\eta_{\bar{h}}\| \leq D_{h+} \|\mathcal{J}_{h+}\|_{\mathcal{L}(V_{\bar{h}}, V_{h+})} \|L_{\bar{h}}^{-1}\|_{\mathcal{L}(V_{\bar{h}}^*, V_{\bar{h}})} \|\eta_{h+}\| \,. \tag{2.94}$$

Proof: With $F_{h+}(u_{h+}) = 0$ and $\mathcal{J}_{h+}[V_{\bar{h}}] \subset V_{h+}$ one gets from the definition (2.84) of error estimate $\eta_{\bar{h}}$:

$$\begin{aligned} < v_{\bar{h}}, L_{\bar{h}} \eta_{\bar{h}} > &\overset{(2.84)}{=} - < \mathcal{J}_{h+} v_{\bar{h}}, F_{h+}(\hat{u}_h) > \\ &= < \mathcal{J}_{h+} v_{\bar{h}}, F_{h+}(u_{h+}) - F_{h+}(\hat{u}_h) > \\ &\leq \|\mathcal{J}_{h+}\|_{\mathcal{L}(V_{\bar{h}}, V_{h+})} \|v_{\bar{h}}\| \|F_{h+}(u_{h+}) - F_{h+}(\hat{u}_h)\|_{V_{h+}^*} \\ &\leq D_{h+} \|\mathcal{J}_{h+}\|_{\mathcal{L}(V_{\bar{h}}, V_{h+})} \|v_{\bar{h}}\| \|u_{h+} - \hat{u}_h\| \\ &\overset{(2.76)}{=} D_{h+} \|\mathcal{J}_{h+}\|_{\mathcal{L}(V_{\bar{h}}, V_{h+})} \|v_{\bar{h}}\| \|\eta_{h+}\| \end{aligned} \tag{2.95}$$

for all $v_{\bar{h}} \in V_{\bar{h}}$. From the the definition of the norm $\|L_{\bar{h}}^{-1}\|_{\mathcal{L}(V_{\bar{h}}^*, V_{\bar{h}})}$ it is

$$\begin{aligned} \|\eta_{\bar{h}}\| &\overset{(2.8)}{\leq} \|L_{\bar{h}}^{-1}\|_{\mathcal{L}(V_{\bar{h}}^*, V_{\bar{h}})} \|L_{\bar{h}} \eta_{\bar{h}}\|_{V^*} \\ &\overset{(2.11)}{\leq} \|L_{\bar{h}}^{-1}\|_{\mathcal{L}(V_{\bar{h}}^*, V_{\bar{h}})} \sup_{v_{\bar{h}} \in V_{\bar{h}}} \frac{< v_{\bar{h}}, L_{\bar{h}} \eta_{\bar{h}} >}{\|v_{\bar{h}}\|} \,. \end{aligned} \tag{2.96}$$

After inserting estimation (2.95) into estimation (2.96) the inequality of the lemma has been proved •

Unfortunately the techniques in the proof of Lemma 2 cannot be applied to obtain a value for q_h in the objected estimation (2.93) since in general it cannot be assumed that $V_{h+} \subset \mathcal{J}_{h+}[V_{\bar{h}}]$. More refined tools have to be used by following the techniques from the analysis of two-level iteration methods, see Eijkhout [34]. In the following the angular distance between the spaces V_h and V_h^c is important. It is measured by the deflection κ_h in the Pythagorean equation for the spaces V_h and V_h^c:

Lemma 3 *Let V_h and V_h^c be finite dimensional subspaces of V with $V_h \cap V_h^c = \{0\}$. Then for the deflection κ_h in the Pythagorean equation for the spaces V_h and V_h^c the following relation holds:*

$$1 \leq \kappa_h^2 := \sup_{v_h \in V_h, v_h^c \in V_h^c} \frac{\|v_h\|^2 + \|v_h^c\|^2}{\|v_h + v_h^c\|^2} < \infty \,. \tag{2.97}$$

Proof: With $v_h = 0$ one gets $\kappa_h \geq 1$. To show that $\kappa_h \in \mathbb{R}$ contradiction is used:

If $\kappa_h = \infty$ there are sequences $(v_h^{(n)})_{n\in\mathbb{N}} \in V_h^{\mathbb{N}}$ and $(w_h^{(n)})_{n\in\mathbb{N}} \in (V_h^c)^{\mathbb{N}}$ with

$$0 < \|v_h^{(n)} + w_h^{(n)}\|^2 \leq \frac{1}{n}(\|v_h^{(n)}\|^2 + \|w_h^{(n)}\|^2) \tag{2.98}$$

for all $n \in \mathbb{N}$. With $\gamma_n := \max(\|v_h^{(n)}\|, \|w_h^{(n)}\|)$ it is set

$$\tilde{v}_h^{(n)} := \frac{1}{\gamma_n} v_h^{(n)} \text{ and } \tilde{w}_h^{(n)} := \frac{1}{\gamma_n} w_h^{(n)} \tag{2.99}$$

for all $n \in \mathbb{N}$. Then

$$max(\|\tilde{v}_h^{(n)}\|, \|\tilde{w}_h^{(n)}\|) = 1 \tag{2.100}$$

holds for all $n \in \mathbb{N}$.

Since the spaces V_h and V_h^c have finite dimensions and the sequence $((\tilde{v}_h^{(n)}, \tilde{w}_h^{(n)}))_{n\in\mathbb{N}}$ is bounded there is a subsequence of the sequence $((\tilde{v}_h^{(n)}, \tilde{w}_h^{(n)}))_{n\in\mathbb{N}}$ which converges to an element $(\tilde{v}_h, \tilde{w}_h) \in V_h \times V_h^c$. For simplification this subsequence is also denoted by $((\tilde{v}_h^{(n)}, \tilde{w}_h^{(n)}))_{n\in\mathbb{N}}$. By using inequality (2.98) one obtains that

$$\begin{aligned} \|\tilde{v}_h^{(n)} + \tilde{w}_h^{(n)}\|^2 &\overset{(2.99)}{=} \|\frac{v_h^{(n)}}{\gamma_n} + \frac{w_h^{(n)}}{\gamma_n}\|^2 \\ &\overset{(2.98)}{\leq} \frac{1}{n}\frac{1}{\gamma_n^2}(\|v_h^{(n)}\|^2 + \|w_h^{(n)}\|^2) \\ &\leq \frac{2}{n} . \end{aligned} \tag{2.101}$$

By taking $\lim_{n\to\infty}$ on this estimate the result is that $\tilde{v}_h = -\tilde{w}_h$. Therefore it has to be $\tilde{v}_h, \tilde{w}_h \in V_h \cap V_h^c = \{0\}$ and consequently $\tilde{v}_h = \tilde{w}_h = 0$. But from equation (2.100) it has to be

$$max(\|\tilde{v}_h\|, \|\tilde{w}_h\|) = 1 . \tag{2.102}$$

This is a contradiction and therefore κ_h has to be finite •

Remark: If V is a Hilbert space with scalar product $< ., . >$ the deflection κ_h has a geometrical interpretation: The value

$$\gamma_h := \sup_{v_h\in V_h, v_h^c \in V_h^c} \frac{< v_h, v_h^c >}{\|v_h\|\|v_h^c\|} < 1 \tag{2.103}$$

is the cosine of the angle between the spaces V_h and V_h^c (The proof for $\gamma_h < 1$ is similar to the proof of Lemma 3). The constant γ_h plays an important role in the multilevel theory, see Eijkhout [34]. There is a relation of γ_h to κ_h:

For all $v_h \in V_h$ and $v_h^c \in V_h^c$ it is

$$\begin{aligned} \frac{\|v_h\|^2 + \|v_h^c\|^2}{\|v_h + v_h^c\|^2} &= \frac{\|v_h\|^2 + \|v_h^c\|^2}{\|v_h\|^2 + \|v_h^c\|^2 + 2 < v_h, v_h^c >} \\ &\overset{(2.103)}{\leq} \frac{\|v_h\|^2 + \|v_h^c\|^2}{\|v_h\|^2 + \|v_h^c\|^2 - 2\gamma_h \|v_h^c\| \, \|v_h\|} \\ &\leq \frac{1}{1 - \gamma_h} \, . \end{aligned} \tag{2.104}$$

In the last estimation the fact is used that

$$\begin{aligned} 2\|v_h^c\| \, \|v_h\| &= \|v_h\|^2 + \|v_h^c\|^2 - (\|v_h\| - \|v_h^c\|)^2 \\ &\leq \|v_h\|^2 + \|v_h^c\|^2 \, . \end{aligned} \tag{2.105}$$

By taking the supreme value over $v_h \in V_h$ and $v_h^c \in V_h^c$ in inequality (2.104) it is shown that $\kappa_h \leq \frac{1}{\sqrt{1-\gamma_h}}$. Moreover γ_h is actually a maximum. By inserting the location of the maximum it can be proved that even

$$\kappa_h = \frac{1}{\sqrt{1 - \gamma_h}} \tag{2.106}$$

holds. If the spaces V_h and V_h^c are orthogonal it is $\kappa_h = 1$.

Lemma 4 (Grosz) *Let $F_{h+} : K_{h+} \to V_{h+}^*$ be well–posed with condition number D_{h+}^2, $\hat{u}_h, u_{h+} \in K_{h+}$ with $F_{h+}(u_{h+}) = 0$ and $L_{\bar{h}} \in \mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)$. Moreover let be $V_{h+} = V_h \oplus V_h^c$ with $V_h \cap V_h^c = \{0\}$ and $\mathcal{J}_{h+} \in \mathcal{L}(V_{\bar{h}}, V_{h+})$ with left hand side inverse $\mathcal{J}_{\bar{h}} \in \mathcal{L}(V_h^c, V_{\bar{h}})$ on the space V_h^c defined by equation (2.92). Then the error estimate $\eta_{\bar{h}}$ defined by equation (2.84) fulfills the following estimate with $\eta_{h+} = u_{h+} - \hat{u}_h$:*

$$\|\eta_{h+}\|^2 \leq D_{h+}^2 \kappa_h^2 (\|F_{h+}(\hat{u}_h)\|_{V_h^*}^2 + \|L_{\bar{h}}\|_{\mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)}^2 \|\mathcal{J}_{\bar{h}}\|_{\mathcal{L}(V_h^c, V_{\bar{h}})}^2 \|\eta_{\bar{h}}\|^2) \tag{2.107}$$

where κ_h is the deflection in the Pythagorean equation for the spaces V_h and V_h^c defined by equation (2.97).

Proof: For every element $v_{h+} \in V_{h+}$ one gets from splitting (2.87)

$$v_{h+} = v_h + v_h^c \tag{2.108}$$

with $v_h \in V_h$ and $v_h^c \in V_h^c$. It is

$$v_h^c = \mathcal{J}_{h+} \mathcal{J}_{\bar{h}} v_h^c \tag{2.109}$$

because of the condition (2.92) for the joining operator $\mathcal{J}_{h+}$ and its right hand side inverse $\mathcal{J}_{\bar{h}}$. By involving the definition (2.84) of the error estimate $\eta_{\bar{h}}$ one obtains:

$$\begin{aligned}
& < v_{h+}, F_{h+}(\hat{u}_h) > \\
\overset{(2.108)}{=} \quad & < v_h, F_{h+}(\hat{u}_h) > + < v_h^c, F_{h+}(\hat{u}_h) > \\
\overset{(2.109)}{=} \quad & < v_h, F_{h+}(\hat{u}_h) > + < \mathcal{J}_{h+}\mathcal{J}_{\bar{h}} v_h^c, F_{h+}(\hat{u}_h) > \\
\overset{(2.84)}{=} \quad & < v_h, F_{h+}(\hat{u}_h) > + < \mathcal{J}_{\bar{h}} v_h^c, L_{\bar{h}}\eta_{\bar{h}} > \\
\leq \quad & \|v_h\| \, \|F_{h+}(\hat{u}_h)\|_{V_h^*} + \\
& \qquad \|v_h^c\| \, \|L_{\bar{h}}\|_{\mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)} \, \|\mathcal{J}_{\bar{h}}\|_{\mathcal{L}(V_h^c, V_{\bar{h}})} \|\eta_{\bar{h}}\| \, .
\end{aligned} \tag{2.110}$$

Taking the Cauchy–Schwartz inequality it turns out that

$$\begin{aligned}
& < v_{h+}, F_{h+}(\hat{u}_h) > \\
\leq \quad & \sqrt{\|v_h\|^2 + \|v_h^c\|^2} \\
& \quad \sqrt{\|F_{h+}(\hat{u}_h)\|_{V_h^*}^2 + \|L_{\bar{h}}\|_{\mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)}^2 \|\mathcal{J}_{\bar{h}}\|_{\mathcal{L}(V_h^c, V_{\bar{h}})}^2 \|\eta_{\bar{h}}\|^2} \\
\overset{(2.97)}{\leq} \quad & \kappa_h \|v_h + v_h^c\| \\
& \quad \sqrt{\|F_{h+}(\hat{u}_h)\|_{V_h^*}^2 + \|L_{\bar{h}}\|_{\mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)}^2 \|\mathcal{J}_{\bar{h}}\|_{\mathcal{L}(V_h^c, V_{\bar{h}})}^2 \|\eta_{\bar{h}}\|^2} \\
\overset{(2.108)}{=} \quad & \kappa_h \|v_{h+}\| \\
& \quad \sqrt{\|F_{h+}(\hat{u}_h)\|_{V_h^*}^2 + \|L_{\bar{h}}\|_{\mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)}^2 \|\mathcal{J}_{\bar{h}}\|_{\mathcal{L}(V_h^c, V_{\bar{h}})}^2 \|\eta_{\bar{h}}\|^2}
\end{aligned} \tag{2.111}$$

by using the definition (2.97) of the deflection κ_h. From the fact that the discrete operator F_{h+} is well–posed with condition number D_{h+} and $F_{h+}(u_{h+}) = 0$ the following estimations hold:

$$\begin{aligned}
\|\eta_{h+}\| \quad \overset{(2.76)}{=} \quad & \|u_{h+} - \hat{u}_h\| \\
\leq \quad & D_{h+} \|F_{h+}(u_{h+}) - F_{h+}(\hat{u}_h)\|_{V_{h+}^*} \\
= \quad & D_{h+} \|F_{h+}(\hat{u}_h)\|_{V_{h+}^*} \\
\overset{(2.11)}{=} \quad & D_{h+} \sup_{v_{h+} \in V_{h+}} \frac{< v_{h+}, F_{h+}(\hat{u}_h) >}{\|v_{h+}\|} \, .
\end{aligned} \tag{2.112}$$

After inserting estimation (2.111) the lemma has been proved •

Remark: If $\|F_{h+}(\hat{u}_h)\|_{V_h^*} = 0$ Lemma 4 actually establishes an estimation for q_h in the wanted inequality (2.93). Even if there is no termination error (i.e $\hat{u}_h = u_h$) the term $\|F_{h+}(\hat{u}_h)\|_{V_h^*}$ does not vanish. The reason is that in the practical implementation it cannot be expected that $F_h(u_h) = F_{h+}(u_h)|_{V_h}$ holds, i.e. in general the discrete operator F_{h+} is not a continuation of the operator F_h from the space V_h to its expansion V_{h+}. However, it has to be

requested that the distance of the discrete operators $F_h(u_h)$ and $F_{h+}(u_h)|_{V_h}$ is small enough compared to the approximation error $\|u - u_h\|$, see condition (2.113) below.

By gathering the results of this section the main theorem of this chapter is stated:

Theorem 4 (Grosz) *Let $u \in V$ be a given element in the Banach space V.*

- *Let for all $h > 0$ $V_h \subset V$ be a finite dimensional subspace of the space V, $K_h \subset V_h$, $F_h : K_h \to V_h^*$ well–posed with condition number D^2 and $u_h \in K_h$ with $F_h(u_h) = 0$ and $u_h \to u$ for $h \to 0$.*

- *Let for all $h > 0$ $V_{h+} \subset V$ be a finite dimensional subspace of the space V and $K_{h+} \subset V_{h+}$ with $V_h \subset V_{h+}$ and $K_h \subset K_{h+}$. Let $F_{h+} : K_{h+} \to V_{h+}^*$ be well–posed with condition number D^2 with*

$$\|F_{h+}(u_h) - F_h(u_h)\|_{V_h^*} \leq s_h \|u - u_h\| \tag{2.113}$$

  and $\lim_{h\to 0} s_h = 0$. Moreover let $(u_h, u_{h+})_{h>0}$ be saturated for the solution u with saturation bound $0 \leq r_0 < 1$ in sense of Definition 2.

- *Let for all $h > 0$ $V_{\bar{h}} \subset V$ and $L_{\bar{h}} \in \mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)$ be an isomorphism with $\limsup_{h\to 0} \|L_{\bar{h}}\|_{\mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)} \leq L$ and $\limsup_{h\to 0} \|L_{\bar{h}}^{-1}\|_{\mathcal{L}(V_{\bar{h}}^*, V_{\bar{h}})} \leq L$.*

- *Let for all $h > 0$ be*

$$V_{h+} = V_h \oplus V_h^c \tag{2.114}$$

  with $V_h \cap V_h^c = \{0\}$ and

$$\sup_{v_h \in V_h, v_h^c \in V_h^c} \frac{\|v_h\|^2 + \|v_h^c\|^2}{\|v_h + v_h^c\|^2} \leq \kappa^2 \tag{2.115}$$

  for fixed $\kappa \in \mathbb{R}_+$. In addition let be $\mathcal{J}_{\bar{h}} \in \mathcal{L}(V_h^c, V_{\bar{h}})$ and $\mathcal{J}_{h+} \in \mathcal{L}(V_{\bar{h}}, V_{h+})$ with

$$\mathcal{J}_{h+}\mathcal{J}_{\bar{h}} v_h^c = v_h^c \text{ for all } v_h^c \in V_h^c \,, \tag{2.116}$$

  $\limsup_{h\to 0} \|\mathcal{J}_{h+}\|_{\mathcal{L}(V_{\bar{h}}, V_{h+})} \leq P$ *and* $\limsup_{h\to 0} \|\mathcal{J}_{\bar{h}}\|_{\mathcal{L}(V_h^c, V_{\bar{h}})} \leq P$.

If $\hat{u}_h \in K_h$ fulfills the stopping criterion

$$\|F_h(\hat{u}_h)\|_{V_h^*} \leq \lambda \|F_{h+}(\hat{u}_h)\|_{\mathcal{J}_{h+}[V_{\bar{h}}]^*} \tag{2.117}$$

with $0 \leq \lambda < \lambda_0 := \frac{1}{D^2}\min(\frac{1-r_0}{2r_0}, \frac{1}{\kappa D^2})$ *then also* $\hat{u}_h \to u$ *for* $h \to 0$ *and the a-posteriori error estimate* $\eta_{\bar{h}} \in V_{\bar{h}}$ *defined by*

$$< v_{\bar{h}}, L_{\bar{h}}\eta_{\bar{h}} >= - < \mathcal{J}_{h+}v_{\bar{h}}, F_{h+}(\hat{u}_h) > \quad \textit{for all } v_{\bar{h}} \in V_{\bar{h}} \tag{2.118}$$

is equivalent to the true error $e_h := u - \hat{u}_h$ *in the sense of Definition 3.*

Proof: Since $\|F_{h+}(\hat{u}_h)\|_{\mathcal{J}_{h+}[V_{\bar{h}}]^*} \leq \|F_{h+}(\hat{u}_h)\|_{V_{h+}^*}$ and the factor λ used in stopping criterion (2.117) is less than $\frac{1}{D^2}\min(\frac{1-r_0}{2r_0}, 1)$ (it is $\kappa D^2 \geq 1$!) the stopping criterion (2.56) in the propositions of Theorem 3 holds. Therefore Theorem 3 shows that the approximations $\hat{u}_h$ converge to the element u for $h \to 0$.

First the existence of an upper bound for the ratio

$$\theta_h := \frac{\|\eta_{\bar{h}}\|}{\|e_h\|} \tag{2.119}$$

is proved when $h \to 0$. From inequality (2.94) in Lemma 2 one obtains

$$\limsup_{h\to 0} \frac{\|\eta_{\bar{h}}\|}{\|\eta_{h+}\|} \leq C \tag{2.120}$$

with

$$C := D\ L\ P \tag{2.121}$$

and the error estimate η_{h+} defined by equation (2.76) in Lemma 1. Moreover by Lemma 1 the inequality (2.77)

$$\limsup_{h\to 0} \frac{\|\eta_{h+}\|}{\|e_h\|} \leq 1 + \hat{r}_0 \tag{2.122}$$

with the factor $\hat{r}_0$ defined by equation (2.58) holds. Therefore an upper bound for the ratio θ_h turns out from

$$\begin{aligned} \limsup_{h\to 0} \theta_h &\overset{(2.119)}{=} \limsup_{h\to 0} \frac{\|\eta_{h+}\|}{\|e_h\|} \cdot \frac{\|\eta_{\bar{h}}\|}{\|\eta_{h+}\|} \\ &\overset{(2.122)+(2.120)}{\leq} (1+\hat{r}_0)\cdot C\ . \end{aligned} \tag{2.123}$$

To find a lower bound for the ratio θ_h Lemma 4 is used but an estimate for the norm $\|F_{h+}(\hat{u}_h)\|_{V_h^*}$ is needed:

Using the condition (2.113) and the fact that the discrete operators F_{h+} and F_h are well–posed it is

$$\begin{aligned} \|F_{h+}(\hat{u}_h)\|_{V_h^*} &= \|F_{h+}(\hat{u}_h) - F_h(u_h)\|_{V_h^*} \\ &\leq \|F_{h+}(\hat{u}_h) - F_{h+}(u_h)\|_{V_{h+}^*} + \\ &\qquad \|F_{h+}(u_h) - F_h(u_h)\|_{V_h^*} \\ &\quad D\|\hat{u}_h - u_h\| + \|F_{h+}(u_h) - F_h(u_h)\|_{V_h^*}\ . \end{aligned} \tag{2.124}$$

To get further estimates the definition of the factor s_h by inequality (2.113) is inserted:

$$\begin{array}{rcl} \|F_{h+}(\hat{u}_h)\|_{V_h^*} & \overset{(2.113)}{\leq} & D\|\hat{u}_h - u_h\| + s_h\|u - u_h\| \\ & \leq & D\|\hat{u}_h - u_h\| + s_h(\|u - \hat{u}_h\| + \|\hat{u}_h - u_h\|) \\ & = & (D + s_h)\|\hat{u}_h - u_h\| + s_h\|e_h\| \\ & \leq & (D + s_h)D\|F_h(\hat{u}_h) - F_h(u_h)\|_{V_h^*} + s_h\|e_h\| \\ & = & (D + s_h)D\|F_h(\hat{u}_h)\|_{V_h^*} + s_h\|e_h\| \ . \end{array} \tag{2.125}$$

Using stopping criterion (2.117) and $F_{h+}(u_{h+}) = 0$ further estimates can be made:

$$\begin{array}{rcl} & & \|F_{h+}(\hat{u}_h)\|_{V_h^*} \\ & \overset{(2.117)}{\leq} & \lambda(D + s_h)D\|F_{h+}(\hat{u}_h)\|_{V_{h+}^*} + s_h\|e_h\| \\ & = & \lambda(D + s_h)D\|F_{h+}(\hat{u}_h) - F_{h+}(u_{h+})\|_{V_{h+}^*} + s_h\|e_h\| \\ & \leq & \lambda(D + s_h)D^2\|\hat{u}_h - u_{h+}\| + s_h\|e_h\| \\ & \overset{(2.76)}{=} & \lambda(D + s_h)D^2\|\eta_{h+}\| + s_h\|e_h\| \ . \end{array} \tag{2.126}$$

Inserting this estimate into the inequality (2.107) of Lemma 4 one gets

$$\|\eta_{h+}\|^2 \leq D^2\kappa^2\Big([\lambda(D + s_h)D^2\|\eta_{h+}\| + s_h\|e_h\|]^2 + C_h^2\kappa^2\|\eta_{\bar{h}}\|^2\Big) \tag{2.127}$$

where it is set

$$C_h := D\|L_{\bar{h}}\|_{\mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)}\|\mathcal{J}_{\bar{h}}\|_{\mathcal{L}(V_h^c, V_{\bar{h}})} \ . \tag{2.128}$$

By solving this inequality for the ratio $\frac{\|\eta_{\bar{h}}\|}{\|\eta_{h+}\|}$ it is

$$\frac{1}{C_h^2\kappa^2}(1 - D^2\kappa^2[\lambda(D + s_h)D^2 + s_h\frac{\|e_h\|}{\|\eta_{h+}\|}]^2) \leq \frac{\|\eta_{\bar{h}}\|^2}{\|\eta_{h+}\|^2} \ . \tag{2.129}$$

Since it is assumed that $\lim_{h\to 0} s_h = 0$, the ratio $\frac{\|e_h\|}{\|\eta_{h+}\|}$ is bounded by Lemma 1 and it is $\lambda < \frac{1}{\kappa D^4}$ the left hand side of inequality (2.129) is positive for a small mesh size h. By Lemma 1 it is

$$\begin{aligned} \liminf_{h\to 0} \frac{\|\eta_{h+}\|}{\|e_h\|} &\geq 1 - \hat{r}_0 > 0 \\ \liminf_{h\to 0} \frac{\|e_h\|}{\|\eta_{h+}\|} &\geq \frac{1}{1 + \hat{r}_0} > 0 \end{aligned} \tag{2.130}$$

and therefore it turns out from inequality (2.129) that

$$\begin{array}{rcl} \liminf_{h\to 0} \theta_h & = & \liminf_{h\to 0} \frac{\|\eta_{h+}\|}{\|e_h\|} \cdot \frac{\|\eta_{\bar{h}}\|}{\|\eta_{h+}\|} \\ & \overset{(2.130)+(2.129)}{\geq} & (1 - \hat{r}_0) \cdot \frac{1}{C\kappa}\sqrt{1 - D^8\kappa^2\lambda^2} \end{array} \tag{2.131}$$

with the constant C defined by equation (2.121). The lower bound has a positive, real value since it is $\lambda < \frac{1}{\kappa D^4}$.

After combining the inequalities (2.123) and (2.129) it has been proved that

$$\begin{aligned}(1-\hat{r}_0)\cdot\tfrac{1}{C\kappa}\sqrt{1-D^8\kappa^2\lambda^2} &\overset{(2.131)}{\leq} \liminf_{h\to 0}\theta_h \leq \\ \limsup_{h\to 0}\theta_h &\overset{(2.123)}{\leq} (1+\hat{r}_0)\cdot C\ .\end{aligned} \tag{2.132}$$

So the inequality (2.74) for the error estimate $\eta_{\bar{h}}$ has been verified. Therefore the error estimate $\eta_{\bar{h}}$ is equivalent to the true error e_h in the sense of Definition 3 •

Remark: If in the condition (2.113) it is

$$\limsup_{h\to 0} s_h = s_0 > 0 \tag{2.133}$$

the results of Theorem 4 are still valid with another constant λ_0 but the limit s_0 has to be small enough.

The proof has explicitly constructed an effectivity index of the a–posteriori error estimate $\eta_{\bar{h}}$. The following corollary of Theorem 4 notes this result for an exactly solved, discrete variational problem (2.30):

Corollary 2 (Grosz) *Under the assumptions of Theorem 4 with $\lambda = 0$ it is*

$$\frac{1-r_0}{\kappa DLP} \leq \liminf_{h\to 0}\frac{\|\eta_{\bar{h}}\|}{\|e_h\|} \leq \limsup_{h\to 0}\frac{\|\eta_{\bar{h}}\|}{\|e_h\|} \leq (1+r_0)DLP\ . \tag{2.134}$$

An effectivity index in the sense of Definition 3 is given by $\frac{\kappa DLP}{1-r_0}$.

Proof: The inequality (2.134) is a direct consequence of inequality (2.132). An effectivity index is obtained from the fact that $\kappa \geq 1$ and $1+r_0 \leq \frac{1}{1-r_0}$ •

Iterative methods, e.g. conjugate gradient methods, do not solve the linear equation (2.118) exactly. A poor accuracy is sufficient to ensure that the approximation $\hat{\eta}_{\bar{h}}$ of the the error estimate $\eta_{\bar{h}}$ is equivalent to the true error:

Corollary 3 (Grosz) *Under the assumptions of Theorem 4 let for all $h > 0$ be $\hat{\eta}_{\bar{h}} \in V_{\hat{h}}$ defined by*

$$\begin{aligned}< v_{\bar{h}}, L_{\bar{h}}\hat{\eta}_{\bar{h}} > = - < \mathcal{J}_{h+}v_{\bar{h}}, F_{h+}(\hat{u}_h) > + < v_{\bar{h}}, d_{\bar{h}} > \\ \text{for all } v_{\bar{h}} \in V_{\bar{h}}\end{aligned} \tag{2.135}$$

with $d_{\bar{h}} \in V_{\bar{h}}^*$. Then there is a constant $\tau_0 > 0$ independent of the mesh size h that for all $\tau_0 > \tau \geq 0$ the following statement holds: If for all $h > 0$

$$\|d_{\bar{h}}\|_{V_{\bar{h}}^*} \leq \tau \|F_{h+}(\hat{u}_h)\|_{\mathcal{J}_{h+}[V_{\bar{h}}]^*} \tag{2.136}$$

the a-posteriori error estimate $\hat{\eta}_{\bar{h}}$ is equivalent to the true error.

Proof: From the definitions of the error estimates $\hat{\eta}_{\bar{h}}$ and $\eta_{\bar{h}}$ it is

$$\begin{array}{rcl} < v_{\bar{h}}, d_{\bar{h}} > & \overset{(2.135)}{=} & < v_{\bar{h}}, L_{\bar{h}} \hat{\eta}_{\bar{h}} > + < \mathcal{J}_{h+} v_{\bar{h}}, F_{h+}(\hat{u}_h) > \\ & \overset{(2.118)}{=} & < v_{\bar{h}}, L_{\bar{h}} [\hat{\eta}_{\bar{h}} - \eta_{\bar{h}}] > \end{array} \tag{2.137}$$

for all $v_{\bar{h}} \in V_{\bar{h}}$. By using the triangle inequality and condition (2.136) it is

$$\begin{array}{rcl} \left| \|\hat{\eta}_{\bar{h}}\| - \|\eta_{\bar{h}}\| \right| & \leq & \|\hat{\eta}_{\bar{h}} - \eta_{\bar{h}}\| \\ & \overset{(2.137)}{\leq} & \|L_{\bar{h}}^{-1}\|_{\mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)} \|d_{\bar{h}}\|_{V_{\bar{h}}^*} \\ & \overset{(2.136)}{\leq} & \tau \|L_{\bar{h}}^{-1}\|_{\mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)} \|F_{h+}(\hat{u}_h)\|_{\mathcal{J}_{h+}[V_{\bar{h}}]^*} \\ & \leq & \tau \|L_{\bar{h}}^{-1}\|_{\mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)} \|F_{h+}(\hat{u}_h)\|_{V_{h+}^*} \ . \end{array} \tag{2.138}$$

With $F_{h+}(u_{h+}) = 0$ and the fact that the discrete operator F_{h+} is well–posed further estimates can be made:

$$\begin{array}{rcl} \left| \|\hat{\eta}_{\bar{h}}\| - \|\eta_{\bar{h}}\| \right| & \leq & \tau \|L_{\bar{h}}^{-1}\|_{\mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)} \|F_{h+}(\hat{u}_h)\|_{V_{h+}^*} \\ & \leq & \tau \|L_{\bar{h}}^{-1}\|_{\mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)} \|F_{h+}(\hat{u}_h) - F_{h+}(u_{h+})\|_{V_{h+}^*} \\ & \leq & \tau D \|L_{\bar{h}}^{-1}\|_{\mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)} \|\hat{u}_h - u_{h+}\| \\ & \leq & \tau D \|L_{\bar{h}}^{-1}\|_{\mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)} (\|\hat{u}_h - u\| + \|u - u_{h+}\|) \\ & \overset{(2.69)}{\leq} & \tau D \|L_{\bar{h}}^{-1}\|_{\mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)} (1 + \hat{r}_h) \|e_h\| \end{array} \tag{2.139}$$

where $\hat{r}_h \leq 1$ is defined by equation (2.69) in the proof of Theorem 3. This establishes that

$$\left| \frac{\|\hat{\eta}_{\bar{h}}\|}{\|e_h\|} - \frac{\|\eta_{\bar{h}}\|}{\|e_h\|} \right| \leq C\tau \tag{2.140}$$

with

$$C := 2DL > 0 \tag{2.141}$$

independent of the factor τ and the mesh size h.

As the error estimate $\eta_{\bar{h}}$ is equivalent to the true error in the sense of Definition 3 there is a constant $Q > 0$ with

$$\frac{1}{Q} \leq \liminf_{h \to 0+} \frac{\|\eta_{\bar{h}}\|}{\|e_h\|} \leq \limsup_{h \to 0+} \frac{\|\eta_{\bar{h}}\|}{\|e_h\|} \leq Q \ . \tag{2.142}$$

Therefore the lower estimates

$$\begin{aligned} \liminf_{h\to 0+} \frac{\|\hat{\eta}_{\bar{h}}\|}{\|e_h\|} &\geq \liminf_{h\to 0+} \left[\frac{\|\eta_{\bar{h}}\|}{\|e_h\|} - \left| \frac{\|\hat{\eta}_{\bar{h}}\|}{\|e_h\|} - \frac{\|\eta_{\bar{h}}\|}{\|e_h\|} \right| \right] \\ &\overset{(2.142)+(2.140)}{\geq} \frac{1}{Q} - C\tau \end{aligned} \tag{2.143}$$

and the upper estimates

$$\begin{aligned} \limsup_{h\to 0+} \frac{\|\hat{\eta}_{\bar{h}}\|}{\|e_h\|} &\leq \limsup_{h\to 0+} \left[\frac{\|\eta_{\bar{h}}\|}{\|e_h\|} + \left| \frac{\|\hat{\eta}_{\bar{h}}\|}{\|e_h\|} - \frac{\|\eta_{\bar{h}}\|}{\|e_h\|} \right| \right] \\ &\overset{(2.142)+(2.140)}{\leq} Q + C\tau \end{aligned} \tag{2.144}$$

can be made. When one selects $0 \leq \tau < \tau_0 := \frac{1}{CQ}$ the inequalities (2.143) and (2.144) can be combined to

$$0 < \frac{1}{Q} - C\tau \overset{(2.143)}{\leq} \liminf_{h\to 0+} \frac{\|\hat{\eta}_{\bar{h}}\|}{\|e_h\|} \leq \limsup_{h\to 0+} \frac{\|\hat{\eta}_{\bar{h}}\|}{\|e_h\|} \overset{(2.144)}{\leq} Q + \frac{1}{Q} \,. \tag{2.145}$$

That proves that the a–posteriori error estimate $\hat{\eta}_{\bar{h}}$ is equivalent to the true error in the sense of Definition 3 if the factor τ is small enough •

The linear functional $d_{\bar{h}} \in V_{\bar{h}}^*$ occuring in equation (2.135) is the defect arising from the inexact solution of the error equation (2.136) defining the a–posteriori error estimate $\eta_{\bar{h}}$. The criterion (2.136) can be used as a stopping criterion for iterative linear solvers, e.g. see LINSOL [65], where $d_{\bar{h}}$ is interpreted as the residual of the current approximation in the iteration procedure.

Remark 1: The value for τ_0 can be very small since C defined by equation (2.141) can be very large. However, a lot of tests have shown that for the most problems $\tau = 10^{-4}$ delivers reliable error estimates though Corollary 3 determines a smaller value τ.

Remark 2: The results of this section are also valid if the very popular problem dependent *energy norm* is used instead of the canonical norm in the Banach space V. In this case D=L=1 and then the effectivity index is closer to 1. However, the factor κ in the effectivity index produced by the reduction of the expansion V_{h+} does not vanish. It is the price which has to be paid to reduce the computational effort for the error estimates.

2.4 Discussion and Summary

Roughly spoken Corollary 2 shows that the effectivity index $\frac{\kappa DLP}{1-r_0}$ for the error estimate $\eta_{\bar{h}}$ depends on the deflection κ in the Pythagorean equation in spaces V_h and the expansion V_h^c, the condition number P^2 of the joining operator $\mathcal{J}_{h+}$, the condition number D^2 of the operator F and the condition number L^2 of the isomorphism $L_{\bar{h}}$. In most of the cases it is $L \approx D$ especially if Newton type methods are used. The uncertainty in the a–posteriori error estimate grows with the increase of the condition numbers and κ.

If a hierarchical a-posteriori error estimate is used the effectivity index is equal to $\frac{\kappa D^2}{1-r_0}$ since $P = 1$. That is the reason why the quality of a hierarchical estimate is better than the quality of the projecting error estimate. Using the inflating a–posteriori error estimate it is additionally $\kappa = 1$ and the best effectivity index $\frac{D^2}{1-r_0}$ of the three discussed types of error estimates can be expected. The costs for the better quality are additional computational effort. A more detailed comparison of the projecting a–posteriori error estimate especially to the hierarchical error estimate is given in Section 3.7.

In the next chapter Theorem 4 is applied to the new projecting a–posteriori error estimate in the range of the finite element discretization of non–linear boundary value problems on a domain Ω. The space V_h is a space of piecewise polynomials of order k and the expansion V_{h+} a space of polynomials of order $2k$. The discrete operators F_h and F_{h+} are constructed by numerical integration schemes which exactly integrate polynomials of degree $2k-2$ and $4k-1$. The construction ensures that the condition (2.113) with $\lim_{h\to 0} s_h = 0$ holds. The joining operators $\mathcal{J}_{h+}$ and $\mathcal{J}_{\bar{h}}$ are polynomial interpolation operators. The isomorphism $L_{\bar{h}}$ is a linearization of F_h, e.g. its Frechet derivative.

The proof that the discrete operators F_h and F_{h+} are well–posed is relatively simple. On the other hand it is more difficult to prove that the condition numbers, the norms $\|\mathcal{J}_{h+}\|_{\mathcal{L}(V_{\bar{h}},V_{h+})}$, $\|\mathcal{J}_{\bar{h}}\|_{\mathcal{L}(V_h^c,V_{\bar{h}})}$, $\|L_{\bar{h}}\|_{\mathcal{L}(V_{\bar{h}},V_{\bar{h}}^*)}$ and $\|L_{\bar{h}}^{-1}\|_{\mathcal{L}(V_{\bar{h}}^*,V_{\bar{h}})}$ and the deflection κ in the Pythagorean equation for the spaces V_h and V_h^c have upper bounds *independent* of the mesh size h. Fortunately this proof can be given for a problem type with a wide scope of applications (for instance, like in the next chapter for the non–linear Neumann problem) independent of the domain Ω and additional loads (see Remark 1 to Definition 1). However, the most crucial condition is that $(u_h, u_{h+})_{h>0}$ has to be saturated for the sought solution u. It will come out that this is related to the smoothness of the solution u which is typically determined by the shape of the domain Ω and additional loads. This aspect is investigated by Example 2, see Section 4.5.

Chapter 3

The Nonlinear Neumann Problem

3.1 Introduction

In this chapter the abstract theory developed in the previous Chapter 2 is applied to the finite element method (FEM) for a model problem namely for a class of non–linear boundary value problems on a polygonal shaped domain. More general formulations of the FEM especially for other boundary value problems are for instance presented in the books of Zienkiewicz [68], Quarteroni [52] and Ciarlet [25]. Naturally this chapter has not the target to introduce the FEM but to show the principles and crucial points of the projecting error estimate in the range of FEMs. The essential result is Theorem 14 which is the FEM formulation of Theorem 4 for the projecting a–posteriori error estimate. Roughly spoken Theorem 14 says that the projecting error estimate is equivalent to the true error in the sense of Definition 3 if the solution is smooth enough. To verify the properties of Theorem 4 the analysis follows closely the well–known linear theory of the FEM for elliptic problems given by Ciarlet [25] but some modifications have to be done to consider non–linear problems and the projecting error estimate. Extensions to other FEM applications are sketched.

3.2 Notations

For any dimension $n \in \mathbb{N}$ and every vector $x = (x_i)_{i=1,n} \in \mathbb{R}^n$ the real value

$$|x| := \sqrt{\sum_{i=1}^{n} x_i^2} \tag{3.1}$$

denotes the Euclidean norm of x.

For any matrix $B \in \mathbb{R}^{n \times n}$ the determinant of the matrix B is denoted by $det(B)$. The real value $|B|$ defined by

$$|B| := \sup_{x \in \mathbb{R}^n} \frac{|Bx|}{|x|} \tag{3.2}$$

denotes the norm of the matrix B. There is a constant $C > 0$ that it is

$$|b_{ij}| \leq C|B| \text{ for all } 1 \leq i, j \leq n \tag{3.3}$$

for all matrices $B = (b_{ij})_{i,j=1,n} \in \mathbb{R}^n$. The constant C depends only on the dimension n.

For any vector $x \in \mathbb{R}^n$ and $\delta > 0$

$$S(x, \delta) := \{y \in \mathbb{R}^n \mid |y - x| < \delta\} \tag{3.4}$$

denotes the ball of radius δ with center x.

For any set $K \subset \mathbb{R}^n$ $cl(K)$ denotes the closure of the set K, $int(K)$ is the open kernel and ∂K is the boundary of the set K. If the set K is bounded and it is $int(K) \neq \emptyset$ the diameter of the set K is denoted by

$$h_K := \inf_{K \subset S(x,\delta)} \delta \tag{3.5}$$

and the diameter of the biggest ball contained in the set K is denoted by

$$\rho_K := \sup_{S(x,\delta) \subset K} \delta \tag{3.6}$$

(see Figure 3.1). These values are used in the following lemma, which will be fundamental in the analysis of the FEM:

Lemma 5 *Let be $K \subset \mathbb{R}^n$ bounded with $int(K) \neq \emptyset$. Then there is a constant $C > 0$ depending on the set K with*

$$|B| \leq C \; h_{\Psi[K]} \tag{3.7}$$

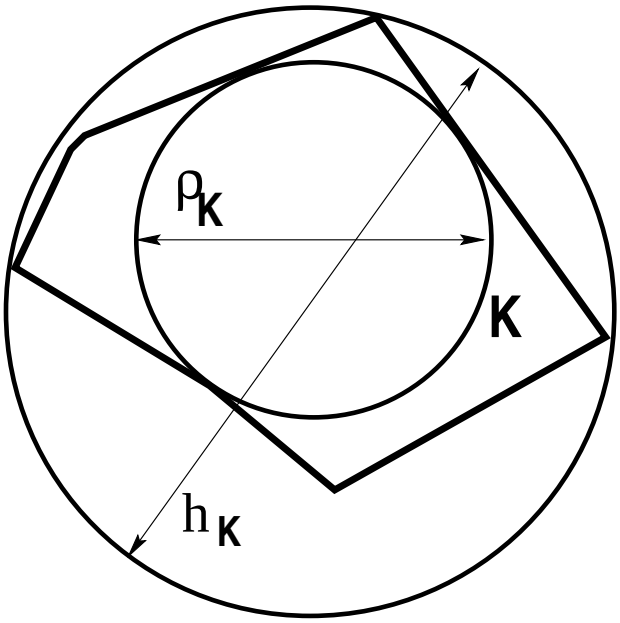

Figure 3.1: The diameter of the set K and the radius of the biggest ball in the set K.

and

$$|B^{-1}| \leq C \, \frac{1}{\rho_{\Psi[K]}} \tag{3.8}$$

for all affine transformation $\Psi : \mathbb{R}^n \to \mathbb{R}^n$ *defined by*

$$\Psi x := Bx + b \text{ for all } x \in \mathbb{R}^n \tag{3.9}$$

with $b \in \mathbb{R}^n$, $B \in \mathbb{R}^{n \times n}$ *and* $det(B) \neq 0$. *The value* $h_{\Psi[K]}$ *denotes the diameter of the set* $\Psi[K]$ *defined by equation (3.5) and the value* $\rho_{\Psi[K]}$ *denotes the diameter of the biggest ball in the set* $\Psi[K]$ *defined by equation (3.6).*

Proof: See Ciarlet [27] •

Remark: The inverse transformation Ψ^{-1} of the transformation Ψ defined by equation (3.9) is given by

$$\Psi^{-1}x = B^{-1}(x - b) \text{ for all } x \in \mathbb{R}^n \, . \tag{3.10}$$

3.2.1 Sobolev Spaces

In this chapter some Sobolev spaces are used, see Adams [2]: Let $n \in \{1, 2, 3\}$ be a spatial dimension, $m \in \mathbb{N}_0$ and $1 \leq q \leq \infty$. In addition $\Omega \subset \mathbb{R}^n$

denotes a domain, i.e. Ω is a bounded, open and connected subset of the real Euclidean space $\mathbb{R}^n$ with a Lipschitz–continuous boundary.

For a multi–index $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0^n$ it is set

$$|\alpha| := \sum_{i=1}^{n} \alpha_i \ . \tag{3.11}$$

For all functions $v : \Omega \to \mathbb{R}$ and all multi–indices $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0^n$ the function

$$D^\alpha v := \frac{\partial^{|\alpha|} v}{\partial^{\alpha_1} x_1 \partial^{\alpha_2} x_2 \cdots \partial^{\alpha_n} x_n} \tag{3.12}$$

denotes the α-th partial derivative of the function v being taken in the sense of distributions. The Sobolev space $W^{m,q}(\Omega)$ is defined by

$$W^{m,q}(\Omega) := \{v : \Omega \to \mathbb{R} \mid \int_\Omega |D^\alpha v|^q \, dx < \infty, \alpha \in \mathbb{N}_0^n, |\alpha| \leq m\} \tag{3.13}$$

if $q < \infty$ and by

$$W^{m,\infty}(\Omega) := \{v : \Omega \to \mathbb{R} \mid \operatorname*{ess\,sup}_{x \in \Omega} |D^\alpha v(x)| < \infty, \alpha \in \mathbb{N}_0^n, |\alpha| \leq m\} \tag{3.14}$$

if $q = \infty$, where 'ess sup' denotes the essential supreme. The Sobolev space $W^{m,q}(\Omega)$ is the set of all functions on the domain Ω whose derivatives up to order m have a finite integral of their q-th power (have a finite essential supreme if $q = \infty$). On the space $W^{m,q}(\Omega)$ the semi-norm

$$|v|_{m,q,\Omega} := \begin{cases} (\sum\limits_{|\alpha|=m} \int_\Omega |D^\alpha v|^q \, dx)^{\frac{1}{q}} & \text{if } q < \infty \\ \sup\limits_{|\alpha|=m} (\operatorname*{ess\,sup}_{x \in \Omega} |D^\alpha v(x)|) & \text{if } q = \infty \end{cases} \tag{3.15}$$

and the norm

$$\|v\|_{m,q,\Omega} := \begin{cases} (\sum\limits_{k=0}^{m} |v|_{k,q,\Omega}^q)^{\frac{1}{q}} & \text{if } q < \infty \\ \sup\limits_{0 \leq k \leq m} |v|_{k,\infty,\Omega} & \text{if } q = \infty \end{cases} \tag{3.16}$$

are used. For all $m \in \mathbb{N}_0$ and all $1 \leq q \leq \infty$ the space $(W^{m,q}(\Omega), \|.\|_{m,q,\Omega})$ is a Banach space. Since the case $q = 2$ is of special interest it is usual to drop the index expressing $q = 2$. Therefore the notations

$$\begin{array}{rcl} H^m(\Omega) & := & W^{m,2}(\Omega) \\ \|.\|_{m,\Omega} & := & \|.\|_{m,2,\Omega} \\ |.|_{m,\Omega} & := & |.|_{m,2,\Omega} \end{array} \tag{3.17}$$

for all $m \in \mathbb{N}_0$ are used. The space $(H^m(\Omega), \|.\|_{m,\Omega})$ is a Hilbert space for all $m \in \mathbb{N}_0$. Keep in mind that for all $1 \leq q \leq \infty$ the following identities hold:

$$\begin{array}{rcl} |.|_{0,q,\Omega} & = & \|.\|_{0,q,\Omega} \\ |.|_{0,\Omega} & = & \|.\|_{0,\Omega} \ . \end{array} \tag{3.18}$$

For all $m \in \mathbb{N}_0$ the set $C^m(\Omega)$ denotes the vector space of the real valued and m-times continuously differentiable functions on the domain Ω. It is $C^m(\Omega) \subset W^{m,\infty}(\Omega)$. The norm of the space $C^m(\Omega)$ is the $\|.\|_{m,\infty,\Omega}$-norm of the Sobolev space $W^{m,\infty}(\Omega)$. Later the following embedding theorem will be used, see Adams [2]:

Theorem 5 *For all $1 \leq q \leq \infty$, $m \in \mathbb{N}_0$ and $s > \frac{n}{q}$ it is*

$$W^{m+s,q}(\Omega) \subset C^m(cl(\Omega)) \subset W^{m,\infty}(\Omega) \ . \tag{3.19}$$

Proof: See Adams [2] •

Estimates for the modification of the Sobolev norm are needed if the domain is transformed by an affine transformation, see Ciarlet [27]. In the following theorem as well as in the further terms it is set $1/\infty := 0$.

Theorem 6 *Let be $1 \leq q \leq \infty$, $m \in \mathbb{N}_0$. There is a constant $C > 0$ depending on the domain Ω with the following property: For all affine transformations $\Psi : \mathbb{R}^n \to \mathbb{R}^n$ defined by*

$$\Psi x := Bx + b \text{ for all } x \in \mathbb{R}^n \tag{3.20}$$

with $B \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$ and $det(B) \neq 0$ hold: If $v \in W^{m,q}(\Psi[\Omega])$ then $v \circ \Psi \in W^{m,q}(\Omega)$ and it is

$$|v \circ \Psi|_{m,q,\Omega} \leq C|B|^m |det(B)|^{-\frac{1}{q}} |v|_{m,q,\Psi[\Omega]} \ . \tag{3.21}$$

If $v \in W^{m,q}(\Omega)$ then $v \circ \Psi^{-1} \in W^{m,q}(\Psi[\Omega])$ and it is

$$|v \circ \Psi^{-1}|_{m,q,\Psi[\Omega]} \leq C|B^{-1}|^m |det(B)|^{\frac{1}{q}} |v|_{m,q,\Omega} \ . \tag{3.22}$$

Proof: See Ciarlet [27] •

Remark: For the norm in the space $H^0(\Omega)$ a stronger result than inequalities (3.21) and (3.22) can be proved. By applying the substitution rule one obtains for all functions $v \in H^0(\Omega)$ and all affine transformations Ψ defined by equation (3.20):

$$|det(B)| \cdot |v \circ \Psi|^2_{0,\Omega} = |v|^2_{0,\Psi[\Omega]} \ . \tag{3.23}$$

3.2.2 Product Spaces

Let $d \in \mathbb{N}$, $1 \leq q \leq \infty$, $m \in \mathbb{N}_0$ and $\mathcal{T}$ be a finite family of pairwise disjoint domains in $\mathbb{R}^n$. The product space $W^{m,q}(\mathcal{T})^d$ is defined by

$$W^{m,q}(\mathcal{T})^d := \{(v_i)_{i=1,d} |\ v_i|_T \in W^{m,q}(T), T \in \mathcal{T}, i = 1, \ldots, d\} \tag{3.24}$$

It is the space of all $\mathbb{R}^d$-valued functions on the set $\bigcup \mathcal{T}$ whose components belong to the Sobolev space $W^{m,q}(T)$ for all sets $T \in \mathcal{T}$. The semi-norm

$$|v|_{m,q,\mathcal{T}} := \begin{cases} (\sum\limits_{1 \leq i \leq d; T \in \mathcal{T}} |v_i|^q_{m,q,T})^{\frac{1}{q}} & \text{if } q < \infty \\ \max\limits_{1 \leq i \leq d; T \in \mathcal{T}} |v_i|_{m,\infty,T} & \text{if } q = \infty \end{cases} \tag{3.25}$$

and the norm

$$\|v\|_{m,q,\mathcal{T}} := \begin{cases} (\sum\limits_{k=0}^{m} |v|^q_{k,q,\mathcal{T}})^{\frac{1}{q}} & \text{if } q < \infty \\ \max\limits_{0 \leq k \leq m} |v|_{k,\infty,\mathcal{T}} & \text{if } q = \infty \end{cases} \tag{3.26}$$

for all functions $v = (v_i)_{i=1,d} \in W^{m,q}(\mathcal{T})^d$ are used. The vector space $W^{m,q}(\mathcal{T})^d$ with the norm $\|.\|_{m,q,\mathcal{T}}$ is a Banach space. The notations and properties of the Sobolev spaces are extended to the product spaces (especially the embedding Theorem 5).

For $d = 1$ it is set $W^{m,q}(\mathcal{T}) := W^{m,q}(\mathcal{T})^1$. If the family $\mathcal{T}$ has a single element, e.g. $\mathcal{T} = \{\Omega\}$, it is set

$$W^{m,q}(\Omega)^d := W^{m,q}(\{\Omega\})^d \text{ and } \|.\|_{m,q,\Omega} := \|.\|_{m,q,\{\Omega\}} \ . \tag{3.27}$$

The notations that were introduced in the previous Chapter 2 are adopted in this chapter. Especially the upper index $'*'$ of Banach spaces denotes still the dual space (e.g. the space $H^1(\Omega)^*$ is the dual space of $H^1(\Omega)$). The lower index of norms indicates Sobolev space norms or norms for operator spaces defined by equation (2.6). They cannot be mixed up as the types of the indices are different.

3.2.3 Basic Error Estimates

Now two theorems are quoted that are essential to prove the convergence order of finite element approximations. They base on the famous Bramble-Hilbert-Lemma, see Bramble [19].

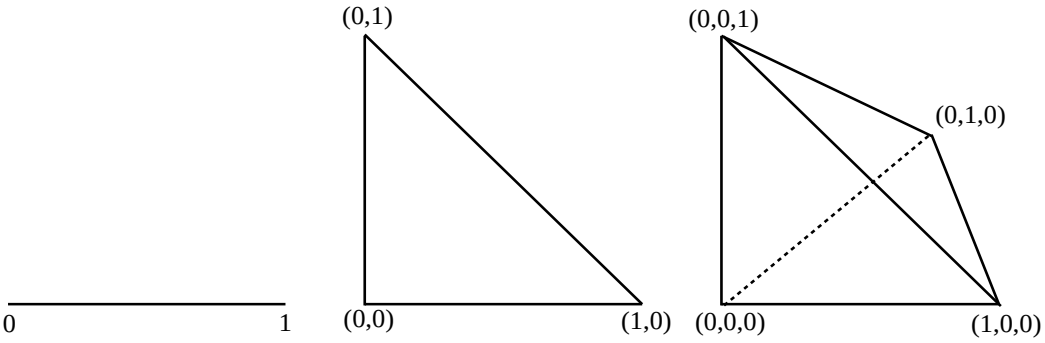

Figure 3.2: The 1-simplex, 2-simplex and 3-simplex.

The set $T^0 \subset \mathbb{R}^n$ denotes the n-simplex defined by

$$T^0 := \begin{cases} [0,1] & \text{if } n = 1 \\ \{(x_1^0, x_2^0) | x_1^0, x_2^0 \geq 0, x_1^0 + x_2^0 \leq 1\} & \text{if } n = 2 \\ \{(x_1^0, x_2^0, x_3^0) | x_1^0, x_2^0, x_3^0 \geq 0, & \text{if } n = 3 \\ \qquad\qquad x_1^0 + x_2^0 + x_3^0 \leq 1\} & \end{cases} . \tag{3.28}$$

The n-simplexes are plotted in Figure 3.2. In the following locations and coordinates which are in the n-simplex T^0 are marked with the upper index 0. The intersections of the n-simplex T^0 with the hyperspaces

$$x_k^0 = 0 \tag{3.29}$$

for all spatial directions $k = 1, \ldots, n$ and

$$\sum_{i=1}^{n} x_i^0 = 1 \tag{3.30}$$

are called the $n+1$ faces of the n-simplex T^0. In the following the n-simplex T^0 and its interior $int(T^0)$ are not distinguished.

The set P_k denotes the space of all polynomials on the set $\mathbb{R}^n$ with maximal order k. In case of $n = 3$ it is

$$P_k := span\{x_1^{k_1} x_2^{k_2} x_3^{k_3} | k_1, k_2, k_3 \in \mathbb{N}_0 \text{ and } k_1 + k_2 + k_3 \leq k\} . \tag{3.31}$$

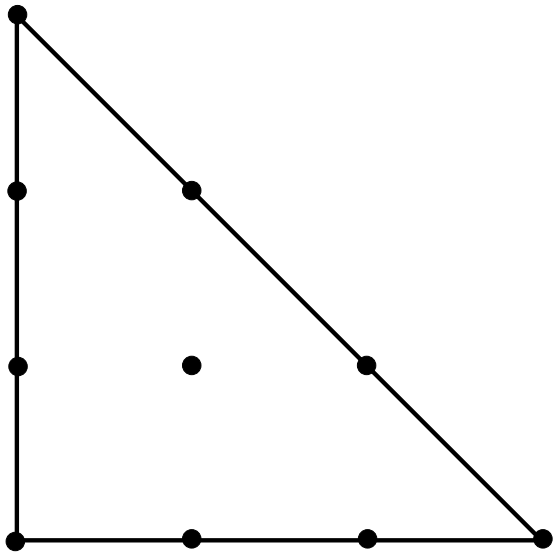

Figure 3.3: The local degrees of freedom for order 3 on the 2-simplex.

The set $X^{0,k} := \{x_i^{0,k}\}_{i=1,d_k} \subset T^0$ defined by

$$X^{0,k} := \begin{cases} \{\frac{k_1}{k} | k_1 \in \mathbb{N}_0, k_1 \leq k\} & \text{if } n = 1 \\ \{(\frac{k_1}{k}, \frac{k_2}{k}) | k_1, k_2 \in \mathbb{N}_0, k_1 + k_2 \leq k\} & \text{if } n = 2 \\ \{(\frac{k_1}{k}, \frac{k_2}{k}, \frac{k_3}{k}) | k_1, k_2, k_3 \in \mathbb{N}_0, & \text{if } n = 3 \\ \qquad\qquad k_1 + k_2 + k_3 \leq k\} & \end{cases} \tag{3.32}$$

denotes the *set of the local degrees of freedom* of order k, see Figure 3.3. A polynomial of order k is uniquely defined by its values at the local degrees of freedom, see Nicolaides [48].

The linear operator $\mathcal{I}^k : C^0(T^0) \to P_k$ defined by $v \to \mathcal{I}^k v$ for all $v \in C^0(T^0)$, where $\mathcal{I}^k v \in P_k$ is the unique solution of the Lagrangean interpolation problem

$$\mathcal{I}^k v(x_i^{0,k}) = v(x_i^{0,k}) \text{ for all } i = 1, \ldots, d_k \ , \tag{3.33}$$

is called the *local interpolation operator* of order k. Taking Theorem 5 the local interpolation operator $\mathcal{I}^k$ is defined on the space $W^{m,q}(T^0) \subset C^0(T^0)$ if $m > \frac{n}{q}$.

The next theorem gives an estimate of the interpolation error, see Ciarlet [27]:

Theorem 7 (Ciarlet 1972) *Let $1 \leq q \leq \infty$ and $\mathcal{I}^k$ be the local interpolation operator of order $k > \frac{n}{q} - 1$. There is a constant $C > 0$ with*

$$|v - \mathcal{I}^k v|_{m,q,T^0} \leq C \ |v|_{k+1,q,T^0} \tag{3.34}$$

for all functions $v \in W^{k+1,q}(T^0)$ and all $0 \leq m \leq k$.

Proof: See Ciarlet [27] •

A numerical quadrature scheme $Q^l : C^0(T^0) \to \mathbb{R}$ on the n-simplex T^0 defined by

$$Q^l(\varphi) := \sum_{i=1}^{q_l} \omega_i^{0,l} \varphi(y_i^{0,l}) \tag{3.35}$$

for all $\varphi \in C^0(T^0)$ approximates the integral $\int_{T^0} \varphi \ dx^0$ by the finite sum $Q^l(\varphi)$. The positive values $\{\omega_i^{0,l}\}_{i=1,q_l} \subset \mathbb{R}_+$ are called integration weights and the points $\{y_i^{0,l}\}_{i=1,q_l} \subset T^0$ are called integration nodes.

Definition 4 *The quadrature scheme $Q^l : C^0(T^0) \to \mathbb{R}$ is called exact of order l if*

$$Q^l(p) = \int_{T^0} p \ dx^0 \ \textit{for all} \ p \in P_l \ . \tag{3.36}$$

In the following the upper index l of a quadrature scheme Q^l indicates that quadrature scheme Q^l is exact of order l in the sense of this definition. Keep in mind that in this definition as well as in the following the quadrature scheme Q^l may exactly integrate polynomials with higher order than l and may also be the exact integration operator.

On the 1-simplex the well–known Gaussian quadrature scheme is the optimal quadrature scheme since it uses the minimal number m of integration nodes to construct a quadrature scheme that is exact of order $2m - 1$, see Davis [28]. For the 2-simplex and 3-simplex the construction of optimal quadrature schemes takes more effort, e.g. see Guessab [41]. Difficulties arise from the requirements that the integration weights have to be positive and the location of the integration nodes should fulfil some symmetry properties. When implemented on a computer the product scheme of Gaussian quadrature schemes on the unit cube $[0,1]^n$ is transformed into the simplex by changing the variables, see Zienkiewicz [68]. Since the transformation is not affine there is a loss of accuracy. Moreover the integration nodes are not symmetrically spaced in the simplex. In spite of this these quadrature schemes are very popular since they are very easy to implement.

If $Q^l : C^0(T^0) \to \mathbb{R}$ is a given quadrature scheme its error functional $E^l : C^0(T^0) \to \mathbb{R}$ is defined by

$$E^l(\varphi) := \int_{T^0} \varphi \, dx^0 - Q^l(\varphi) \tag{3.37}$$

for all functions $\varphi \in C^0(T^0)$. It is obvious that the quadrature scheme Q^l defined by equation (3.35) is a continuous, linear functional on $C^0(T^0)$, i.e. $Q^l \in C^0(T^0)^*$. Therefore it is also $E^l \in C^0(T^0)^*$. Taking Theorem 5 the linear functionals Q^l and E^l are defined on the Sobolev space $W^{k,\infty}(T^0)$ for all $k \geq 1$. Keep in mind that $E^l(p) = 0$ for all $p \in P_l$ if and only if the quadrature scheme Q^l is exact of order l in the sense of Definition 4.

Analogously to the estimate of the interpolation error in Theorem 7 there is an estimate of the error by a quadrature scheme, see Ciarlet [26]:

Theorem 8 (Ciarlet 1972) *Let be $E^l \in C^0(T^0)^*$ with $E^l(p) = 0$ for all $p \in P_l$ and $k \in \mathbb{N}$ with $l \geq k \geq 1$. Then there is a real value $C > 0$ with*

$$|E^l(f \cdot p)| \leq C(|f|_{l-k+1,\infty,T^0}|p|_{1,T^0} + |f|_{l-k+2,\infty,T^0}|p|_{0,T^0}) \tag{3.38}$$

and

$$|E^l(f \cdot \frac{\partial p}{\partial x_i^0})| \leq C|f|_{l-k+2,\infty,T^0}|p|_{1,T^0} \tag{3.39}$$

for all functions $f \in W^{k,\infty}(T^0)$, all polynomials $p \in P_k$ and all spatial directions $i = 1, \ldots, n$.

Proof: See Ciarlet [26] •

Remark: In Lemma 5, Theorem 6, Theorem 7 and Theorem 8 the values of the constants C are unknown.

3.3 The Variational Problem

This section introduces a class of variational problems arising from the non–linear version of the model boundary value problem (1.1) presented in the introducing Chapter 1. It will be verified that these variational problems are well–posed in the sense of Definition 1.

In the rest of this chapter $n \in \{1, 2, 3\}$ denotes the spatial dimension and $\Omega \subset \mathbb{R}^n$ denotes a fixed domain.

To get simpler formulas the following notation is introduced: For all functions $u \in W^{1,q}(\Omega)$ $(1 \leq q \leq \infty)$ the vector valued function $u_;$ denotes the function of $n+1$ components created by the function u and its spatial derivatives:

$$u_; := (u_{;i})_{i=1,n+1} := (u, (\frac{\partial u}{\partial x_j})_{j=1,n}) \; . \tag{3.40}$$

It is $u_; \in W^{0,q}(\Omega)^{n+1}$ with

$$\begin{aligned} |u_;|_{0,q,\Omega} = \|u_;\|_{0,q,\Omega} = \|u\|_{1,q,\Omega} \qquad & \text{and} \\ |u_;|_{0,\Omega} = \|u_;\|_{0,\Omega} = \|u\|_{1,\Omega} \qquad & \text{if } q = 2 \end{aligned} \tag{3.41}$$

where the product space $W^{0,q}(\Omega)^{n+1}$ and its norm are defined by equations (3.24) and (3.26).

Definition 5 (Grosz) *Let be $G : \Omega \times \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$. The function G is called uniform positive definite with positivity bound $D > 0$, if*

- *for all vectors $\zeta \in \mathbb{R}^{n+1}$ the function $G(\zeta, .) : \Omega \to \mathbb{R}^{n+1}$ defined by $x \to G(\zeta, x)$ for all $x \in \Omega$ belongs to $H^0(\Omega)^{n+1}$*
- *for all vectors $x \in \Omega$ the function $G(., x) : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ defined by $\zeta \to G(\zeta, x)$ for all $\zeta \in \mathbb{R}^{n+1}$ belongs to $C^1(\mathbb{R}^{n+1})^{n+1}$ and the estimates*

$$\chi \cdot \partial G(\zeta, x)\xi \leq D|\chi||\xi| \tag{3.42}$$

and

$$\frac{1}{D}|\chi|^2 \leq \chi \cdot \partial G(\zeta, x)\chi \tag{3.43}$$

hold for all $x \in \Omega$ and all $\zeta, \chi, \xi \in \mathbb{R}^{n+1}$. In both inequalities (3.42) and (3.43) the real $(n+1) \times (n+1)$ matrix

$$\partial G(\zeta, x) = (\partial_j G_i(\zeta, x))_{i,j=1,n+1} := (\frac{\partial G_i}{\partial \zeta_j}(\zeta, x))_{i,j=1,n+1} \tag{3.44}$$

denotes the Jacobi–matrix of the function G with respect to the first $n+1$ variables ζ at the location (ζ, x) for all $x \in \Omega$ and $\zeta = (\zeta_i)_{i=1,n+1}$.

The mapping $F : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ defined by

$$< v, F(u) > := \int_\Omega v_; \cdot G(u_;, .) \, dx \tag{3.45}$$

for all $u, v \in H^1(\Omega)$ is called the operator generated by the kernel G.

In the equations (3.42), (3.43) and (3.45)

$$\chi \cdot \xi := \sum_{i=1}^{n+1} \chi_i \xi_i \tag{3.46}$$

denotes the scalar product of the vectors $\chi = (\chi_i)_{i=1,n+1}$ and $\xi = (\xi_i)_{i=1,n+1}$ in $\mathbb{R}^{n+1}$ and

$$\partial G(\zeta, x)\chi := (\sum_{j=1}^{n+1} \partial_j G_i(\zeta, x)\chi_j)_{i=1,n+1} \tag{3.47}$$

denotes the matrix-vector product of the Jacobi matrix of the function G with the vector $\chi = (\chi_i)_{i=1,n+1} \in \mathbb{R}^{n+1}$.

Remark 1: The positivity bound D is not unique. Every constant greater than D can be used as well.

Remark 2: The operator F defined by equation (1.3) in the introducing Chapter 1 is generated by the kernel

$$G(\zeta, x) := (b(x)\zeta_1 - f(x), a(x)\zeta_2, \ldots, a(x)\zeta_{n+1}) \tag{3.48}$$

for all $x \in \Omega$ and $\zeta = (\zeta_i)_{i=1,n+1} \in \mathbb{R}^{n+1}$. The kernel G is uniform positive definite with positivity bound

$$D := \max(C, \frac{1}{c}) \tag{3.49}$$

if $a, b, f \in H^0(\Omega)$ and $C \geq a(x) \geq c > 0$ and $C \geq b(x) \geq c > 0$ for all $x \in \Omega$.

In the following the finite element approximation of the solution $u \in H^1(\Omega)$ of the non–linear variational problem

$$< v, F(u) >= 0 \text{ for all } v \in H^1(\Omega) \tag{3.50}$$

is discussed when the operator F is generated by a uniform positive definite kernel G. This variational problem is produced by the weak formulation of the homogeneous *Neumann boundary value problem*, see Quarteroni [52]: find a solution $u : \Omega \to \mathbb{R}$

$$\begin{aligned} G_1(u_{;}(x), x) - \sum_{i=1}^{n} \frac{\partial G_{i+1}(u_{;}(x), x)}{\partial x_i} &= 0 \quad \text{for all } x \in \Omega \\ \sum_{i=1}^{n} n_i(x) G_{i+1}(u_{;}(x), x) &= 0 \quad \text{for all } x \in \partial\Omega \ . \end{aligned} \tag{3.51}$$

The mapping $x \to (n_i(x))_{i=1,n}$ denotes the outer unit field of the boundary $\partial\Omega$ of the domain Ω. The second condition prescribes that the normal component of the vector field $(G_2, \ldots, G_{n+1})$ for the sought solution u has to

vanish at the boundary of the domain Ω. It is a boundary condition of the Neumann type.

The first equation is a partial differential equation for the sought, scalar function u. By using the chain rule (it is assumed that the function u and the kernel G are smooth enough) this equation is equal to

$$G_1 - \sum_{i=1}^{n} \frac{\partial G_{i+1}}{\partial x_i} - \sum_{i=1}^{n} \partial_1 G_{i+1} \frac{\partial u}{\partial x_i} - \sum_{i,j=1}^{n} \partial_{j+1} G_{i+1} \frac{\partial^2 u}{\partial x_i \partial x_j} = 0 \qquad (3.52)$$

where the argument $(u_{,}(x), x)$ of the kernel G is dropped. As it follows from condition (3.43) that $\partial_{i+1} G_{i+1} > 0$ for all spatial directions $i = 1, \ldots, n$ the partial differential equation (3.52) has the order two. As the matrix $(\partial_{i+1} G_{i+1})_{i,j=1,n}$ is even positive definite the partial differential equation has the characteristics of an elliptic differential equation.

By modifying slightly Definition 5 and the following discussion systems of n_c coupled Neumann boundary value problems for the sought solution $u \in H^1(\Omega)^{n_c}$ can be considered. The value $n_c \in \mathbb{N}$ denotes the number of components of the solution. Mainly a summation over the solution components has to be added in the proofs. Especially the kernel G is now a $\mathbb{R}^{(n+1)\times n_c}$–valued function, more exactly $G : \mathbb{R}^{(n+1)\times n_c} \times \Omega \to \mathbb{R}^{(n+1)\times n_c}$. Other important modifications are the consideration of non–homogeneous Neumann boundary conditions, which are introduced by additional boundary integrals in the definition of the operator F. Moreover Dirichlet boundary conditions can be introduced by restricting the generated operator F to a suitable subspace of the Sobolev space $H^1(\Omega)$, see Quarteroni [52], or by using Lagrangean multiplier, see Babuška [4]. It has to be pointed out that the same results as for the model problem can be verified for these modifications by using the well–known techniques of the analysis of FEMs for the corresponding linear variational problems.

At first it has to be guaranteed that the operator generated by an uniform positive definite kernel G is well–posed and the variational problem (3.50) has exactly one solution:

Theorem 9 (Grosz) *Let G be uniform positive definite with positivity bound D and F the operator generated by the kernel G defined by equation (3.45). Then it holds:*

- *For all fixed functions $u \in H^1(\Omega)$ the linear functional $F(u) : H^1(\Omega) \to \mathbb{R}$ defined by $v \to < v, F(u) >$ for all $v \in H^1(\Omega)$ belongs to the dual space $H^1(\Omega)^*$ of the Sobolev space $H^1(\Omega)$.*

- *The operator $F : H^1(\Omega) \to H^1(\Omega)^*$ defined by $u \to F(u)$ for all $u \in H^1(\Omega)$ is well–posed with condition number D^2.*
- *The variational problem (3.50) has exactly one solution $u \in H^1(\Omega)$.*

Proof: Let be $u_1, u_2, w \in H^1(\Omega)$. It is for all $x \in \Omega$:

$$\begin{aligned} & w_{;}(x) \cdot [G(u_{1;}(x), x) - G(u_{2;}(x), x)] \\ = & \int_0^1 w_{;}(x) \cdot \partial G(t \cdot u_{1;}(x) + (1-t) \cdot u_{2;}(x), x) \\ & \qquad\qquad [u_{1;}(x) - u_{2;}(x)] dt \\ \leq & D|u_{1;}(x) - u_{2;}(x)||w_{;}(x)| \end{aligned} \tag{3.53}$$

where in the last estimation the condition (3.42) of Definition 5 is used.

By setting $u_1 := u$ and $u_2 := 0$ it follows from inequality (3.53) that

$$\begin{aligned} w_{;}(x) \cdot G(u_{;}(x), x) &\leq w_{;}(x) \cdot G(0, x) + D|u_{;}(x)| \; |w_{;}(x)| \\ &\leq |w_{;}(x)| \; (|G(0, x)| + D|u_{;}(x)|) \end{aligned} \tag{3.54}$$

for all $x \in \Omega$. After this inequality has been integrated over the domain Ω the following estimates can be made

$$\begin{aligned} < w, F(u) > &\overset{(3.45)}{=} \int_\Omega w_{;} \cdot G(u_{;}, .) \; dx \\ &\overset{(3.54)}{\leq} \int_\Omega |w_{;}| \; (|G(0, .)| + D|u_{;}|) \; dx \\ &\leq |w_{;}|_{0,\Omega} \; (|G(0, .)|_{0,\Omega} + D|u_{;}|_{0,\Omega}) \end{aligned} \tag{3.55}$$

where in the last estimate the Cauchy-Schwartz-inequality in the Hilbert space $H^0(\Omega)$ is used. As from equation (3.41) it is $|w_{;}|_{0,\Omega} = \|w\|_{1,\Omega}$ it can be shown that the norm of $F(u)$ is bounded:

$$\begin{aligned} \|F(u)\|_{H^1(\Omega)^*} &\overset{(2.11)}{=} \sup_{w \in H^1(\Omega)} \frac{< w, F(u) >}{\|w\|_{1,\Omega}} \\ &\overset{(3.55)}{\leq} |G(0, .)|_{0,\Omega} + D\|u\|_{1,\Omega} \; . \end{aligned} \tag{3.56}$$

This proves that the functional $F(u)$ belongs to the dual space $H^1(\Omega)^*$.

To prove the second and third claim of the theorem the propositions of Theorem 1 are verified for the operator F in the Hilbert space $H^1(\Omega)$:

By integrating the inequality (3.53) over the domain Ω and using the Cauchy-Schwartz inequality in the Hilbert space $H^0(\Omega)$ one gets

$$\begin{array}{ll} & < w, F(u_1) - F(u_2) > \\ \overset{(3.45)}{=} & \int_\Omega w_{;} \cdot [G(u_{1;}, .) - G(u_{2;}, .)]dx \\ \overset{(3.53)}{\leq} & D \int_\Omega |u_{1;} - u_{2;}||w_{;}|dx \\ \leq & D|u_{1;} - u_{2;}|_{0,\Omega}|w_{;}|_{0,\Omega} \\ \overset{(3.41)}{=} & D\|u_1 - u_2\|_{1,\Omega}\|w\|_{1,\Omega} \ . \end{array} \tag{3.57}$$

So condition (2.22) of Theorem 1 has been proved:

$$\|F(u_1) - F(u_2)\|_{H^1(\Omega)^*} \leq D\|u_1 - u_2\|_{1,\Omega} \ . \tag{3.58}$$

If $u_1, u_2 \in H^1(\Omega)$ it is for all $x \in \Omega$

$$\begin{array}{ll} & [u_{1;}(x) - u_{2;}(x)] \cdot [G(u_{1;}(x), x) - G(u_{2;}(x), x)] \\ = & \int_0^1 [u_{1;}(x) - u_{2;}(x)] \cdot \partial G(t \cdot u_{1;}(x) + (1-t) \cdot u_{2;}(x), x) \\ & \qquad\qquad [u_{1;}(x) - u_{2;}(x)]dt \\ \geq & \frac{1}{D}|u_{1;}(x) - u_{2;}(x)|^2 \end{array} \tag{3.59}$$

where in the last estimation the condition (3.43) of Definition 5 is used. By integrating over the domain Ω it is

$$\begin{array}{ll} & < u_1 - u_2, F(u_1) - F(u_2) > \\ \overset{(3.45)}{=} & \int_\Omega [u_{1;} - u_{2;}] \cdot [G(u_{1;}, .) - G(u_{2;}, .)]dx \\ \overset{(3.59)}{\geq} & \frac{1}{D} \int_\Omega |u_{1;} - u_{2;}|^2 dx \\ = & \frac{1}{D}|u_{1;} - u_{2;}|^2_{0,\Omega} \\ \overset{(3.41)}{=} & \frac{1}{D}\|u_1 - u_2\|^2_{1,\Omega} \ . \end{array} \tag{3.60}$$

This verifies condition (2.23) of Theorem 1. From this theorem one obtains that the operator F is well–posed with condition number D^2 and the variational problem (3.50) has exactly one solution •

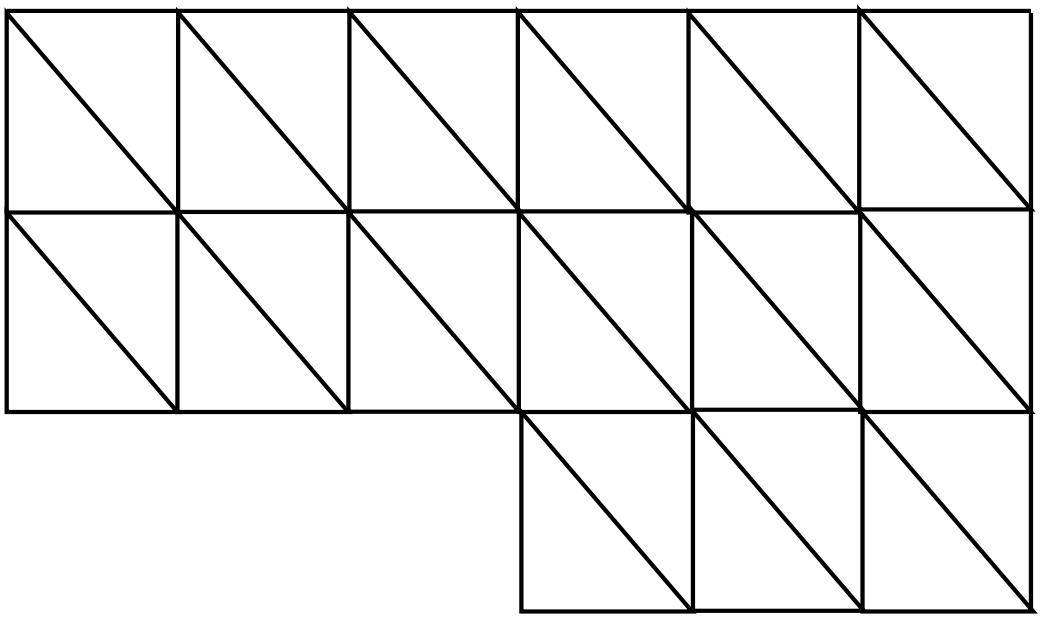

Figure 3.4: Example of a triangulation of a 2-dimensional domain.

3.4 The Finite Element Space

This section deals with the construction of the approximation space $V_h \subset H^1(\Omega)$ of the finite element method (FEM). The construction bases on a subdivision of the domain Ω into small subdomains, called *elements*. Essential results of this section are two theorems on the approximation properties of the finite element space basing on the application of Theorem 7 and Theorem 8. Here only simplex elements of a fixed polynomial order are considered. More general approaches are presented in Ciarlet [25].

The starting point is the triangulation of the domain Ω, see Figure 3.4:

Definition 6 *The family $\mathcal{T}_h$ of subsets of $\mathbb{R}^n$ is called a triangulation of the domain Ω, if the following conditions hold:*

1. *The family $\mathcal{T}_h$ is a subdivision of the domain Ω: $cl(\Omega) = \bigcup_{T \in \mathcal{T}_h} cl(T)$*
2. *The elements are disjoint: for all $T_1, T_2 \in \mathcal{T}_h$: $int(T_1) \cap int(T_2) = \emptyset$*
3. *The elements have an affine representation: for all $T \in \mathcal{T}_h$ there is a transformation $\Psi_T : T^0 \to T$ defined by*

$$\Psi_T x^0 := B_T x^0 + b_T \tag{3.61}$$

for all $x^0 \in T^0$ *with* $B_T \in \mathbb{R}^{n\times n}$, $det(B_T) \neq 0$ *and* $b_T \in \mathbb{R}^n$. T^0 *denotes the* n*-simplex defined by equation (3.28).*

4. *The elements are adjacent: any face of any* $T_1 \in \mathcal{T}_h$ *is either a subset of the boundary* $\partial\Omega$ *of the domain* Ω *or it is a face of an other* $T_2 \in \mathcal{T}_h$. *The faces of* $T \in \mathcal{T}_h$ *are the ranges of the faces of the* n*-simplex* T^0 *mapped by its parametrical representation* Ψ_T.

$T \in \mathcal{T}_h$ *is called element. The affine transformation* Ψ_T *defined by equation (3.61) is called the parametrical representation of the element* T. *The value*

$$h := \max_{T\in\mathcal{T}_h} h_T \tag{3.62}$$

names the mesh size of the triangulation $\mathcal{T}_h$ *and the value*

$$\sigma_h := \sup_{T\in\mathcal{T}_h} \frac{h_T}{\rho_T} \tag{3.63}$$

names its mesh quality, where the values h_T *and* ρ_T *denote the diameter of element* T *and the diameter of the biggest ball in the element* T *defined by equations (3.5) and (3.6), see Figure 3.1.*

Remark: In any case it is $\sigma_h \geq 1$ and $h \leq h_\Omega$. Taking Lemma 5 (with $K = T^0$) the mesh quality σ_h is mainly the maximal condition number of the matrices B_T over all elements T in the triangulation $\mathcal{T}_h$. Keep in mind that for the one dimensional case $n = 1$ it is $h_T = \rho_T$ for all elements T and therefore it is always $\sigma_h = 1$.

There are a lot of powerful program packages to generate triangulations of a given domain, e.g. see I–DEAS [44], PATRAN [50]. Figure 3.5 shows the subdivision of the 2-dimensional unit circle by I–DEAS. This example demonstrates that a triangulation in the sense of Definition 6 exists only for polygonal domains. If the boundary of the domain is curved the subdivision can only be an approximation of the domain. To improve the approximation of the boundary curved elements can be used. In the following discussion curved triangulations are not considered but the results can be adapted to the more general situation especially when using isoparametrical elements, see Ciarlet [25].

The behavior of a family of triangulations with mesh sizes having the unique accumulation point zero is analyzed. The notation with respect of the index h for the space V_h introduced in Chapter 2 is adopted to the family of triangulations $(\mathcal{T}_h)_{h>0}$. Analogously to the n-simplex T^0 the element T and its interior is not distinguished.

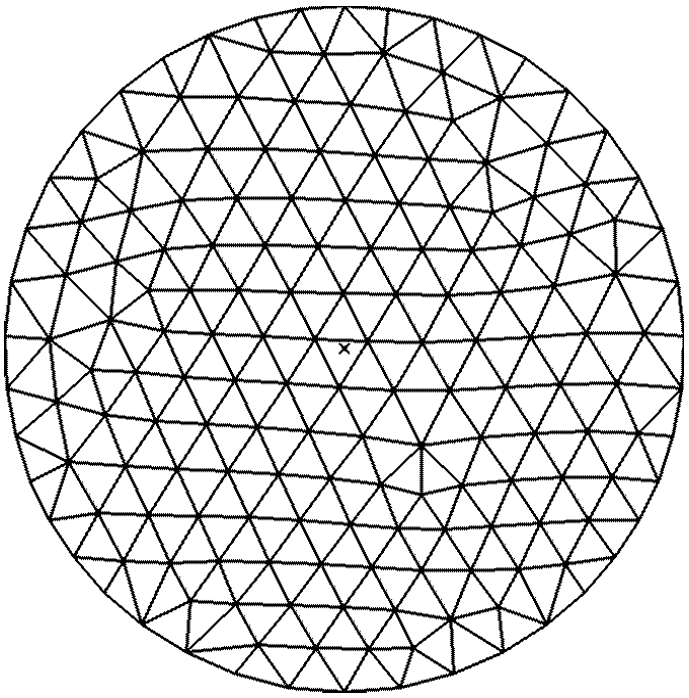

Figure 3.5: Triangulation of the unit circle by I–DEAS ($h \approx .15$, $\sigma_h \approx 4$).

For a triangulation $\mathcal{T}_h$ and order $k \in \mathbb{N}$ the space

$$V^{h,k} := \{v \in C^0(cl(\Omega))| \text{ for all } T \in \mathcal{T}_h : v|_T \circ \Psi_T^{-1} \in P_k\} \tag{3.64}$$

is called the *finite element space of order* k by the triangulation $\mathcal{T}_h$. As the transformations Ψ_T^{-1} are affine transformations the function $v|_T \circ \Psi_T^{-1}$ is a polynomial of order k if and only if the function $v|_T$ is a polynomial of order k. Therefore the space $V^{h,k}$ is the set of all continuous functions on the domain Ω which are piecewise polynomials of order k. The following lemma ensures that the space $V^{h,k}$ suites when discretizing the variational problem (3.50):

Lemma 6 *For any triangulation $\mathcal{T}_h$ of the domain Ω and $k \in \mathbb{N}$ it is*

$$V^{h,k} \subset H^1(\Omega) \; . \tag{3.65}$$

Proof: See Ciarlet [25] •

In addition to an approximation space V_h an approximation for the integral in the functional equation (3.50) has to be introduced as the integral cannot be evaluated on a computer. If a local quadrature scheme Q^l on the n-simplex is given this scheme can be extended to a quadrature scheme over

the domain Ω in the following way: For any function $\varphi \in C^0(cl(\Omega))$ it is by the substitution rule:

$$\begin{aligned} \int_\Omega \varphi \, dx &= \sum_{T \in \mathcal{T}_h} \int_{\Psi_T[T^0]} \varphi \, dx \\ &= \sum_{T \in \mathcal{T}_h} |det(B_T)| \int_{T^0} \varphi \circ \Psi_T \, dx^0 \ . \end{aligned} \tag{3.66}$$

Therefore a quadrature scheme Q^{h,Q^l} approximating the integral $\int_\Omega \varphi \, dx$ is introduced by the quadrature scheme Q^l on the n-simplex T^0 by setting

$$Q^{h,Q^l}(\varphi) := \sum_{T \in \mathcal{T}_h} |det(B_T)| Q^l(\varphi \circ \Psi_T) \tag{3.67}$$

for all functions $\varphi \in C^0(cl(\Omega))$.

The discrete variational problem which is solved to get an approximative solution for the variational problem (3.50) is now: find a discrete solution $u_h \in V^{h,k}$ with

$$Q^{h,Q^l}(v_{h;} \cdot G(u_{h;}, .)) = 0 \text{ for all } v_h \in V^{h,k} \ . \tag{3.68}$$

It has to be verified that the FEM approximations u_h converge to the sought solution u if the mesh size h goes to zero. A corresponding result can be obtained from Corollary 1 in the previous Chapter 2. Therefore now the propositions (2.48) and (2.49) of Corollary 1 have to be verified where the global interpolation operator $\mathcal{I}^{h,k}$ in the space $V^{h,k}$ (see below) is used for the operator $\mathcal{I}_h$:

The *set of the global degrees of freedom* for order k by the triangulation $\mathcal{T}_h$ denoted by

$$X^{h,k} := \{\Psi_T x_i^{0,k} | T \in \mathcal{T}_h, x_i^{0,k} \in X^{0,k}\} \tag{3.69}$$

are the images of the local degrees of freedom $X^{0,k}$ defined by equation (3.32) under the parametrical representations of the elements in the triangulation $\mathcal{T}_h$. The number of points in the set of the global degrees of freedom $X^{h,k}$ is denoted by the integer value $d^{h,k}$. The global degrees of freedom are enumerated from 1 to $d^{h,k}$:

$$X^{h,k} = \{x_i^{h,k}\}_{i=1,d^{h,k}} \ . \tag{3.70}$$

For all elements $T \in \mathcal{T}_h$ the key list $\pi^{h,k}(T) \in \mathbb{N}^{d_k}$ joins the local degrees of freedom $X^{0,k}$ in the n-simplex T^0 to those global degrees of freedom belonging to element T. Exactly the list $\pi^{h,k}(T)$ is defined by

$$\Psi_T x_j^{0,k} = x^{h,k}_{\pi_j^{h,k}(T)} \tag{3.71}$$

for all $j = 1, \ldots, d_k$, i.e. the number $\pi_j^{h,k}(T)$ is the id number of the point assigned to the j-th local degree of freedom via the parametrical representation Ψ_T of element T.

In practical implementations the $\pi^{h,k}$-list is used to gather values given at the global degrees of freedom for the local degrees of freedom. The following lemma shows that an interpolation problem in the space $V^{h,k}$ at the global degrees of freedom can be broken into many interpolation problems in the space of polynomials P_k at the n-simplex T^0 by using the $\pi^{h,k}$-list:

Lemma 7 *Let $\mathcal{T}_h$ be a triangulation, $k \in \mathbb{N}$ and $\{v_i\}_{i=1,d^{h,k}} \in \mathbb{R}^{d^{h,k}}$. Then the global interpolation problem*

$$v_h(x_i^{h,k}) = v_i \text{ for } i = 1, \ldots, d^{h,k} \tag{3.72}$$

at the global degrees of freedom has exactly one solution $v_h \in V^{h,k}$. For all elements $T \in \mathcal{T}_h$ the restriction $v_h|_T$ of the function v_h onto the element T is given by

$$v_h|_T := v_T \circ \Psi_T^{-1} \tag{3.73}$$

where the polynomial $v_T \in P_k$ is the unique solution of the interpolation problem

$$v_T(x_j^{0,k}) = v_{\pi_j^{h,k}(T)} \text{ for all } j = 1, \ldots, d_k \tag{3.74}$$

on n-simplex T^0.

Proof: see Nicolaides [49]. In the proof the location of the local degrees of freedom as defined in equation (3.32) is essential to ensure that the function v_h defined by the equations (3.73) and (3.74) belongs to the space $V^{h,k}$ •

The linear operator $\mathcal{I}^{h,k} : C^0(cl(\Omega)) \to V^{h,k}$ defined by the unique solution $\mathcal{I}^{h,k}v \in V^{h,k}$ of the global interpolation problem

$$\mathcal{I}^{h,k}v(x_i^{h,k}) = v(x_i^{h,k}) \text{ for } i = 1, \ldots, d^{h,k} \tag{3.75}$$

for all $v \in C^0(cl(\Omega))$ is called the *global interpolation operator of order k* by the triangulation $\mathcal{T}_h$. From Lemma 7 and the definition of the local interpolation operator $\mathcal{I}^k$ in equation (3.33) the global interpolation operator can be represented in the following manner:

$$(\mathcal{I}^{h,k}v) \circ \Psi_T = \mathcal{I}^k(v \circ \Psi_T) \text{ on } T^0 \tag{3.76}$$

for all functions $v \in C^0(cl(\Omega))$ and all elements $T \in \mathcal{T}_h$. This property shows the fundamental localization principle of the finite element method: A

property on the domain Ω is restricted to an element of a given triangulation and then transformed to the n-simplex where handling is easier.

The next theorem is another, very typical application of this principle. It gives an error estimate of the global interpolation operator which is an extension of the local version in Theorem 7, see Ciarlet [27]:

Theorem 10 (Ciarlet 1972) *Let be $1 \leq q \leq \infty$ and $k > \frac{n}{q} - 1$. Then there is a constant $C > 0$ with*

$$|v - \mathcal{I}^{h,k} v|_{m,q,\Omega} \leq C \, \sigma_h^m h^{k-m+1} |v|_{k+1,q,\Omega} \tag{3.77}$$

for all triangulations $\mathcal{T}_h$ with mesh quality σ_h, all functions $v \in W^{k+1,q}(\Omega)$ and all $0 \leq m \leq k$.

Proof: The proof is given in Ciarlet [27] but to show the so–called 'scaling argument' that is a standard argument in the FEM analysis the proof is presented here:

Let be $v \in W^{k+1,q}(\Omega)$ and $T \in \mathcal{T}_h$. The parametrical representation of the element T denoted by Ψ_T is defined by equation (3.61).

By applying Theorem 6 (with $\Omega := T^0$) and Lemma 5 (with $K := T^0$) it is

$$\begin{aligned} |v - \mathcal{I}^{h,k} v|_{m,q,T} &= |(v - \mathcal{I}^{h,k} v) \circ \Psi_T \circ \Psi_T^{-1}|_{m,q,T} \\ &\overset{(3.22)+(3.8)}{\leq} C_1 \rho_T^{-m} |det(B_T)|^{\frac{1}{q}} \\ &\qquad |(v - \mathcal{I}^{h,k} v) \circ \Psi_T|_{m,q,T^0} \ . \end{aligned} \tag{3.78}$$

After using the local representation (3.76) of the global interpolation operator $\mathcal{I}^{h,k}$ one can profit from the error estimate of the local interpolation operator $\mathcal{I}^k$ in Theorem 7:

$$\begin{aligned} |(v - \mathcal{I}^{h,k} v) \circ \Psi_T|_{m,q,T^0} &\overset{(3.76)}{=} |(v \circ \Psi_T) - \mathcal{I}^k (v \circ \Psi_T)|_{m,q,T^0} \\ &\overset{(3.34)}{\leq} C_2 |v \circ \Psi_T|_{k+1,q,T^0} \ . \end{aligned} \tag{3.79}$$

Theorem 7 can be applied since the function $v \circ \Psi_T$ belongs to the Sobolev space $W^{k+1,q}(T^0)$ by Theorem 6. More over it holds

$$|v \circ \Psi_T|_{k+1,q,T^0} \overset{(3.21)+(3.7)}{\leq} C_3 h_T^{k+1} |det(B_T)|^{-\frac{1}{q}} |v|_{k+1,q,T} \ . \tag{3.80}$$

By combining estimates (3.78), (3.79) and (3.80) it comes out that

$$\begin{aligned} |v - \mathcal{I}^{h,k}v|_{m,q,T} &\overset{(3.78)}{\leq} C_1 \rho_T^{-m} |det(B_T)|^{\frac{1}{q}} |(v - \mathcal{I}^{h,k}v) \circ \Psi_T|_{m,q,T^0} \\ &\overset{(3.79)}{\leq} C_4 \rho_T^{-m} |det(B_T)|^{\frac{1}{q}} |v \circ \Psi_T|_{k+1,q,T^0} \\ &\overset{(3.80)}{\leq} C_5 \rho_T^{-m} h_T^{k+1} |v|_{k+1,q,T} \\ &\overset{(3.63)}{\leq} C_5 \sigma_h^m h_T^{k+1-m} |v|_{k+1,q,T} \end{aligned} \tag{3.81}$$

where in the last estimation the definition (3.63) of the mesh quality σ_h is inserted. By summing over all elements $T \in \mathcal{T}_h$ (when $q = \infty$ the $\sum_{T \in \mathcal{T}_h}$ has to be replaced by $\operatorname*{ess\,sup}_{T \in \mathcal{T}_h}$) one obtains

$$\begin{aligned} |v - \mathcal{I}^{h,k}v|^q_{m,q,\Omega} &= \sum_{T \in \mathcal{T}_h} |v - \mathcal{I}^{h,k}v|^q_{m,q,T} \\ &\overset{(3.81)}{\leq} \sum_{T \in \mathcal{T}_h} C_6 \sigma_h^{q \cdot m} h_T^{q(k+1-m)} |v|^q_{k+1,q,T} \\ &\overset{(3.62)}{\leq} C_6 h^{q(k-m+1)} \sigma_h^{q \cdot m} |v|^q_{k+1,q,\Omega} \, . \end{aligned} \tag{3.82}$$

In the last estimation the definition (3.62) of mesh size h is used. So the theorem is proved •

If Theorem 10 is applied for $m = 0, 1$ and $q = 2$ it turns out that for all functions $u \in H^{k+1}(\Omega)$ the interpolation $\mathcal{I}^{h,k}u$ converges to the function u with convergence order k when the step size h goes to zero. Therefore, if the function u is smooth enough, the functions $\mathcal{I}_h u := \mathcal{I}^{h,k}u$ fulfills the proposition (2.48) of the Corollary 1 which will be used to prove the convergence of the discrete solution u_h of the discrete variational problem (3.68) to the sought solution u of variational problem (3.50). It remains to prove that the discrete variational problem converges to the original problem in the sense of proposition (2.49) of Corollary 1.

For a given quadrature scheme Q^{h,Q^l} defined by equation (3.67) the error functional $Q^{h,E^l} : C^0(cl(\Omega)) \to \mathbb{R}$ is defined by

$$Q^{h,E^l}(\varphi) := \int_\Omega \varphi \, dx - Q^{h,Q^l}(\varphi) \tag{3.83}$$

for all $\varphi \in C^0(cl(\Omega))$. By using the local error functional E^l defined by equation (3.37) the error functional Q^{h,E^l} can be written as

$$Q^{h,E^l}(\varphi) = \sum_{T \in \mathcal{T}_h} |det(B_T)| E^l(\varphi \circ \Psi_T) \tag{3.84}$$

for all functions $\varphi \in C^0(cl(\Omega))$. Since the local error functional E^l is a continuous, linear functional on the space $C^0(T^0)$ the error functional Q^{h,E^l} is also continuous on the space $C^0(cl(\Omega))$, i.e. $Q^{h,E^l} \in C^0(cl(\Omega))^*$.

The following theorem is the global version of Theorem 8, see Ciarlet [26]. It is pointed out that the local error functional E^l may be any continuous, linear functional on the space $C^0(T^0)$. It does not need to be the error functional of a quadrature scheme on the n-simplex T^0.

Theorem 11 (Ciarlet 1972) *Let be $l \geq k \geq 1$ and $E^l \in C^0(T^0)^*$ with $E^l(p) = 0$ for all $p \in P_l$. Then there is a constant $C > 0$ so that for all triangulations $\mathcal{T}_h$ of the domain Ω with mesh quality σ_h the linear functional $Q^{h,E^l} : C^0(cl(\Omega)) \to \mathbb{R}$ defined by*

$$Q^{h,E^l}(\varphi) := \sum_{T \in \mathcal{T}_h} |det(B_T)| E^l(\varphi \circ \Psi_T) \tag{3.85}$$

for all $\varphi \in C^0(cl(\Omega))$ belongs to the dual space $C^0(cl(\Omega))^*$. Moreover for all functions $v_h \in V^{h,k}$ and $f \in W^{l-k+2,\infty}(\mathcal{T}_h)$ the following estimates hold $(i = 1, \ldots, n)$:

$$|Q^{h,E^l}(f \cdot v_h)| \leq C h^{l-k+2} \max(|f|_{l-k+1,\infty,\mathcal{T}_h}, |f|_{l-k+2,\infty,\mathcal{T}_h}) \|v_h\|_{1,\Omega} \tag{3.86}$$

and

$$|Q^{h,E^l}(f \cdot \frac{\partial v_h}{\partial x_i})| \leq C\ \sigma_h h^{l-k+2} |f|_{l-k+2,\infty,\mathcal{T}_h} |v_h|_{1,\Omega}\ . \tag{3.87}$$

Proof: The proof can be found in Ciarlet [26]. Here only the proof for inequality (3.87) is presented to show the 'scaling argument' used for quadrature schemes since it is not standard to consider the integration errors in the FEM analysis.

Let be $v_h \in V^{h,k}$, $f \in W^{l-k+2,\infty}(\mathcal{T}_h)$ and $i \in \{1, \ldots, n\}$. From the definition of the error functional Q^{h,E^l} by equation (3.85) it is

$$\begin{aligned} |Q^{h,E^l}(f \cdot \frac{\partial v_h}{\partial x_i})| &\leq \sum_{T \in \mathcal{T}_h} |det(B_T)|\ |E^l((f \cdot \frac{\partial v_h}{\partial x_i}) \circ \Psi_T)| \\ &= \sum_{T \in \mathcal{T}_h} |det(B_T)|\ |E^l((f \circ \Psi_T) \cdot (\frac{\partial v_h}{\partial x_i} \circ \Psi_T))|\ . \end{aligned} \tag{3.88}$$

Now let $T \in \mathcal{T}_h$ be fixed. Then it is

$$p_T := v_h \circ \Psi_T \in P_k\ . \tag{3.89}$$

By Theorem 6 (with $q := 2$ and $m := 1$) and Lemma 5 the inequality

$$|p_T|_{1,T^0} \leq C_1 h_T |det(B_T)|^{-\frac{1}{2}} |v_h|_{1,T} \tag{3.90}$$

holds. In addition let be

$$f_T := f \circ \Psi_T \in W^{l-k+2,\infty}(T^0) \, . \tag{3.91}$$

Again by using Theorem 6 (with $q := \infty$ and $m := l - k + 2$) and Lemma 5 the following inequality holds:

$$|f_T|_{l-k+2,\infty,T^0} \leq C_2 h_T^{l-k+2} |f|_{l-k+2,\infty,T} \, . \tag{3.92}$$

If it is $B_T^{-1} = (\beta_{ij}^T)_{i,j=1,n}$ it results from Lemma 5 that

$$|\beta_{ij}^T| \overset{(3.3)}{\leq} C_3 |B_T^{-1}| \overset{(3.7)}{\leq} C_4 \rho_T^{-1} \tag{3.93}$$

for all $i, j = 1, \ldots, n$. In addition it is by the chain rule

$$\frac{\partial v_h}{\partial x_i}(\Psi_T x^0) = \sum_{j=1}^{n} \beta_{ij}^T \, \frac{\partial p_T}{\partial x_j^0}(x^0) \text{ for all } x^0 \in T^0 \, . \tag{3.94}$$

Using this equation and the fact that the local error functional E^l is linear the following estimates can be made:

$$\begin{aligned} |E^l((f \circ \Psi_T) \cdot (\frac{\partial v_h}{\partial x_i} \circ \Psi_T))| &\overset{(3.94)}{\leq} \sum_{j=1}^{n} |\beta_{ij}^T| |E^l(f_T \cdot \frac{\partial p_T}{\partial x_j^0})| \\ &\overset{(3.93)}{\leq} \sum_{j=1}^{n} C_4 \rho_T^{-1} |E^l(f_T \cdot \frac{\partial p_T}{\partial x_j^0})| \, . \end{aligned} \tag{3.95}$$

Theorem 8 is applied to the term $|E^l(f_T \cdot \frac{\partial p_T}{\partial x_j^0})|$ to get

$$\begin{aligned} & |E^l((f \circ \Psi_T) \cdot (\frac{\partial v_h}{\partial x_i} \circ \Psi_T))| \\ \overset{(3.39)}{\leq} & \sum_{j=1}^{n} C_5 \rho_T^{-1} |f_T|_{l-k+2,\infty,T^0} |p_T|_{1,T^0} \\ \overset{(3.90)+(3.92)}{\leq} & C_5 \rho_T^{-1} h_T^{l-k+2} |f|_{l-k+2,\infty,T} \; h_T |det(B_T)|^{-\frac{1}{2}} |v_h|_{1,T} \, . \end{aligned} \tag{3.96}$$

When inserting this estimate into estimate (3.88) the consequence is

$$\begin{aligned} & |Q^{h,E^l}(f \cdot \frac{\partial v_h}{\partial x_i})| \\ \leq & \sum_{T \in \mathcal{T}_h} C_5 h_T^{l-k+2} |f|_{l-k+2,\infty,T} \frac{h_T}{\rho_T} |det(B_T)|^{\frac{1}{2}} |v_h|_{1,T} \\ \leq & C_5 \sigma_h h^{l-k+2} |f|_{l-k+2,\infty,\mathcal{T}_h} \sum_{T \in \mathcal{T}_h} |det(B_T)|^{\frac{1}{2}} |v_h|_{1,T} \end{aligned} \tag{3.97}$$

where in the last estimate the definition (3.63) of the mesh quality σ_h is used. A further estimate is done by applying the Cauchy-Schwartz inequality for sums:

$$|Q^{h,E^l}(f \cdot \frac{\partial v_h}{\partial x_i})| \leq C_5 \sigma_h h^{l-k+2} |f|_{l-k+2,\infty,\mathcal{T}_h} \sqrt{\sum_{T \in \mathcal{T}_h} |det(B_T)|} \sqrt{\sum_{T \in \mathcal{T}_h} |v_h|^2_{1,T}} \; . \tag{3.98}$$

As the volume $vol(\Omega)$ of the domain Ω is given by

$$vol(\Omega) = \sum_{T \in \mathcal{T}_h} |det(B_T)| \tag{3.99}$$

and it is

$$\sum_{T \in \mathcal{T}_h} |v_h|^2_{1,T} = |v_h|^2_{1,\Omega} \tag{3.100}$$

inequality (3.98) can be written as

$$|Q^{h,E^l}(f \cdot \frac{\partial v_h}{\partial x_i})| \leq C_5 \sigma_h h^{l-k+2} |f|_{l-k+2,\infty,\mathcal{T}_h} \sqrt{vol(\Omega)} |v_h|_{1,\Omega} \; . \tag{3.101}$$

By setting $C := C_5 \sqrt{vol(\Omega)}$ inequality (3.87) is proved.

The proof for inequality (3.86) is analogous to the proof of inequality (3.87). As it is obvious from the definition (3.85) that $Q^{h,E^l} \in C^0(\Omega)^*$ the theorem is proved •

Theorem 11 shows that the discrete variational problem (3.68) converges to the original problem (3.50) for $h \to 0$ if a quadrature scheme on the n-simplex T^0 with an accuracy greater than $k-2$ is used (assuming the function $x \to G(u_{h;}(x), x)$ on the domain Ω is smooth enough and its derivatives up to order $l - k + 2$ are bounded for $h \to 0$, see Lemma 9).

Remark 1: In the proofs of Theorem 10 and Theorem 11 it was essential that the parametrical representations of the elements are *affine* transformations. The proofs for non–linear parametrical representations are much more difficult but the results are essentially the same, see Ciarlet [25].

Remark 2: The actual values of the constants C occurring in the estimates (3.77), (3.86) and (3.87) are unknown as only the existence but not the values of the corresponding constants in the underlying Theorem 7 and Theorem 8 are known.

3.5 The Finite Element Discretization

Theorem 10 shows that for a function u on the domain Ω there are elements in the spaces $V^{h,k}$, namely $\mathcal{I}^{h,k}u$, which converge to the function u if the mesh size h goes to zero. Moreover Theorem 11 ensures that the approximation of the integrals by numerical integration converges to the integral if the mesh size goes to zero. Both hold if the involved functions are smooth enough and the quadrature scheme is exact of order $k-1$ in the sense of Definition 4. From this point of view the discrete variational problem (3.68) has the potential to deliver approximative solutions converging to the sought solution if the mesh size decreases.

To confirm this expectation it has to be proven that the discrete variational problem (3.68) has an unique solution and that the involved operator is well–posed with a condition number independent of the mesh size. Then the convergence of the FEM approximations results from Corollary 1.

The next lemma is essential to obtain condition numbers independent of the mesh size, see Ciarlet [26]:

Lemma 8 *Let Q^{2k-2} be a quadrature scheme that is exact of order $2k-2$ in the sense of Definition 4. Then a real constant $C > 0$ exists with*

$$\frac{1}{C}\|u_h\|_{1,\Omega}^2 \leq Q^{h,Q^{2k-2}}(|u_{h;}|^2) \leq C\|u_h\|_{1,\Omega}^2 \tag{3.102}$$

for all triangulations $\mathcal{T}_h$ of the domain Ω and all functions $u_h \in V^{h,k}$.

Proof: See Ciarlet [26]. First it is shown that the mapping $p \to \sqrt{Q^{2k-2}(|p_{;}|^2)}$ defines a norm on the finite dimensional space P_k. Therefore inequality (3.102) holds for all polynomials of order k on the n-simplex T^0. Using the scaling argument in the proof of Theorem 11 the inequality is shifted from polynomial space P_k to the space $V^{h,k}$ •

The following theorem which is the discrete version of Theorem 9 guarantees the existence of the FEM approximation $u_h \in V^{h,k}$. More important for the discussions in this thesis is the result that the involved discrete operator is well–posed with a condition number that is independent of the mesh size:

Theorem 12 (Grosz) *Let $k \geq 1$ and Q^{2k-2} be a quadrature scheme on the n-simplex T^0 which is exact of order $2k-2$ in the sense of Definition 4. Then there is a constant $C > 0$ so that for all uniform positive definite kernels G*

with positivity bound D, all triangulations $\mathcal{T}_h$ of the domain Ω and all spaces $V_h \subset V^{h,k}$ the operator $F_h : V_h \to V_h^*$ defined by

$$< v_h, F_h(u_h) >:= Q^{h,Q^{2k-2}}(v_{h;} \cdot G(u_{h;}, .)) \tag{3.103}$$

for all $v_h, u_h \in V_h$ is well–posed with condition number $D^2 \cdot C$. Moreover the variational problem

$$< v_h, F_h(u_h) >= 0 \text{ for all } v_h \in V_h \tag{3.104}$$

has exactly one solution $u_h \in V_h$.

Proof: The proof resembles the proof of Theorem 9 but some modifications have to be introduced to verify the propositions of Theorem 1 in the Hilbert space $V_h \subset V^{h,k} \subset H^1(\Omega)$.

Analogously to inequality (3.57) in the proof of Theorem 9 it is for all functions $w_h, u_{1h}, u_{2h} \in V_h \subset V^{h,k}$:

$$\begin{aligned} & < w_h, F_h(u_{1h}) - F_h(u_{2h}) > \\ = \;& Q^{h,Q^{2k-2}}(w_{h;} \cdot [G(u_{1h;}, .) - G(u_{2h;}, .)]) \\ \overset{(3.53)}{\leq} \;& D\sqrt{Q^{h,Q^{2k-2}}(|u_{1h;} - v_{2h;}|^2)}\sqrt{Q^{h,Q^{2k-2}}(|w_{h;}|^2)} \\ \overset{(3.102)}{\leq} \;& C\ D\|u_{1h} - u_{2h}\|_{1,\Omega}\|w_h\|_{1,\Omega} \end{aligned} \tag{3.105}$$

where in the last estimate Lemma 8 is applied. So it has been shown that for all functions $u_{1h}, u_{2h} \in V_h$:

$$\|F_h(u_{1h}) - F_h(u_{2h})\|_{V_h^*} \leq C\ D\|u_{1h} - u_{2h}\|_{1,\Omega}\ . \tag{3.106}$$

To prove the condition (2.23) the estimate (3.60) in the proof of Theorem 9 is slightly changed: For all functions $u_{1h}, u_{2h} \in V_h$ it is

$$\begin{aligned} & < u_{1h} - u_{2h}, F_h(u_{1h}) - F_h(u_{2h}) > \\ = \;& Q^{h,Q^{2k-2}}([u_{1h;} - u_{2h;}] \cdot [G(u_{1h;}, .) - G(u_{2h;}, .)]) \\ \overset{(3.59)}{\geq} \;& \frac{1}{D}Q^{h,Q^{2k-2}}(|u_{1h;} - u_{2h;}|^2) \\ \overset{(3.102)}{\geq} \;& \frac{1}{CD}\|u_{1h} - u_{2h}\|_{1,\Omega}^2\ . \end{aligned} \tag{3.107}$$

In the last estimate Lemma 8 is used again. As the assumptions of Theorem 1 were verified in the Hilbert space V_h the theorem is proved •

By applying this Theorem 12 to the uniform positive definite mapping

$$(\chi, x) \to \chi \cdot \partial G(u_{h;}(x), x) := (\sum_{j=1}^{n+1} \chi_j \partial_j G_i(u_{h;}(x), x))_{i=1,n+1} \tag{3.108}$$

for all $x \in \Omega$ and $\chi = (\chi_j)_{j=1,n+1} \in \mathbb{R}^{n+1}$ with positivity bound D (the function $u_{h;} \in V^{h,k}$ is fixed) the following corollary is derived:

Corollary 4 *With the propositions of Theorem 12 there is a constant $C > 0$, so that for all uniform positive definite kernels G with positivity bound D, all triangulations $\mathcal{T}_h$ of the domain Ω, all fixed functions $u_h \in V^{h,k}$ and all subspaces $V_h \subset V^{h,k}$ the linear operator $DF_h(u_h) : V_h \to V_h^*$ defined by*

$$< v_h, DF_h(u_h) w_h >= Q^{h,Q^{2k-2}}(v_{h;} \cdot \partial G(u_{h;}, .) w_{h;}) \tag{3.109}$$

for all $v_h, w_h \in V_h$ is an isomorphism in $\mathcal{L}(V_h, V_h^*)$ with

$$\begin{array}{lcl} \|DF_h(u_h)\|_{\mathcal{L}(V_h, V_h^*)} & \leq & D \cdot C \\ \|DF_h(u_h)^{-1}\|_{\mathcal{L}(V_h^*, V_h)} & \leq & D \cdot C \ . \end{array} \tag{3.110}$$

Remark: This Corollary 4 is the classical existence theorem for the finite element approximation of linear variational problems, see Strang [60].

The propositions of Corollary 1 are verified to prove the convergence of the FEM approximations to the sought solution u. For this the next lemma is important:

Lemma 9 (Grosz) *Let be $l + 2 \geq k > 0$, $E^l \in C^0(T^0)^*$ with $E^l(p) = 0$ for all $p \in P_l$, $G \in C^{l-k+2}(\mathbb{R}^{n+1} \times cl(\Omega))^{n+1}$ and $M > 0$ fixed. Then there is a real constant $C > 0$ with*

$$\sup_{v_h \in V^{h,k}} \frac{1}{\|v_h\|_{1,\Omega}} |Q^{h,E^l}(v_{h;} \cdot G(u_{;}, .))| \leq C \sigma_h h^{l-k+2} \tag{3.111}$$

for all triangulations $\mathcal{T}_h$ of the domain Ω with mesh quality σ_h and all functions $u \in C^{l-k+3}(\mathcal{T}_h)$ with $\|u\|_{l-k+3,\infty,\mathcal{T}_h} \leq M$.

Proof: Let $\mathcal{T}_h$ be a triangulation of the domain Ω and $u \in C^{l-k+3}(\mathcal{T}_h)$ with

$$\|u\|_{l-k+3,\infty,\mathcal{T}_h} \leq M \ . \tag{3.112}$$

Theorem 11 is applied to the terms of the sum on the left hand side of the inequality (3.111). It remains to prove that for a fixed component $i \in \{1, \ldots, n+1\}$ the function $f : \Omega \to \mathbb{R}$ defined by

$$f(x) = G_i(u_{;}(x), x) \text{ for all } x \in \Omega \tag{3.113}$$

belongs to the Sobolev space $W^{l-k+2,\infty}(\mathcal{T}_h)$ and $\|f\|_{l-k+2,\infty,\mathcal{T}_h} \leq C_1$ with a constant $C_1 > 0$ that may depend on the value M but must be independent of the function u.

Let be $\alpha \in \mathbb{N}_0^n$ with $|\alpha| \leq l-k+2$ and

$$d_\alpha := card(\{\beta \in \mathbb{N}^n | \ |\beta| \leq |\alpha|+1\}) \tag{3.114}$$

where $card(Z)$ denotes the number of elements in the set Z. Then there is a function $g_\alpha \in C^{l-k+2-|\alpha|}(cl(\Omega) \times \mathbb{R}^{d_\alpha})$ with

$$D^\alpha f(x) = g_\alpha(x, (D^\beta u(x))_{\beta \in \mathbb{N}_0^n, |\beta| \leq |\alpha|+1}) \tag{3.115}$$

for all $x \in \Omega$. This can be easily proved by using induction over $|\alpha|$ and the chain rule.

Since for all multi–indices α with $|\alpha| \leq l-k+2$ the function g_α is continuous on the compact set $cl(\Omega) \times [-M, M]^{d_\alpha}$ it exists the constant

$$C_2 := \max_{|\alpha| \leq l-k+2} |g_\alpha|_{0,\infty,cl(\Omega)\times[-M,M]^{d_\alpha}} \ . \tag{3.116}$$

For all elements $T \in \mathcal{T}_h$ it is $D^\alpha f|_T \in C^0(T)$ as $g_\alpha \in C^0((cl(\Omega) \times \mathbb{R}^{d_\alpha})$ and $u \in C^1(\mathcal{T}_h)$. Moreover it is

$$|D^\alpha f|_{0,\infty,T} \overset{(3.115)}{=} |g_\alpha(x, (D^\beta u(x))_{\beta \in \mathbb{N}_0^n, |\beta| \leq |\alpha|+1})|_{0,\infty,T} \overset{(3.116)}{\leq} C_2 \tag{3.117}$$

since for all multi–indices $\beta \in \mathbb{N}_0^n$ with $|\beta| \leq |\alpha|+1 \leq l-k+3$ and all $x \in T$ the estimate

$$|D^\beta u(x)| \leq \|u\|_{l-k+3,\infty,\mathcal{T}_h} \leq M \tag{3.118}$$

holds. It is proved that $f \in W^{l-k+2,\infty}(\mathcal{T}_h)$ with $\|f\|_{l-k+2,\infty,\mathcal{T}_h} \leq C_2$ where the constant C_2 depends on the bound M and the kernel G.

Therefore the lemma is proved by Theorem 11 •

Now the convergence of the finite element approximations which are the solution of the discrete variational problems (3.104) to the solution of the variational problem (3.50) for decreasing mesh sizes h is proved. It is the non–linear version of the famous result of Zlamal [74] for linear variational problems and the result of Ciarlet [26] considering in addition the integration error.

Theorem 13 (Grosz) *Let $l \geq k \geq 1$, $F : H^1(\Omega) \to H^1(\Omega)^*$ be the operator generated by the uniform positive kernel $G \in C^l(\mathbb{R}^{n+1} \times cl(\Omega))^{n+1}$ and $u \in W^{k+1,\infty}(\Omega)$ be the unique solution of the variational problem*

$$< v, F(u) >:= \int_\Omega v_{;} \cdot G(u_{;}, .) \, dx = 0 \tag{3.119}$$

for all $v \in H^1(\Omega)$. In addition let $(\mathcal{T}_h)_{h>0}$ be a family of triangulations of the domain Ω with bounded mesh quality $\sigma_h \leq \sigma$, V_h be a subspace of $H^1(\Omega)$ with $V^{h,k} \subset V_h \subset V^{h,l}$ and Q^{2l-2} be a quadrature scheme on the n-simplex T^0 which is exact of order $2l-2$ in the sense of Definition 4.

Then the discrete variational problem

$$< v_h, F_h(u_h) >:= Q^{h,Q^{2l-2}}(v_{h;} \cdot G(u_{h;}, .)) = 0 \tag{3.120}$$

for all $v_h \in V_h$ has exactly one solution $u_h \in V_h$. Moreover there is a constant $C > 0$ independent of the mesh size h with

$$\|u - u_h\|_{1,\Omega} \leq C h^k \tag{3.121}$$

for all $h > 0$, i.e $(u_h)_{h>0}$ converges to the solution u with minimal order k.

Proof: The propositions of Corollary 1 are verified for $K_h := V_h$ and $\mathcal{I}_h := \mathcal{I}^{h,k}$: By Theorem 12 the operator $F_h : V_h \to V_h^*$ defined by

$$< v_h, F_h(u_h) >:= Q^{h,Q^{2l-2}}(v_{h;} \cdot G(u_{h;}, .)) \tag{3.122}$$

for all $v_h, u_h \in V_h$ is well–posed with a condition number which is independent of the mesh size. Moreover the discrete variational problem (3.120) has exactly one solution $u_h \in V_h$. From Theorem 10 (for $m := 0, 1$ and $q := 2$) it is

$$\begin{array}{rcl} \|u - \mathcal{I}^{h,k}u\|_{1,\Omega} & \overset{(3.77)}{\leq} & C_1 h^k \sqrt{h^2 + \sigma_h^2}\,|u|_{k+1,\Omega} \\ & \leq & C_2 h^k |u|_{k+1,\infty,\Omega} \end{array} \tag{3.123}$$

as $h \leq h_\Omega$ and $\sigma_h \leq \sigma$. That is proposition (2.48) of Corollary 1.

To verify proposition (2.49) Theorem 10 (for $m := 0, \ldots, k$ and $q := \infty$) is used again and one gets

$$\begin{array}{rcl} \|\mathcal{I}^{h,k}u\|_{l+1,\infty,\mathcal{T}_h} & = & \|\mathcal{I}^{h,k}u\|_{k,\infty,\mathcal{T}_h} \\ & \leq & \|u\|_{k,\infty,\Omega} + \|u - \mathcal{I}^{h,k}u\|_{k,\infty,\mathcal{T}_h} \\ & \overset{(3.77)}{\leq} & \|u\|_{k,\infty,\Omega} + C_3 \max\limits_{0 \leq m \leq k} \sigma_h^m h^{k-m+1} |u|_{k+1,\infty,\Omega} \\ & \leq & C_4 \|u\|_{k+1,\infty,\Omega} =: M \end{array} \tag{3.124}$$

since all terms $\sigma_h^m h^{k-m+1}$ are bounded for all $0 \leq m \leq k$ and all mesh sizes $h \leq h_\Omega$.

The linear functional $E^{2l-2} \in C^0(T^0)^*$ defined by

$$E^{2l-2}(\varphi) := \int_{T^0} \varphi \, dx^0 - Q^{2l-2}(\varphi) \tag{3.125}$$

for all $\varphi \in C^0(T^0)$ has the property

$$E^{2l-2}(p) = 0 \text{ for all } p \in P_{2l-2} \tag{3.126}$$

as the quadrature scheme Q^{2l-2} is exact of order $2l-2$. Lemma 9 for $l := 2l-2$ and M defined by inequality (3.124) shows that

$$\begin{aligned}
& \|F(\mathcal{I}^{h,k}u) - F_h(\mathcal{I}^{h,k}u)\|_{V_h^*} \\
\overset{(3.120)+(3.125)}{=} & \sup_{v_h \in V_h} \frac{1}{\|v_h\|_{1,\Omega}} |Q^{h,E^{2l-2}}(v_{h;} \cdot G([\mathcal{I}^{h,k}u]_{;}, .))| \\
\leq & \sup_{v_h \in V^{h,l}} \frac{1}{\|v_h\|_{1,\Omega}} |Q^{h,E^{2l-2}}(v_{h;} \cdot G([\mathcal{I}^{h,k}u]_{;}, .))| \\
\overset{(3.111)}{\leq} & C_5 h^l
\end{aligned} \tag{3.127}$$

with a constant $C_5 > 0$ which is independent of the mesh size h. After using Corollary 1 the lemma has been proved •

Remark 1: In the propositions of Theorem 13 it has been assumed that the solution u belongs to the Sobolev space $W^{k+1,\infty}(\Omega)$ to guarantee that the values $\|\mathcal{I}^{h,k}u\|_{l+1,\infty,\mathcal{T}_h}$ are bounded independently of the mesh size h, see inequality (3.124). For linear problems it is sufficient to have that $u \in H^{k+1}(\Omega)$, see Ciarlet [26]. That can be achieved by the smoothness of the kernel G, see Grisvard [51].

Remark 2: Even if the finite element space $V^{h,k'} \subset V^{h,l}$ with $k' > k$ is used to construct approximations of the solution u (i.e. piecewise polynomials of higher order than k are used) but $u \notin W^{k`+1}$ no better convergence order than k can be achieved, see also Example 2 in Section 4.5.

Theorem 13 states: By using piecewise linear polynomial FEM approximations and a quadrature scheme which is exact for polynomials with constant values the FEM approximations converge to the sought solution u of order one if $u \in W^{2,\infty}(\Omega)$. If even $u \in W^{3,\infty}(\Omega)$ and piecewise quadratic polynomials together with a quadrature scheme that is exact for quadratical polynomials are used the FEM approximations converge with order two. Even after all used estimates have been gathered the constant C in inequality (3.121) is unknown as it contains constants whose values cannot be computed. Moreover the true error would be highly overestimated as some rough estimates are applied. Therefore it is necessary to have an a–posteriori error estimate to get an acceptable estimate for the true error.

In the next section the projecting error estimate introduced in Section 2.3 is applied to the FEM in the scope of the variational problem (3.119). The

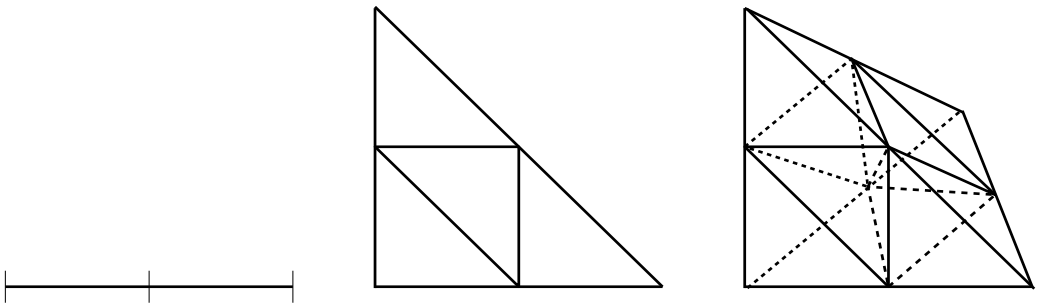

Figure 3.6: The subdivision of the n-simplex for the global refinement.

result of Theorem 13 that the FEM approximation by higher order polynomials has a higher convergence order shows how the expansion V_{h+} has to be selected namely by adding higher order polynomials to the space V_h to obtain that $(u_h, u_{h+})_{h>0}$ is saturated for the sought solution u in the sense of Definition 2.

3.6 The Projecting Error Estimate

A special kind of triangulation $\mathcal{T}_h$ which is produced by a global refinement of a given triangulation $\mathcal{T}_{2h}$ is introduced as follows:

The n-simplex T^0 is subdivided into 2^n subsets $T_1^0, \ldots, T_{2^n}^0$ as it is shown in Figure 3.6. The set

$$\mathcal{T}_0 := \{T_i^0\}_{i=1,2^n} \tag{3.128}$$

is a triangulation of the n-simplex T^0. For all $i = 1, \ldots, 2^n$ the subelement T_i^0 has an affine representation $\Psi_{T_i^0} : T^0 \to T_i^0$ defined by

$$\Psi_{T_i^0} x := B_{T_i^0} x + b_{T_i^0} \text{ for all } x \in T^0 \tag{3.129}$$

where it is $b_{T_i^0} \in \mathbb{R}^n$ and $B_{T_i^0} \in \mathbb{R}^{n \times n}$ with $det(B_{T_i^0}) \neq 0$.

This subdivision of the n-simplex creates a new triangulation out of a given triangulation:

Definition 7 *Let be $\mathcal{T}_{2h}$ a given triangulation of mesh size $2h$. Then*

$$\mathcal{T}_h := \{(\Psi_{T_2} \circ \Psi_{T_i^0})[T^0] \mid i = 1, \ldots, 2^n \text{ and } T_2 \in \mathcal{T}_{2h}\} \tag{3.130}$$

is called the global refined triangulation of the triangulation $\mathcal{T}_{2h}$ of mesh size h. The triangulation $\mathcal{T}_{2h}$ is called the coarse mesh and the triangulation $\mathcal{T}_h$ is called the fine mesh.

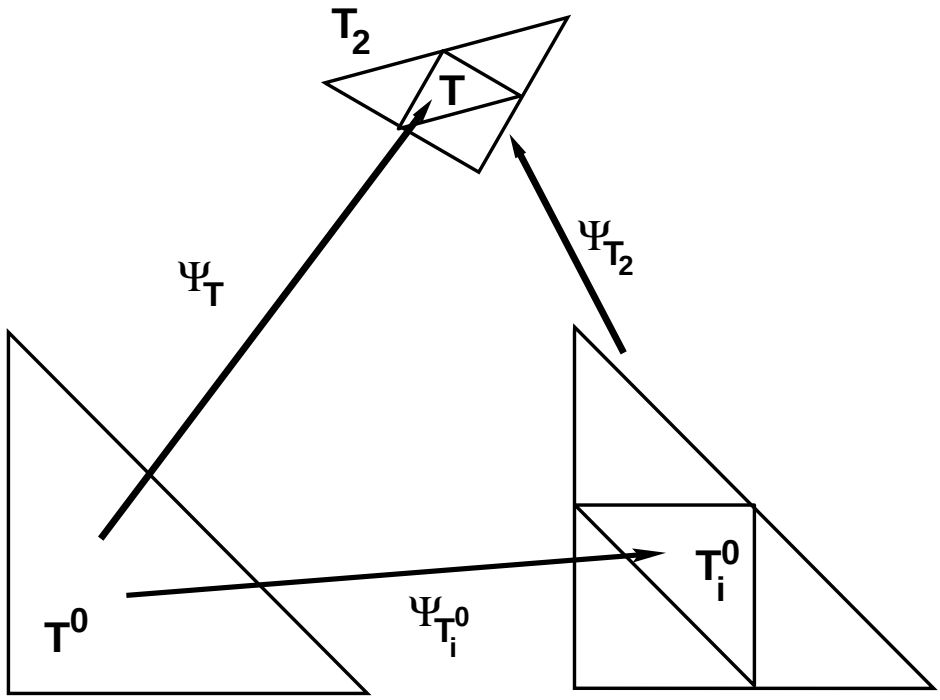

Figure 3.7: The parametrical representations Ψ_T of a refined element T_2.

It becomes clear that the family $\mathcal{T}_h$ is again a triangulation. The parametrical representation Ψ_T of an element $T \in \mathcal{T}_h$ is given by

$$\Psi_T = \Psi_{T_2} \circ \Psi_{T_i^0} \tag{3.131}$$

for any $T_2 \in \mathcal{T}_{2h}$ and any $i \in \{1, \ldots, 2^n\}$, see Figure 3.7.

For the following discussions it is useful to introduce some additional spaces: The set S_k defined by

$$S_k := \{s \in C^0(T^0) | \text{ for all } i = 1, \ldots, 2^n : s|_{T_i^0} \in P_k\} \tag{3.132}$$

denotes the space of all continuous function on the n-simplex T^0 which are polynomials of order k on the subelements in $\mathcal{T}_0$. The space

$$S^{h,k} := \{v \in C^0(cl(\Omega)) | \text{ for all } T \in \mathcal{T}_h : v|_T \circ \Psi_T^{-1} \in S_k\} \tag{3.133}$$

is the set of all continuous and piecewise S^k–functions on the domain Ω. As the functions in the space S^k are piecewise polynomials the space $S^{h,k}$ is not a new space but it is the space of piecewise polynomials of order k on the fine mesh.

Lemma 10

$$V^{h,k} = S^{2h,k} \ . \tag{3.134}$$

Proof: By Definition 7 of the global refined triangulation $\mathcal{T}_h$ of the coarse mesh $\mathcal{T}_{2h}$ the function v_h belongs to space $V^{h,k}$ if and only if for all elements $T_2 \in \mathcal{T}_{2h}$ in the coarse mesh and all subelements $i = 1, \ldots, 2^n$

$$v_h \circ \Psi_{T_2} \circ \Psi_{T_i^0} \in P_k \ . \tag{3.135}$$

As all the parametric representations $\Psi_{T_1^0}, \ldots, \Psi_{T_{2^n}^0}$ are affine transformations this holds if and only if

$$v_h \circ \Psi_{T_2} \in S_k \tag{3.136}$$

for all elements $T_2 \in \mathcal{T}_{2h}$. This exactly means that $v_h \in S^{2h,k}$. So the lemma has been proved •

Lemma 10 allows to use the space $V^{h,k}$ defined on the fine mesh $\mathcal{T}_h$ as a space basing on the coarse grid $\mathcal{T}_{2h}$.

As discussed in the foregoing section an approximation u_h of the sought solution u is computed from the space $V_h := V^{h,k}$ by using piecewise polynomials of order k on the fine mesh. The space V_{h+} involved in the definition for the a–posteriori error estimate introduced in Section 2.3 is the expansion of the space V_h in such a way that the space $V^{2h,2k}$ becomes a subspace of the expansion V_{h+}, i.e. the space V_{h+} contains piecewise polynomials of order $2k$ on the coarse mesh. The approach is motivated by the fact that a better approximation of the solution u from piecewise polynomials of higher order than k can be expected.

Lemma 12 will show that it is sufficient for the construction of the expansion V_{h+} by equation (2.92) to add

$$V_h^c := V_0^{2h,2k} := \{v_{h+} \in V^{2h,2k} \mid \mathcal{I}^{2h,k} v_{h+} = 0\} \tag{3.137}$$

to the space $V^{h,k}$ to achieve that the space $V^{2h,2k}$ is a subset of expansion V_{h+}. The space $V_0^{2h,2k}$ is the set of all piecewise polynomials of order $2k$ on the coarse mesh that vanish at the global degrees of freedom of order k on the same mesh.

At first Lemma 12 is proved for polynomials on the n-simplex T^0:

Lemma 11 *Let be*

$$P_{2k\,0} := \{p \in P_{2k} \mid \mathcal{I}^k p = 0\} \tag{3.138}$$

the space of all polynomials of order $2k$ that vanish at the local degrees of freedom of order k. Then the following identities hold:

$$P_k = P_{2k} \cap S_k \tag{3.139}$$

and

$$P_{2k} \subset P_{2k\,0} \oplus S_k \subset S_{2k} \tag{3.140}$$

with $P_{2k\,0} \cap S_k = \{0\}$.

Proof: The equation (3.139) is obvious.

To prove inclusion (3.140) it has to be shown that $P_{2k\,0} \cap S_k = \{0\}$: Let be $p \in P_{2k} \cap S_k$ with $\mathcal{I}^k p = 0$. From equation (3.139) it is $p \in P_k$ and therefore it is $p = \mathcal{I}^k p = 0$. Hence it is actually $P_{2k\,0} \cap S_k = \{0\}$.

In addition for any $p \in P_{2k}$ the function $\mathcal{I}^k p$ belongs to $P_k \subset S_k$ and it is

$$q := p - \mathcal{I}^k p \in P_{2k} \tag{3.141}$$

with $\mathcal{I}^k q = 0$. Therefore it has been verified that the polynomial $p = q + \mathcal{I}^k p$ is in the space $P_{2k\,0} \oplus S_k$. The second inclusion of inclusion (3.140) is evident. So the lemma has been proved •

Now the proof of the following lemma becomes very simple:

Lemma 12 *With the space $V_0^{2h,2k}$ defined by equation (3.137) it is*

$$V^{2h,2k} \subset V_0^{2h,2k} \oplus V^{h,k} \subset V^{h,2k} \tag{3.142}$$

where it is $V_0^{2h,2k} \cap V^{h,k} = \{0\}$.

Proof: By using Lemma 10 inclusion (3.142) can be reformulated to

$$V^{2h,2k} \subset V_0^{2h,2k} \oplus S^{2h,k} \subset S^{2h,2k} \ . \tag{3.143}$$

Moreover property (3.76) allows to break off the global interpolation operator $\mathcal{I}^{2h,k}$ into the local interpolation operator $\mathcal{I}^k$ on the n-simplex T^0. Then the statements of the lemma are shifted to the n-simplex by the parametrical representations of the elements in the coarse mesh $\mathcal{T}_{2h}$. The effort now is to prove that

$$P_{2k} \subset P_{2k\,0} \oplus S_k \subset S_{2k} \ . \tag{3.144}$$

As this is exactly equation (3.140) in Lemma 11 the lemma is proved •

It has to be shown that the norm of the global interpolation operators $\mathcal{I}^{2h,2k}$ and $\mathcal{I}^{h,k}$ which will play the rule of the joining operator $\mathcal{J}_{h+}$ and its right hand side inverse $\mathcal{J}_{\bar{h}}$ have a norm that is independent of the mesh size h. To do so the quotient space technique is used, see Heuser [43]:

For any function $v \in H^1(T^0)$ the set

$$[v] := \{v + c \mid c \in \mathbb{R}\} \tag{3.145}$$

contains all functions in the Sobolev space $H^1(T^0)$ which differ from the function v by a constant. By introducing vector addition and scalar multiplication in a canonical manner the *quotient space*

$$H^1(T^0)/\mathbb{R} := \{[v] \mid v \in H^1(T^0)\} \tag{3.146}$$

is a vector space. Moreover it is a Banch space when using the norm defined by

$$\|[v]\|_{1,T^0} := |v|_{1,T^0} \tag{3.147}$$

for all $v \in H^1(T^0)$. For any subset $Z \subset H^1(T^0)$ it is set

$$Z/\mathbb{R} := \{[v] \mid v \in Z\} \subset H^1(T^0)/\mathbb{R}\,. \tag{3.148}$$

It is evident that the global interpolation operators $\mathcal{I}^{2h,2k}$ and $\mathcal{I}^{h,k}$ are continuous on the finite dimensional space $V^{h,2k}$. Moreover there are upper bounds for their norms that are independent of the mesh size:

Lemma 13 *Let be $k \geq 1$. Then there is a constant $C > 0$ with*

$$\begin{array}{rl} \|\mathcal{I}^{2h,2k} u_{h+}\|_{1,\Omega} & \leq C\sigma_h \|u_{h+}\|_{1,\Omega} \\ \|\mathcal{I}^{h,k} u_{h+}\|_{1,\Omega} & \leq C\sigma_h \|u_{h+}\|_{1,\Omega} \end{array} \tag{3.149}$$

for all triangulations $\mathcal{T}_h$ with mesh quality σ_h and all functions $u_{h+} \in V^{h,2k}$.

Proof: Since the space S_{2k} has a finite dimension there is a constant $C_1 > 0$ depending on the n-simplex and the order k with

$$|\mathcal{I}^{2k} p|_{0,T^0} \leq C_1 |p|_{0,T^0} \tag{3.150}$$

for all functions $p \in S_{2k}$. The linear operator $[\mathcal{I}^{2k}] : S_{2k}/\mathbb{R} \to H^1(T^0)/\mathbb{R}$ is defined by

$$[p] \to [\mathcal{I}^{2k}][p] := [\mathcal{I}^{2k} p] \tag{3.151}$$

for all $[p] \in S_{2k}/\mathbb{R}$. This definition is senseful as $\mathcal{I}^{2k}[\mathbb{R}] \subset \mathbb{R}$. Since the quotient space $S_{2k}/\mathbb{R}$ is finite dimensional there is a constant $C_2 > 0$ with

$$\|[\mathcal{I}^{2k}][p]\|_{1,T^0} \leq C_2 \|[p]\|_{1,T^0} \tag{3.152}$$

for all $[p] \in S_{2k}/\mathbb{R}$. Then it holds for all functions $p \in S_{2k}$:

$$\begin{aligned} |\mathcal{I}^{2k}p|_{1,T^0} &\overset{(3.147)}{=} \|[\mathcal{I}^{2k}p]\|_{1,T^0} \\ &\overset{(3.151)}{=} \|[\mathcal{I}^{2k}][p]\|_{1,T^0} \\ &\overset{(3.152)}{\leq} C_2\|[p]\|_{1,T^0} \\ &\overset{(3.147)}{=} C_2|p|_{1,T^0} \ . \end{aligned} \tag{3.153}$$

Inequality (3.149) is now shown by the scaling argument used in the proof of Theorem 10:

Let be $u_{h+} \in V^{h,2k} = S^{2h,2k}$. As for all elements $T \in \mathcal{T}_{2h}$ it is $u_{h+} \circ \Psi_T \in S_{2k}$ the inequalities (3.150) and (3.153) can be used in the following manner $(m = 0, 1)$:

$$\begin{aligned} & |\mathcal{I}^{2h,2k}u_{h+}|^2_{m,\Omega} \\ = \quad & \sum_{T\in\mathcal{T}_{2h}} |\mathcal{I}^{2h,2k}u_{h+}|^2_{m,T} \\ \overset{(3.22)+(3.8)}{\leq} \quad & C_3 \sum_{T\in\mathcal{T}_{2h}} \rho_T^{-m}|det(B_T)|\ |\mathcal{I}^{2k}(u_{h+}\circ\Psi_T)|^2_{m,T^0} \\ \overset{(3.150)or(3.153)}{\leq} \quad & C_4 \sum_{T\in\mathcal{T}_{2h}} \rho_T^{-m}|det(B_T)|\ |u_{h+}\circ\Psi_T|^2_{m,T^0} \\ \overset{(3.21)+(3.7)}{\leq} \quad & C_5 \sum_{T\in\mathcal{T}_{2h}} h_T^m\rho_T^{-m}|u_{h+}|^2_{m,T} \\ \leq \quad & C_5\sigma_h^m\|u_{h+}\|^2_{m,\Omega} \\ \leq \quad & C_5\sigma_h\|u_{h+}\|^2_{1,\Omega} \ . \end{aligned} \tag{3.154}$$

By combining the both cases $m = 1$ and $m = 0$ the first estimate in inequality (3.149) is proved.

To prove the second estimate the same proof like for the first estimate can be used but the space of piecewise polynomials S_{2k} is replaced by the polynomial space P_{2k}, the global interpolation operator $\mathcal{I}^{2h,2k}$ by the interpolation operator $\mathcal{I}^{h,k}$ and the coarse mesh $\mathcal{T}_{2h}$ by the fine mesh $\mathcal{T}_h$ •

The next lemma confirms the proposition (2.115) of Theorem 4 with a deflection κ in the Pythagorean equation that is independent of the mesh size. In the range of multilevel methods Eijkhout [34] has shown a similar result with a more difficult proof but for a more general situation.

Lemma 14 *Let be $k \geq 1$. Then there is a constant $C > 0$ with*

$$\|v_h\|^2_{1,\Omega} + \|v_{h+}\|^2_{1,\Omega} \leq C\sigma_h^2\|v_h + v_{h+}\|^2_{1,\Omega} \tag{3.155}$$

for all triangulations $\mathcal{T}_h$ with mesh quality σ_h, all functions $v_h \in V^{h,k}$ and $v_{h+} \in V_0^{2h,2k}$.

Proof: It is known from Lemma 11 that

$$S_k \cap P_{2k\,0} = \{0\}\ . \tag{3.156}$$

Therefore by Lemma 3 there is a constant $C_1 > 0$ with

$$|p|^2_{0,T^0} + |s|^2_{0,T^0} \leq C_1 |p+s|^2_{0,T^0} \tag{3.157}$$

for all functions $s \in S_k$ and all polynomials $p \in P_{2k\,0}$. From equation (3.156) it is

$$S_k/\mathbb{R} \cap P_{2k\,0}/\mathbb{R} = \{[0]\} \tag{3.158}$$

and so, again by Lemma 3, there is $C_2 > 0$ with

$$\begin{aligned} |p|^2_{1,T^0} + |s|^2_{1,T^0} &\overset{(3.147)}{=} \|[p]\|^2_{1,T^0} + \|[s]\|^2_{1,T^0} \\ &\overset{(2.97)}{\leq} C_2 \|[p+s]\|^2_{1,T^0} \\ &\overset{(3.147)}{=} C_2 |p+s|^2_{1,T^0} \end{aligned} \tag{3.159}$$

for all functions $s \in S_k$ and all polynomials $p \in P_{2k\,0}$. Using the scaling argument in the proof of Lemma 13 the inequalities (3.157) and (3.159) are shifted from the space S_k to the space $V^{h,k}$ and from the space $P_{2k\,0}$ to the space $V_0^{2h,2k}$ via the triangulation $\mathcal{T}_{2h}$. So the estimate of the lemma has been proved •

The following theorem is the main result of this chapter. It introduces the projecting a–posteriori error estimate to the FEM.

Theorem 14 (Grosz) *Let be $k \geq 1$, $G \in C^{2k}(\times \mathbb{R}^{n+1} \times cl(\Omega))^{n+1}$ be uniform positive definite with positivity bound D and let be Q^{2k-1} and Q^{4k-2} quadrature schemes on the n-simplex T^0 exact of order $2k-1$ and $4k-2$ in the sense of Definition 4. Let $u \in W^{k+2,\infty}(\Omega)$ be the solution of the variational problem*

$$< v, F(u) > := \int_\Omega v_{;} \cdot G(u_{;}, .)\, dx = 0 \tag{3.160}$$

for all $v \in H^1(\Omega)$. Let $(\mathcal{T}_{2h})_{h>0}$ be a family of triangulations of the domain Ω with global refined triangulations $(\mathcal{T}_h)_{h>0}$ and bounded mesh quality $\sigma_h \leq \sigma$. Let for all mesh sizes $h > 0$ $u_h \in V^{h,k}$ be the solution of the discrete variational problem

$$Q^{h,Q^{2k-1}}(v_{h;} \cdot G(u_{h;}, .)) = 0 \tag{3.161}$$

for all $v_h \in V^{h,k}$ *and let be* $C_1 > 0$ *with*

$$\|u_h\|_{k,\infty,\mathcal{T}_h} \leq C_1 \tag{3.162}$$

for all mesh sizes $h > 0$.

If $u_h \to u$ *for* $h \to 0$ *with maximal order* k, *e.g. there is constant* $C_2 > 0$ *with*

$$C_2 h^k \leq \|u - u_h\|_{1,\Omega} \text{ for all } h > 0 \tag{3.163}$$

then the projecting a–posteriori error estimate $\eta_h^P \in V^{h,k}$ *defined by the discrete linear variational problem*

$$Q^{h,Q^{2k-1}}(v_{h;} \cdot \partial G(u^0_{h;}, .)\eta^P_{h;}) = -Q^{h,Q^{4k-2}}([\mathcal{I}^{2h,2k} v_h]_{;} \cdot G(u_{h;}, .)) \tag{3.164}$$

for all $v_h \in V^{h,k}$ *is equivalent to the true error* $e_h := u - u_h$ *in the sense of Definition 3, where* u_h^0 *is an arbitrary function in* $V^{h,k}$. *More exactly there is a constant* $C > 0$ *which depends only on the order* k *and the quadrature schemes with*

$$\frac{1}{C\sigma^3 D^4} \leq \liminf_{h \to 0+} \frac{\|\eta_h^P\|}{\|e_h\|} \leq \limsup_{h \to 0+} \frac{\|\eta_h^P\|}{\|e_h\|} \leq C\sigma^2 D^4 \ . \tag{3.165}$$

Therefore an effectivity index is given by the value $C\sigma^3 D^4 = \max(C\sigma^3 D^4, C\sigma^2 D^4)$.

Proof: It has to be verified that the propositions of Theorem 4 in the Banach space $V := H^1(\Omega)$ with $\lambda = 0$ hold if it is set

$$\begin{array}{l} V_h := V_{\bar{h}} := V^{h,k} \\ V_{h+} := V_h^c \oplus V^{h,k} \subset V^{h,2k} \text{ with} \\ V_h^c := V_0^{2h,2k} \end{array} \tag{3.166}$$

where the space $V_0^{2h,2k}$ is defined by equation (3.137). By Theorem 12 the discrete operator $F_h : V_h \to V_h^*$ defined by $u_h \to F_h(u_h)$ for all $u_h \in V_h$ with

$$< v_h, F_h(u_h) >:= Q^{h,Q^{2k-1}}(v_{h;} \cdot G(u_{h;}, .)) \tag{3.167}$$

for all $v_h \in V_h$ is well–posed with condition number $D^2 \cdot C_3$. Again by Theorem 12 the discrete operator $F_{h+} : V_{h+} \to V_{h+}^*$ defined by $u_{h+} \to F_{h+}(u_{h+})$ for all $u_{h+} \in V_{h+}$ with

$$< v_{h+}, F_{h+}(u_{h+}) >:= Q^{h,Q^{4k-2}}(v_{h+;} \cdot G(u_{h+;}, .)) \tag{3.168}$$

for all $v_{h+} \in V_{h+}$ is well–posed with condition number $D^2 \cdot C_4$. Moreover there is a solution $u_{h+} \in V_{h+}$ with $F_{h+}(u_{h+}) = 0$. Theorem 12 can be applied as

$V_{h+} \subset V^{h,2k}$ and the quadrature scheme Q^{4k-2} is exact of order $2 \cdot (2k) - 2$. The constants C_3 and C_4 are independent of the mesh size and the mesh quality σ_h.

It has to be proved that $(u_h, u_{h+})_{h>0}$ is saturated for the solution u in the sense of Definition 2. If it can be shown that

$$\|u - u_{h+}\|_{1,\Omega} \leq C_5 h^{k+1} \text{ for all } h > 0 \,. \tag{3.169}$$

Then condition (2.54) holds with $r_h := C_5 C_2^{-1} h$ because of assumption (3.163). Therefore $(u_h, u_{h+})_{h>0}$ is saturated for the solution u with saturation bound 0.

To verify estimate (3.169) Theorem 13 cannot be applied directly since the approximation space for the better solution u_{h+} and the quadrature scheme for the definition of the operator F_{h+} base on different triangulations. But it is possible to use a slight modification:

By Lemma 10 for $m := 0, 1$ and $q := 2$ there is $C_6 > 0$ with

$$\|u - \mathcal{I}^{2h,k+1} u\|_{1,\Omega} \leq C_6 h^{k+1} \text{ for all } h > 0 \tag{3.170}$$

(analogously to estimate (3.123)). In addition one obtains analogously to estimate (3.124) that $\mathcal{I}^{2h,k+1} u \in W^{k+2,\infty}(\mathcal{T}_h)$ with

$$\|\mathcal{I}^{2h,k+1} u\|_{2k+1,\infty,\mathcal{T}_h} \leq \|\mathcal{I}^{2h,k+1} u\|_{k+1,\infty,\mathcal{T}_h} \leq C_7 \tag{3.171}$$

where $C_7 > 0$ is a constant independent of h. The error functional $E^{4k-2} : C^0(T^0) \to \mathbb{R}$ defined by

$$E^{4k-2}(\varphi) := \int_{T^0} \varphi \, dx^0 - Q^{4k-2}(\varphi) \tag{3.172}$$

for all $\varphi \in C^0(T^0)$ is equal to zero for all polynomials of order $4k-2$. Taking Lemma 9 for $l := 4k-2$ and $k := k+1$ one gets

$$\begin{aligned} & \|F(\mathcal{I}^{2h,k+1} u) - F_{h+}(\mathcal{I}^{2h,k+1} u)\|_{V_{h+}^*} \\ \leq \quad & \sup_{v_{h+} \in V^{h,2k}} \frac{1}{\|v_{h+}\|_{1,\Omega}} |Q^{h,E^{4k-2}}(v_{h;} \cdot G((\mathcal{I}^{2h,k+1} u)_{;}, .))| \\ \overset{(3.111)}{\leq} \quad & C_8 h^{2k} \end{aligned} \tag{3.173}$$

for all mesh sizes $h > 0$ with a constant $C_8 > 0$ which is independent of the mesh size h. Since it is

$$\mathcal{I}^{2h,k+1} u \in V^{2h,k+1} \subset V_{h+} \tag{3.174}$$

and the estimates (3.170) and (3.173) hold the inequality (3.169) is proved by Corollary 1.

As next proposition (2.113) of Theorem 4 is verified: The error functional $E^{2k-1} \in C^0(T^0)^*$ defined by

$$E^{2k-1}(\varphi) := Q^{4k-2}(\varphi) - Q^{2k-1}(\varphi) \tag{3.175}$$

for all $\varphi \in C^0(T^0)$ is equal to zero for all polynomials of order $2k-1$. Using proposition (3.163) it is

$$\|u_h\|_{k+2,\infty,\mathcal{T}_h} = \|u_h\|_{k,\infty,\mathcal{T}_h} \leq C_1 \tag{3.176}$$

for all mesh sizes $h > 0$. Especially the discrete solution u_h belongs to the Sobolev space $W^{k+2,\infty}(\mathcal{T}_h)$. Lemma 9 is applied for $l := 2k-1$ and $k := k$ to obtain

$$\begin{aligned} & \|F_{h+}(u_h) - F_h(u_h)\|_{V_h^*} \\ = & \sup_{v_h \in V_h} \frac{1}{\|v_h\|_{1,\Omega}} |Q^{h,E^{2k-2}}(v_{h;} \cdot G(u_{h;}, .))| \\ \overset{(3.111)}{\leq} & C_9 h^{k+1} \end{aligned} \tag{3.177}$$

with a constant $C_9 > 0$ which is independent of the mesh size h. Applying proposition (3.163) the condition (2.113) holds with $s_h := C_9 C_2^{-1} h$. Actually it is $\lim_{h \to 0} s_h = 0$.

Corollary 4 shows that for any $u_h \in V^{h,k}$ the linear operator $L_{\bar{h}} := DF_h(u_h) : V_h \to V_h^*$ defined by

$$< v_h, DF_h(u_h) w_h >= Q^{h,Q^{2k-1}}(v_{h;} \cdot \partial G(u^0_{h;}, .) w_{h;}) \tag{3.178}$$

for all $v_h, w_h \in V_{\bar{h}} = V_h = V^{h,k}$ is an isomorphism from the space V_h to its dual space V_h^* with

$$\begin{aligned} \|L_{\bar{h}}\|_{\mathcal{L}(V_{\bar{h}}, V_{\bar{h}}^*)} & \leq C_{10} D \\ \|L_{\bar{h}}^{-1}\|_{\mathcal{L}(V_{\bar{h}}^*, V_{\bar{h}})} & \leq C_{10} D \end{aligned} \tag{3.179}$$

where $C_{10} > 0$ is a constant. It is independent of the mesh size h and the mesh quality σ_h.

Taking Lemma 14 a bound for the deflection in the Pythagorean equation defined by inequality (2.115) is given by $\kappa = C_{11} \cdot \sigma$. The constant $C_{11} > 0$ is independent of the mesh size and the mesh quality.

The operators

$$\begin{aligned} \mathcal{J}_{\bar{h}} & := \mathcal{I}^{h,k}|_{V_h^c} \\ \mathcal{J}_{h+} & := \mathcal{I}^{2h,2k}|_{V_{\bar{h}}} \end{aligned} \tag{3.180}$$

fulfil the condition $\mathcal{J}_{\bar{h}}[V_h^c] \subset V^{h,k} = V_{\bar{h}}$ and by Lemma 12 the condition $\mathcal{J}_{h+}[V_{\bar{h}}] \subset V^{2h,2k} \subset V_{h+}$. Therefore it is $\mathcal{J}_{\bar{h}} \in \mathcal{L}(V_h^c, V_{\bar{h}})$ and $\mathcal{J}_{h+} \in \mathcal{L}(V_{\bar{h}}, V_{h+})$. Moreover Lemma 13 turns out that

$$\begin{aligned} \|\mathcal{J}_{\bar{h}}\|_{\mathcal{L}(V_h^c, V_{\bar{h}})} &\leq C_{12}\sigma \\ \|\mathcal{J}_{h+}\|_{\mathcal{L}(V_{\bar{h}}, V_{h+})} &\leq C_{12}\sigma \ . \end{aligned} \tag{3.181}$$

The constant $C_{12} > 0$ is independent of the mesh size and the mesh quality. Since the global interpolation operators $\mathcal{I}^{h,k}$ and $\mathcal{I}^{2h,2k}$ use the same global degrees of freedom $X^{2h,2k} = X^{h,k}$ (see equation (3.69)) it is

$$\mathcal{I}^{2h,2k} \circ (\mathcal{I}^{h,k}|_{V^{2h,2k}}) = I_{V^{2h,2k}} \ . \tag{3.182}$$

Since it is $V_h^c \subset V^{2h,2k}$ the proposition (2.116) of Theorem 4 holds, too.

As all propositions of Theorem 4 and Corollary 2 have been verified the theorem has been proved •

Theorem 14 establishes that the projecting a–posteriori error estimate based on higher order approximation on a coarser mesh is equivalent to the true error in the sense of Definition 3 if the kernel G and the sought solution u are smooth enough. To calculate the projecting error estimate from equation (3.164) the variational problem (3.160) has to be evaluated with a quadrature scheme of higher exactness than used for the calculation of the discrete solution u_h of the discrete variational problem (3.161). It is not necessary to assemble a new stiffness matrix as the stiffness matrix in equation (3.164) is the same like the stiffness matrix in the Newton-Raphson iteration to calculate the discrete solution $u_h \in V_h$.

Remark 1: To simplify the formulation of the Theorem 14 it is assumed that the discrete variational problem (3.161) is solved exactly. Certainly Theorem 14 holds for the more general situation if the stopping criterion (2.56) with sufficiently small factor λ is used when solving the discrete variational problem (3.161). Naturally the bounds for the effectivity index are different.

Remark 2: It has to be pointed out that the quadrature scheme Q^{2k-1} used for the calculation of the discrete solution u_h is exact for polynomials of order $2k-1$ although it is sufficient that it is exact of order $2k-2$ to get a convergence of order k. The greater exactness is necessary to ensure that condition (2.113) holds with $\lim_{h\to 0} s_h = 0$ (see also the remark to Theorem 4). However, it is possible to use a quadrature scheme of exactness $2k-2$ to mount the stiffness matrix in the equation (3.164) when calculating the projecting error estimate.

Remark 3: It is essential to build the expansion of the space $V^{h,k}$ by piecewise polynomials of order $2k$ on the coarse mesh as it is then $X^{2h,2k} = X^{h,k}$,

i.e. the global degrees of freedom for order $2k$ on the coarse mesh are the same like the degrees of freedom for order k on the fine mesh. This condition is fundamental to prove equation (3.182) and proposition (2.116). The order of the polynomials has to be doubled to get a reliable projecting a–posteriori error estimate. If the order k is large this increases extremely the computational effort when mounting the right hand side for the linear system defining the error estimate as the kernel G has to be evaluated at plenty of quadrature nodes. Moreover stability problems can appear.

Remark 4: To prove that $(u_h, u_{h+})_{h>0}$ is saturated for the solution u it is essential that $u \in W^{k+2,\infty}(\Omega)$. If the solution u is not belonging to the Sobolev space $W^{k+2,\infty}(\Omega)$ there is the danger that the saturation bound is positive or even the situation occurs that $(u_h, u_{h+})_{h>0}$ is not saturated for the solution u. Example 2 in Section 4.5 illustrates that under these conditions the error estimate becomes fuzzy.

Remark 5: The propositions (3.163) and (3.162) have a very technical character. In practice both conditions are mostly fulfilled. Mainly proposition (3.163) says that the solution u is not a polynomial of order k. A handy criterion has been given by Babuška [7]. Condition (3.162) can be shown from $u \in W^{k+2,\infty}(\Omega)$ if the kernel G meets additional requirements.

3.7 Discussion

In this section the results of this chapter are compared with well–known a–posteriori error estimates. To simplify the presentation the discussion is restricted to the model problem (1.2) considered in the introducing Chapter 1 for the two dimensional case and the FEM approximation by piecewise linear polynomials $V^{h,1}$. Moreover it is assumed that all integrals are computed exactly, i.e. the error from the numerical integration is ignored. As shown in Chapter 1, see equation (1.9), the error $e_h := u - u_h$ is given by the variational problem

$$\begin{aligned} \int_\Omega \Big(a(\nabla v)(\nabla e_h) + bve_h \Big) dx = \\ - \int_\Omega \Big(a(\nabla v)(\nabla u_h) + (bu_h - f)v \Big) dx \text{ for all } v \in H^1(\Omega) \, . \end{aligned} \tag{3.183}$$

The right hand side of this error equation defines the residual of the discrete solution u_h which is a linear functional on the space $H^1(\Omega)$. One possibility to interpret some a–posteriori error estimates is the approach to estimate the norm of this residual functional, see Verfuerth [64]. An alternative view, that

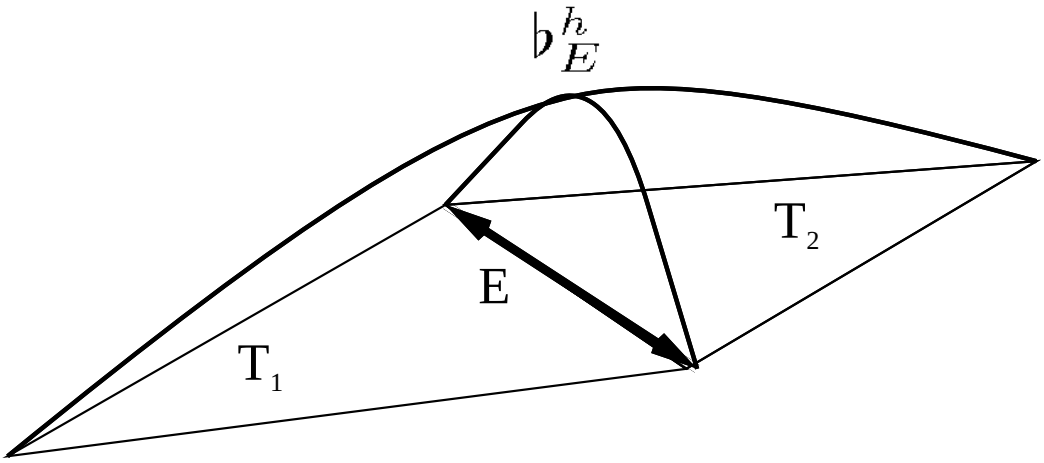

Figure 3.8: The edge bubble function $\flat_E^h$ on edge E on the fine mesh $\mathcal{T}_h$.

is followed here, is the approximative solution of the error equation (3.183). The error estimates vary in the selection of the used approximation space and the method to solve the discretized error equation.

For the assumptions made here the error equation (3.164) defining the projecting error estimate $\eta_h^P \in V^{h,1}$ is the following:

$$\begin{aligned}\int_\Omega \Big(a(\nabla v_h)(\nabla \eta_h^P) + bv\eta_h^P\Big)dx = \\ -\int_\Omega \Big(a(\nabla[\mathcal{I}^{2h,2}v_h])(\nabla u_h) + (bu_h - f)[\mathcal{I}^{2h,2}v_h]\Big)dx\end{aligned} \tag{3.184}$$

for all $v_h \in V^{h,1}$. It is obvious that the residual functional for the discrete solution u_h is only evaluated for the space $\mathcal{I}^{2h,2}[V^{h,1}] = V^{2h,2}$, i.e. for the space of piecewise quadratic functions on the coarse grid, instead of the whole space $H^1(\Omega)$. On the other hand the error equation is not solved in $V^{2h,2}$ but is interpolated to the space of piecewise linear functions on the fine mesh. This can be interpreted as a reduction step in a multi–level procedure going from a quadratic to a linear approximation. The difference to standard multi–level methods is that the global degrees of freedom are kept which has the effect that high frequencies can be represented as well on the lower level as on the higher level.

A similar idea can be found in the hierarchical error estimate technique, see Zienkiewicz [69], Deufelhard [31], Bank [15]. Here Deufelhard's representation is quoted:

The error equation (3.183) is solved on the space

$$V_{h+}^H := V^{h,1} \oplus V_h^H \text{ with } V_h^H := span\{\flat_E^h\}_{E\in\mathcal{E}_h} \tag{3.185}$$

where $\mathcal{E}_h$ denotes the set of the edges of the elements in the triangulation $\mathcal{T}_h$. The function $\flat_E^h$ is the *edge bubble function* for the edge $E \in \mathcal{E}_h$ which is a continuous, piecewise quadratic function on the triangles T_1 and T_2 sharing the edge E and has the value one at the middle point of the edge E, see Figure 3.8. The expansion of the usual nodal basis of piecewise linear functions by the edge bubble functions $\{\flat_E^h\}_{E\in\mathcal{E}_h}$ can be taken as a hierarchical basis of order 2 for the space V_{h+}^H. The solution of the error equation in space V_{h+}^H requires an assemblage of a new stiffness matrix for the hierarchical basis and the solution of a linear system with a higher effort than needed for the calculation of the discrete solution u_h. Therefore the stiffness matrix is reduced to its diagonal. The significant solution components of this simplified linear system are given by

$$\eta_E^H := -\frac{1}{B(\flat_E^h, \flat_E^h)} \int_\Omega \Big(a(\nabla \flat_E^h)(\nabla u_h) + (bu_h - f)\flat_E^h) \Big) dx \tag{3.186}$$

for all edges $E \in \mathcal{E}_h$ where the bilinear form $B : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ is defined by

$$B(v_1, v_2) := -\int_\Omega \Big(a(\nabla v_1)(\nabla v_2) + bv_1 v_2 \Big) dx \text{ for all } v_1, v_2 \in H^1(\Omega) \,. \tag{3.187}$$

The value η_E^H delivers an estimation for the discretization error at the middle point of the edge $E \in \mathcal{E}_h$. It can be proved that the a–posteriori error estimate

$$\eta_h^H := \sum_{E\in\mathcal{E}_h} \eta_E^H \flat_E^h \tag{3.188}$$

is equivalent to the true error in the sense of Definition 3, see Deufelhard [31], Bank [15]. (Remark: Theorem 4 can be used to get this result). For the proof two assumptions are needed: the saturation condition in the sense of Definition 2 and the fact that the bilinear form B defined by equation (3.187) is symmetric and positive definite. This last condition restricts the application of the hierarchical error estimate drasticly as in many applications, especially for non–linear problems, the involved bilinear form B is neither symmetric nor positive definite. However, Bornemann [18] and Verfuerth [64] have shown that the error estimator η_h^H cannot be expressed in terms of the Babuška–Miller residual error estimator which is equivalent to the true error without using any saturation condition (if the material functions a and b and the right hand side f are piecewise constant, see Babuška [7]). Therefore the saturation condition is an indispensable assumption for the hierarchical a–posteriori error estimate η_h^H.

It is obvious that the space $V_h^c = V_0^{2h,2}$ added to the approximation space $V^{h,1}$ to define the expansion V_{h+} for the projecting error estimate can be

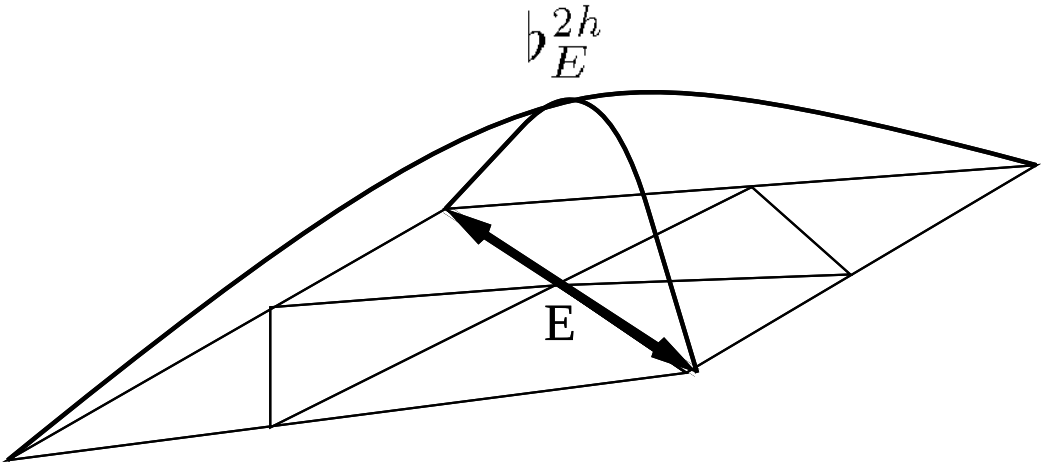

Figure 3.9: The edge bubble function $\flat_E^{2h}$ on edge E on the coarse mesh $\mathcal{T}_{2h}$.

expressed by using edge bubble functions:

$$V_0^{2h,2} = span\{\flat_E^{2h}\}_{E \in \mathcal{E}^{2h}} \ . \tag{3.189}$$

Differing from the space V_{h+}^H used to define the hierarchical error estimate η_h^H the space $V^{h,1}$ is expanded by edge bubble functions on the coarse mesh $\mathcal{T}_{2h}$ instead of using the fine mesh $\mathcal{T}_h$, see Figure 3.9. The construction of the space $V_0^{2h,2}$ builds up the macro–elements in the coarse mesh $\mathcal{T}_{2h}$ from the elements in fine mesh $\mathcal{T}_h$. At the end no new node coordinates have to be generated as the middle points of the edges in the coarse mesh that are associated with the edge bubble functions are vertices of elements in the fine mesh. However, it is

$$V_{h+} = V^{h,1} + V_0^{2h,2} \subset V_{h+}^H \ . \tag{3.190}$$

This confirms that the projecting error estimate η_h^P and the hierarchical error estimate η_h^H are closely related. The property (3.190) shows that it is not possible to express the projecting error estimate in terms of the Babuška–Miller residual error estimator. Therefore the saturation condition in the sense of Definition 2 has to be assumed to prove that the projecting error estimate is equivalent to the true error in the sense of Definition 3.

The reflection on the hierarchical error estimate has emphasized some advantages of the projecting error estimate compared to other error estimates. The calculation of the projecting error estimate is possible as for most FEM schemes a higher order scheme and a suitable interpolation operator from the lower to the higher order scheme is available. In this sense the projecting

error estimate is defined for most FEM problems without additional assumptions, like symmetry or positivity. Since the higher order scheme uses the double polynomial order than the lower order scheme stability problems can occur if the projecting error estimate is applied to a FEM approximation based on polynomials of high order.

The special quality of the projecting error estimate is the new idea to reuse the stiffness matrix for the calculation of the a–posteriori error estimate. This approach saves the assemblage of a new stiffness matrix. Moreover there is the possibility to profit from the reordering of the matrix, the LU-factorization or the preconditioner matrix, that were set up to calculate the discrete solution u_h efficiently, for a second time when solving the error equation.

3.8 Summary

The projecting a–posteriori error estimate for the finite element method of order k is equivalent to the true error in the sense of Definition 3 if polynomials of order $2k$ for the error estimate are used. The effectivity index indicating the quality of the error estimate depends on the condition number D^2 of the Jacobi matrix of the kernel G and the mesh quality σ of the finite element mesh. This holds under the assumption that the sought solution is smooth enough. As proved in Corollary 2 the effectivity index is increased for a non-smooth solution u and the a-posteriori error estimate becomes more fuzzy.

Though only the homogeneous Neumann boundary value problem has been discussed the result is also valid for non–homogeneous Neumann boundary value problems including homogeneous Dirichlet boundary conditions and systems of such boundary value problems. The reason is that for these problem types a modified uniform positivity of the kernel G involved in the formulation of the boundary value problems can be specified as well. Non–homogeneous Dirichlet boundary conditions can be considered by introducing a Lagrangean multiplier which produces a saddle–point problem.

Saddle–point problems, e.g. see Brezzi [21], are not belonging to the class of problems investigated in this chapter but they can be treated by the abstract analysis of Chapter 2. The essential problem is to verify that the involved operators are well–posed. It is necessary to use various polynomial orders for the components of the solution. However the projecting a–posteriori error estimate can be applied to this kind of problems, too. By using Theorem 4

it can be shown that this projecting error estimate is also equivalent to the true error (see Example 4 in Section 4.7).

Further extensions consider curved domains. That requires the introduction of curved elements with non-affine parametrical representations (e.g. isoparametrical elements). In principle the results of this chapter can be adapted with some modifications considering the bending of the elements, see e.g. Ciarlet [24, 25]. If the projecting error estimate shall consider the error from the approximation of the domain the formulation of the algorithm, especially the definition of the global refined triangulation, as well as the analysis becomes more difficult but the line of the thoughts and the results remain mostly unchanged.

Chapter 4

Examples

4.1 Introduction

In the following some examples are presented to demonstrate the projecting a-posteriori error estimate for piecewise linear FEMs based on an expanding by piecewise quadratic polynomials in practice. For the calculations a modified version of the program package VECFEM [38] is used, see Section 4.2. The first example is a very smooth problem to get a feeling for the actual values of the effectivity index given in Theorem 14. In the second example the influence of the smoothness of the solution on the effectivity index is examined. In the third example the calculation of the displacements of a loaded linear elastic body is presented. To illustrate that the projecting error estimate also works for saddle–point problems the fourth example is the solution of the two–dimensional Navier–Stokes equations.

4.2 The VECFEM Program Package

VECFEM [38] is a program package to solve non–linear variational problems by the finite element method. The solution can have more than one component. The user can select between isoparametrical elements up to order three and mixed finite elements of arbitrary order on lines, quadrilaterals, triangles, hexahedrons, prisms and tetrahedrons. The variational problem has to be entered in the formulation (3.50). Among other terms surface integrals can be additionally introduced to consider non–homogeneous Neumann boundary conditions. Moreover Dirichlet boundary conditions can be

considered. The variational problem and Dirichlet boundary conditions are specified symbolically. A code generator transforms the variational problem into a FORTRAN program which solves the problem by calling suitable routines of the VECFEM library. In particular the code generator calculates the Jacobi matrix ∂G by using the computer algebra program MAPLE [22].

By introducing a suitable basis of V_h the discrete variational problem is reduced to the system of non–linear equations (2.29). It is solved by the Newton–Raphson method (2.41). This requires the assemblage of the stiffness matrix (2.37) by evaluating the Jacobi matrix ∂G. For the numerical integration on lines, quadrilaterals and hexahedrons Gaussian quadrature schemes are used. The quadrature schemes for triangles, prisms and tetrahedrons are constructed by transforming the Gaussian quadrature schemes in a suitable manner. The mounted linear system is solved by the program package LINSOL [65]. LINSOL uses iterative methods of the conjugate gradient type. The stopping criterion for LINSOL is optimally set by VECFEM to compute the Newton–Raphson correction with a minimal number of conjugate gradient steps to not destroy the quadratic convergence order of the Newton–Raphson.

The basis of the approximation space $V^{h,k}$ is constructed by a basis for the polynomial space P_k on the n-simplex. This local basis is assembled to a global basis of $V^{h,k}$ by using the parametrical representations of the elements in the triangulation $\mathcal{T}_h$. The local basis is defined by a table that gives the values of the basis functions and their first derivatives at the integration nodes. Therefore it is very simple to modify the basis and the quadrature scheme for the space $V^{h,k}$ without changing other parts of the code. More details are presented in Grosz [40].

A simple implementation of the projecting error estimate for FEM approximations of order two could be found by exchanging the standard table of VECFEM: FEM data for an order two approximation on a coarse triangulation $\mathcal{T}_{2h}$ are handed over. By using piecewise linear polynomials on subelements, see Figure 4.1, a FEM approximation of order one on the refined triangulation $\mathcal{T}_h$ is calculated. The right hand side of the error equation (3.164) bsed on the order two method on the coarse triangulation $\mathcal{T}_{2h}$ is assembled by using the original FEM data. In detail this procedure works as follows:

When implementing the FEM approximation and its projecting error estimate by error equation (3.164) the main difficulty arises from the fact that two triangulations are needed, namely the triangulation $\mathcal{T}_h$ for the FEM approximation and the coarse triangulation $\mathcal{T}_{2h}$ for the error estimation. But Lemma 10 allows to interpret the space $V^{h,1}$ as a FEM space based on the

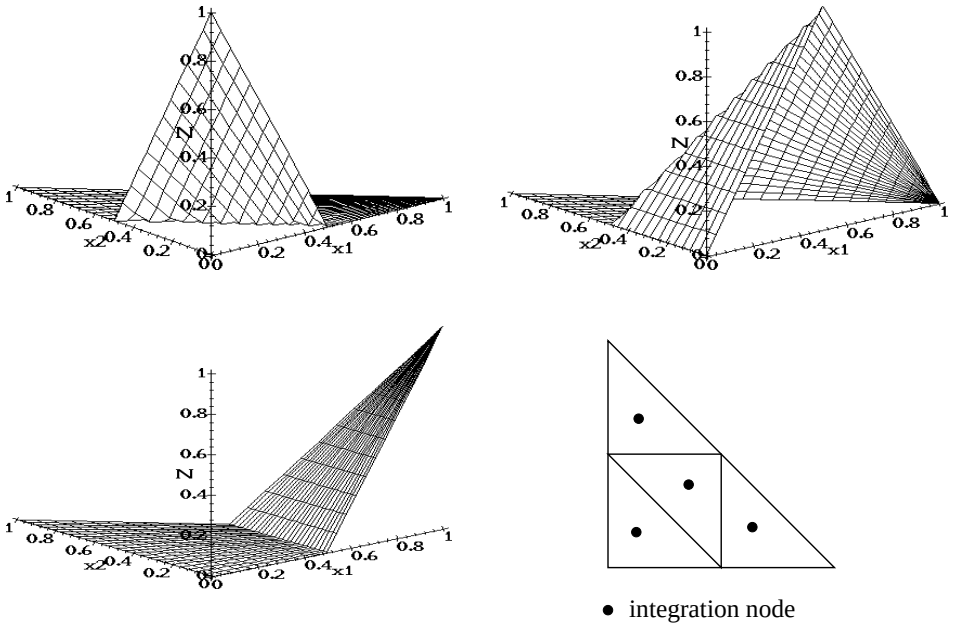

Figure 4.1: Three local basis functions of the space S_1 on the 2-simplex and integration nodes for a quadrature scheme that is exact of order 1.

coarse triangulation $\mathcal{T}_{2h}$ with local basis of the space S_1 defined by equation (3.132). Therefore for this test implementation of the projecting error estimate it is assumed that FEM data are handed over to VECFEM which are normally used to construct an approximation by piecewise polynomials of order two on a coarse triangulation $\mathcal{T}_{2h}$. Yet the local basis for polynomials of order two belonging to this FEM data is replaced by a basis for the space S_1. This can be done since both spaces have the same dimension. Figure 4.1 shows three of the needed six local basis functions for the space S_1 of piecewise linear functions on the 2-simplex.

For a given local quadrature scheme Q^l and the triangulation $\mathcal{T}_h$ the global quadrature scheme has to be calculated by formula (3.67). Because of the identity (3.131) the global quadrature scheme can be interpreted as a global quadrature scheme of the coarse triangulation $\mathcal{T}_{2h}$ and a local quadrature scheme that is composed by quadrature schemes on the sets of the subdivision $\mathcal{T}_0$ of the n-simplex defined by equation (3.128). The composed quadrature scheme is constructed by formula (3.67) where the subdivision $\mathcal{T}_0$ plays the role of the triangulation $\mathcal{T}_h$. Figure 4.1 shows the location of the integration nodes for a composed quadrature scheme that is exact of order 1 in the sense of Definition 4.

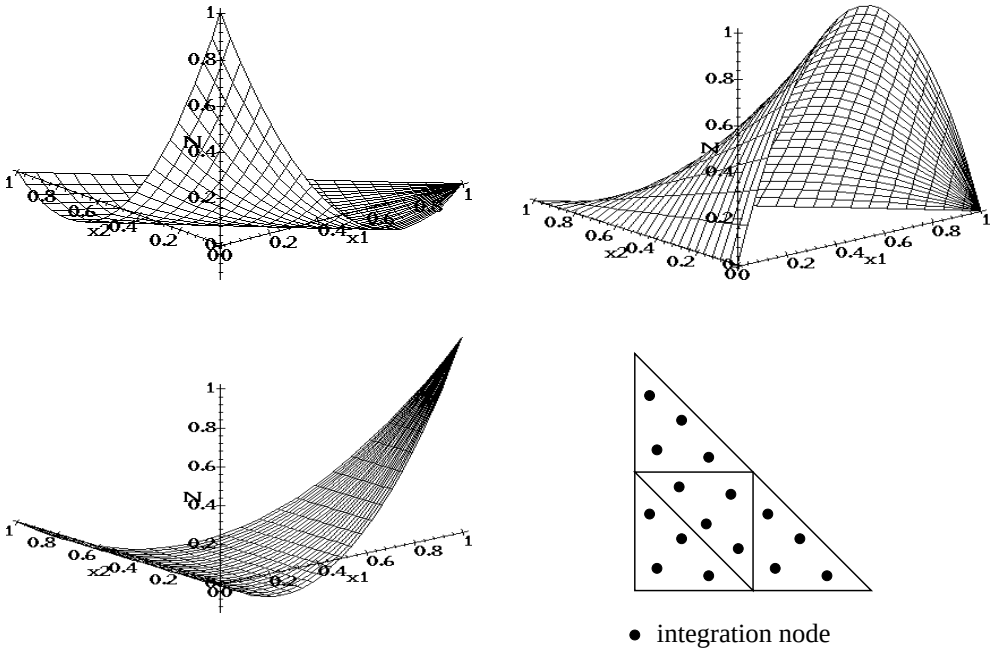

Figure 4.2: $\mathcal{I}_2$ interpolation of the basis of the space S_1 shown in Figure 4.1 and integration nodes for a quadrature scheme that is exact of order 3.

After the FEM approximation in the space $V^{h,1}$ has been calculated the right hand side of the error equation (3.164) has to be assembled to calculate the projecting error estimate. For that purpose the image of the basis of the space $V^{h,1}$ for the global interpolation operator $\mathcal{I}^{2h,2}$ has to be specified. By applying formula (3.76) this global interpolation is reduced to an interpolation of the local basis of S_1 by the local interpolation operator $\mathcal{I}^2$. Therefore the assembling routine of VECFEM can be used for the right hand side of the error equation if the local basis is selected to the $\mathcal{I}^2$-interpolation of the basis of the space S_1. As above the quadrature scheme is constructed by a composed quadrature scheme. Figure 4.2 shows the $\mathcal{I}^2$-interpolation of the S_1–basis functions shown in Figure 4.1 and the location of the integration nodes that are exact for polynomials of order two on the subelements.

The new linear system of the new right hand side and the stiffness matrix used for the solution of the discrete variational problem is solved by LINSOL to get the projecting error estimate. In the same manner FEM data for an order 4 method on a coarse triangulation $\mathcal{T}_{2h}$ could be processed. In this case a FEM approximation of order two on the refined triangulation $\mathcal{T}_h$ is calculated by using a local basis of piecewise polynomials of order two. The right hand side for the error equation is assembled by using the given FEM data for the

order 4 method on the coarse triangulation $\mathcal{T}_{2h}$. It is obvious that in any case the selected implementation based on the coarse triangulation is not the most efficient way to implement the projecting error estimate. However, it is a simple way as one does not need to write a new FEM code and it is sufficient to check if the projecting error estimate works successfully.

4.3 Common Terms

The following test problems are solved by finite element approximations of order one with various mesh sizes. The meshes are generated by hand or by the commercial mesh generator I–DEAS [44]. The small problems are solved on a workstation IBM RS6000 and large scale problems on a vector computer Fujitsu VPP300. Some comparative computations for problems of medium order ($\approx$ 5000 unknowns) have shown that the results are independent of the used platform.

The discrete variational problems are solved with an accuracy $TOL = 10^{-10}$ on the level of solution, see Grosz [38, 39]. This very small accuracy ensures that the error from terminating the Newton–Raphson iteration can be neglected compared to the discretization error. The stopping criterion (2.136) presented in Corollary 3 is used when solving the equation (3.164) that defines the projecting error estimate. For all problems $\tau = 10^{-4}$ is set.

Since the exact solutions of the example problems are known the dependence of the $\|.\|_{1,\Omega}$-norm of the true error (that is the difference of the exact solution and the calculated FEM approximation) on the mesh size is presented. The values of the errors are the absolute errors, i.e. they are not scaled by the norm of the solution. In the diagrams logarithmic scales for the mesh sizes and the true errors are used. In a second diagram the ratio of the $\|.\|_{1,\Omega}$-norm of the projecting error estimate and the $\|.\|_{1,\Omega}$-norm of the true error with a linear scale is shown. In all diagrams the actually measured values are marked by points. Points which are connected by lines are produced by meshes with approximately the same mesh quality, see equation (3.63).

4.4 Example 1: The Model Problem

The first example should get the actual order of the effectivity index given in Theorem 14:

Let $\Omega := [0,1]^n$ be the n-dimensional unit cube ($n = 1, 2, 3$). The sought solution $u : \Omega \to \mathbb{R}$ is determined by the linear Neumann boundary value problem

$$\begin{aligned} -\Delta u + u - f &= 0 \quad \text{on} \quad \Omega \\ \frac{\partial u}{\partial n} &= 0 \quad \text{on} \quad \partial\Omega \ . \end{aligned} \tag{4.1}$$

$\frac{\partial u}{\partial n}$ denotes the derivative of the function u with respect of the outer normal of the boundary $\partial\Omega$ of the domain Ω. The function $f : \Omega \to \mathbb{R}$ is determined by the exact solution

$$u(x) = \prod_{i=1}^{n} \cos(m_i \pi x_i) \text{ for all } x = (x_i)_{i=1,n} \in \Omega \tag{4.2}$$

with fixed $(m_1, \ldots, m_n) \in \mathbb{N}_0^n$. The value $m := m_1 + \ldots + m_n$ scales the number of oscillations of the function u. The $H^1(\Omega)$-norm of the solution is in the order of $(\pi m)^n$.

Actually the boundary value problem is solved in its weak formulation: find the solution $u \in H^1(\Omega)$ with

$$\int_\Omega (u - f)v + \sum_{i=1}^{n} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} \, dx = 0 \text{ for all } v \in H^1(\Omega) \ . \tag{4.3}$$

In the notations of Chapter 3 it is set

$$G(\zeta, x) = (\zeta_1 - f(x), \zeta_2, \ldots, \zeta_{n+1}) \tag{4.4}$$

for all $x \in \Omega$ and all $\zeta = (\zeta_i)_{i=1,n+1} \in \mathbb{R}^{n+1}$. This kernel G is uniform positive definite with a positivity bound 1. For the construction of the finite element space the n-dimensional unit cube is subdivided into n-simplexes basing on a rectangular grid. For the three as well as for the two dimensional case the values for the mesh qualities s are in the order of 4.

In the first test the dependence of the true error on the mesh size is investigated. Various numbers of oscillations of the solution indicated by the value m are selected to inspect the influence of the solution on the results. Figure 4.3 shows the dependence of the true error on the mesh size for the 1-dimensional case, Figure 4.4 for the 2-dimensional case and Figure 4.5 for

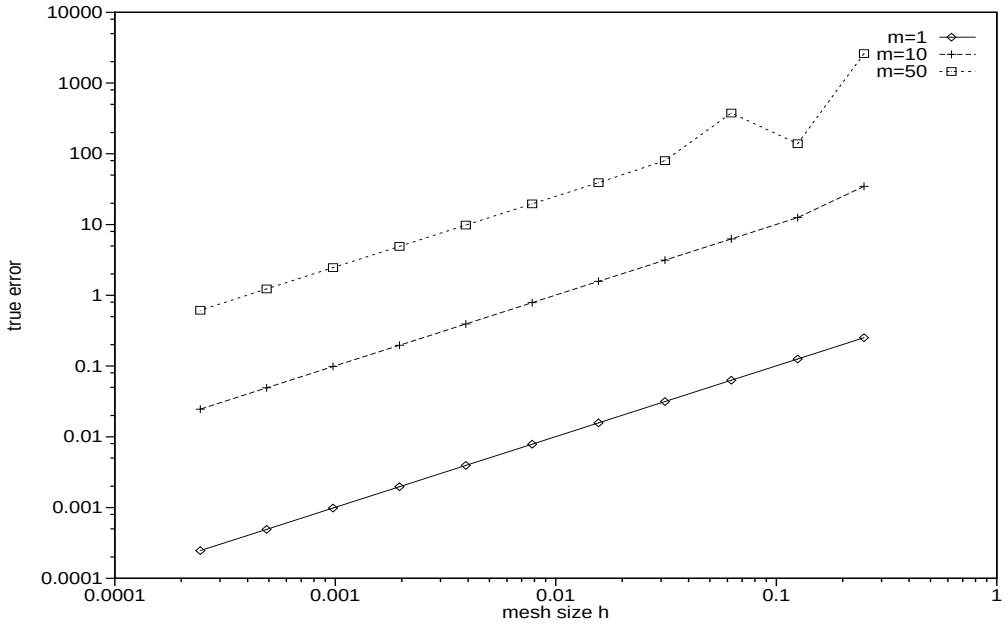

Figure 4.3: Example 1, $n = 1$: The true errors in the $H^1(\Omega)$-norm for various numbers of solution oscillations m.

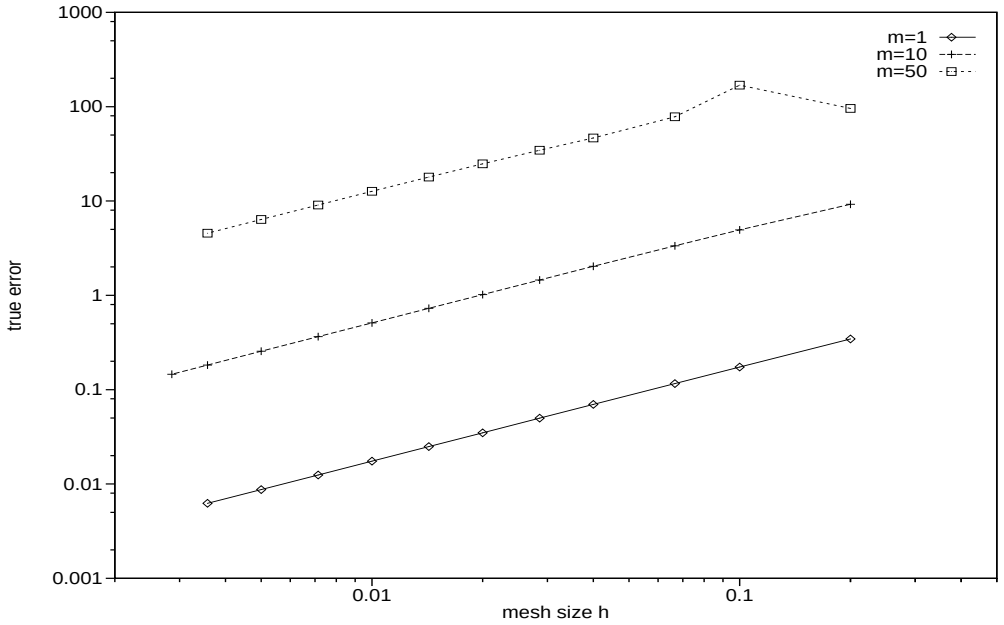

Figure 4.4: Example 1, $n = 2$: The true errors in the $H^1(\Omega)$-norm for various numbers of solution oscillations m.

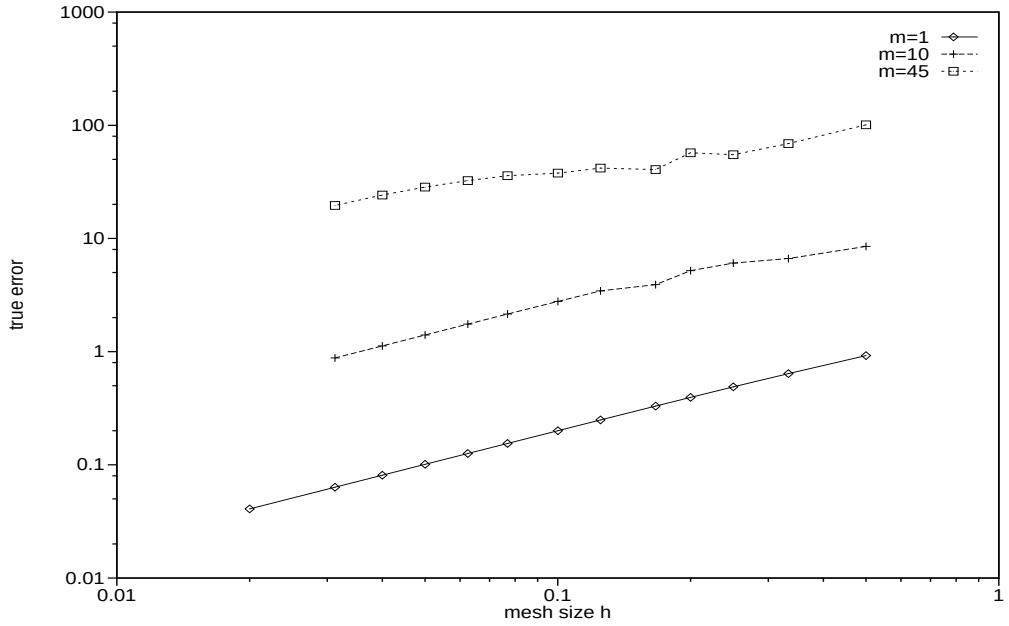

Figure 4.5: Example 1, $n = 3$: The true errors in the $H^1(\Omega)$-norm for various numbers of solution oscillations m.

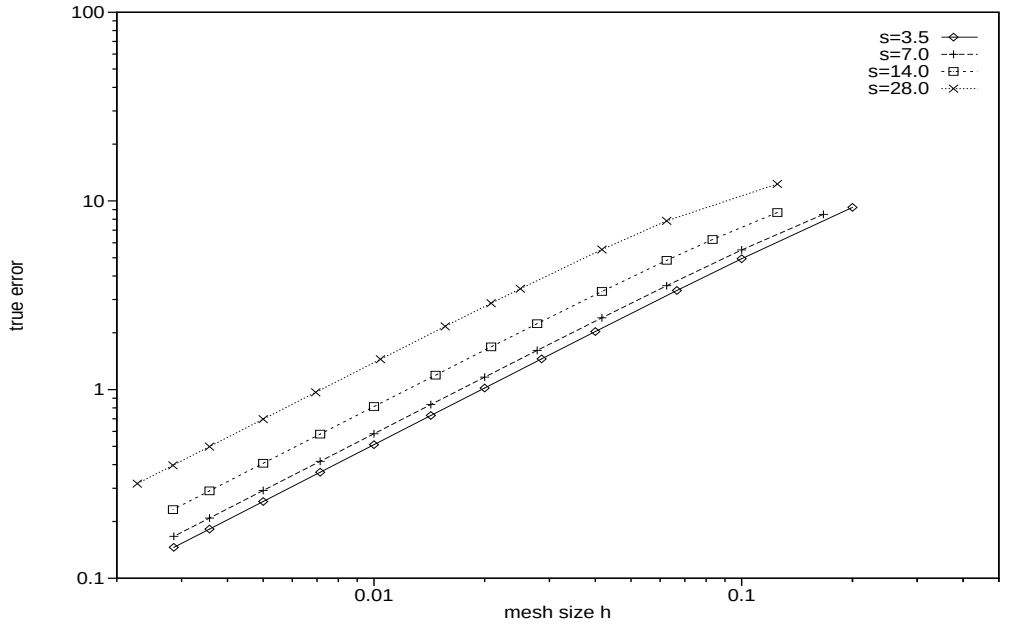

Figure 4.6: Example 1, $n = 2$: The true errors in the $H^1(\Omega)$-norm for various mesh qualities s ($m = 10$).

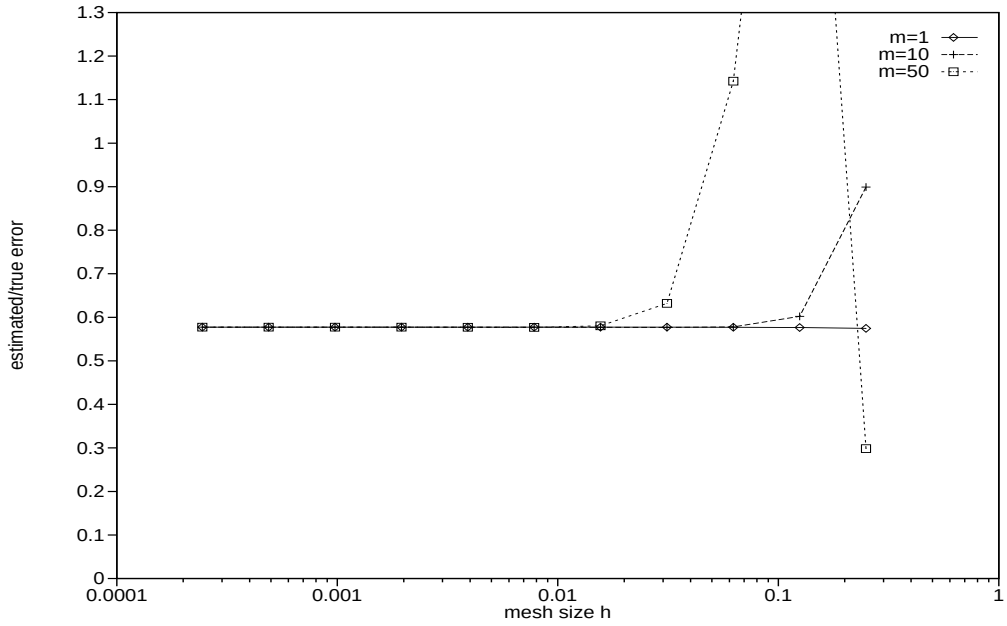

Figure 4.7: Example 1, $n = 1$: The ratios of estimated and true errors in the $H^1(\Omega)$-norm for various numbers of solution oscillations m.

the 3-dimensional case. Corresponding to the result of Theorem 13 the true errors converge to zero if the mesh size goes to zero. The convergence order is independent of the solution and the mesh quality (see Figure 4.6). On the other hand the actual error depends on the solution as well as the mesh quality. The tests confirm the well–known behavior of the FEM.

The Figures 4.7 to 4.10 present the dependencies of the ratio of the estimated and true error in the $H^1(\Omega)$–norm on the mesh size for the projecting a-posteriori error estimate. As shown in Figures 4.7 and 4.8 the ratios of estimated and true error seem to converge to the value $0.577 \approx \frac{\sqrt{3}}{3}$ for a one or two–dimensional domain. Even if the mesh quality is increased this value does not change, see Figure 4.10. For a three dimensional domain the situation is undetermined since the computational effort to process FEM meshes with a mesh size less than 0.01 exceeds the limit of the available computer capacity. But the results give no counterargument to conclude that the ratios of estimated and true errors converge to a value in the order of 0.577, too. Summarizing these tests it has to be stated that at least for small mesh sizes the projecting a–posteriori error estimate underestimates the true error by a factor in the order of 0.577.

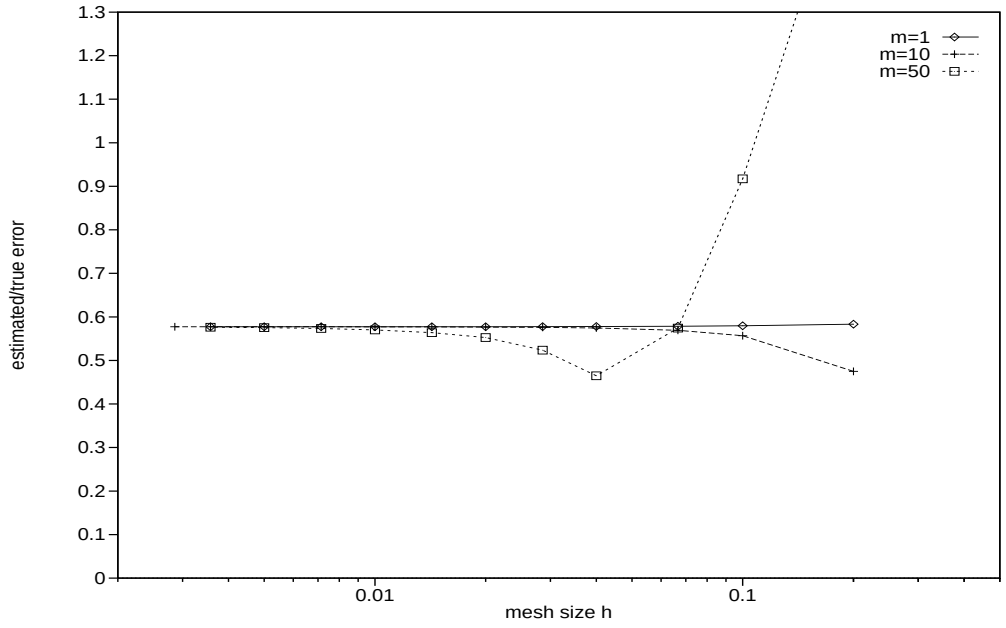

Figure 4.8: Example 1, $n = 2$: The ratios of estimated and true errors in the $H^1(\Omega)$-norm for various numbers of solution oscillations m.

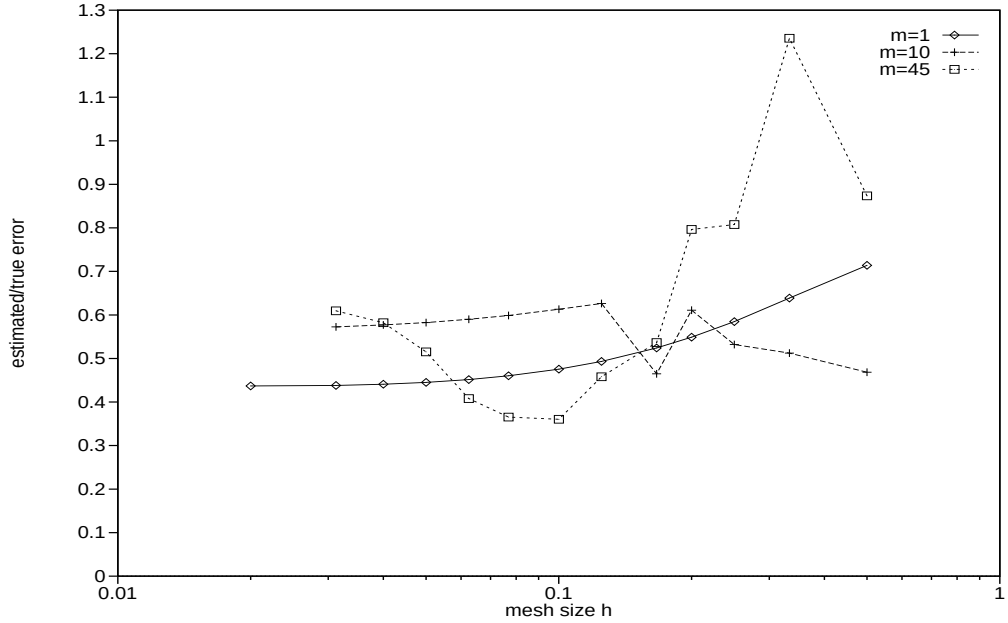

Figure 4.9: Example 1, $n = 3$: The ratios of estimated and true errors in the $H^1(\Omega)$-norm for various numbers of solution oscillations m.

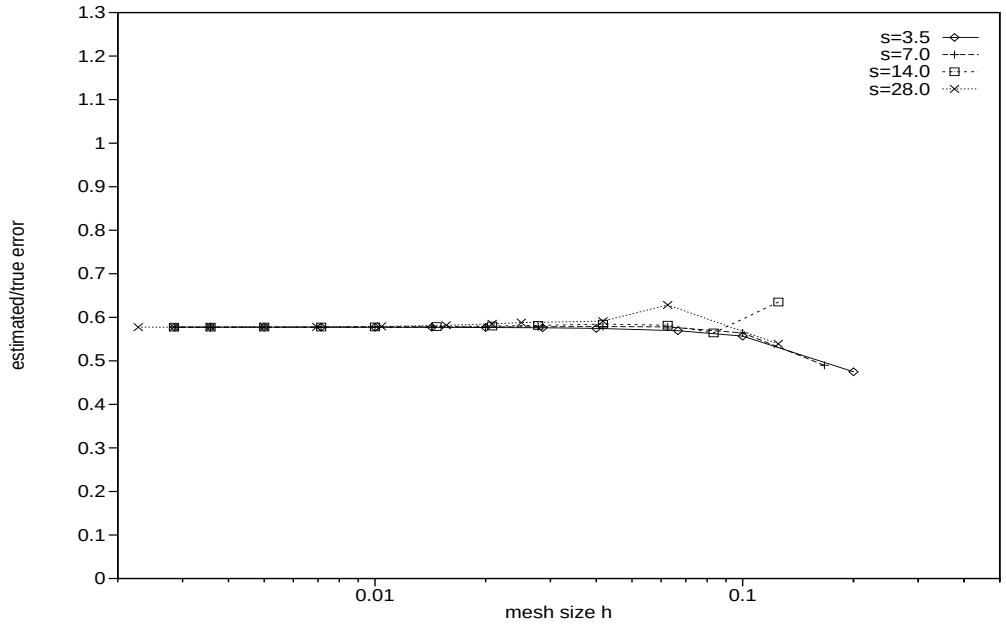

Figure 4.10: Example 1, $n = 2$: The ratios of estimated and true errors in the $H^1(\Omega)$-norm for various mesh qualities s ($m = 10$).

4.5 Example 2: Singularities

Now the dependence of the effectivity index on the smoothness of the solution is investigated. The domain is the two-dimensional, L-shaped domain

$$\Omega := [-1,1]^2 \backslash [-1,0]^2 \ , \tag{4.5}$$

see Figure 4.11. It is set

$$?_N = [-1,0] \times \{0\} \cup \{0\} \times [-1,0] \tag{4.6}$$

and $?_D := \partial\Omega \backslash ?_N$. The test problem is the Poisson equation with Neumann and Dirichlet conditions for the sought solution $u : \Omega \to \mathbb{R}$:

$$\begin{aligned} -\Delta u + f &= 0 \quad \text{on} \quad \Omega \\ u &= 0 \quad \text{on} \quad ?_D \\ \frac{\partial u}{\partial n} &= 0 \quad \text{on} \quad ?_N \ . \end{aligned} \tag{4.7}$$

f is a given function on the domain Ω.

The corresponding weak formulation is given by the variational problem on the space $H_0^1(\Omega) := \{v \in H^1(\Omega) \mid v|_{\Gamma_D} = 0\}$: find the solution $u \in H_0^1(\Omega)$

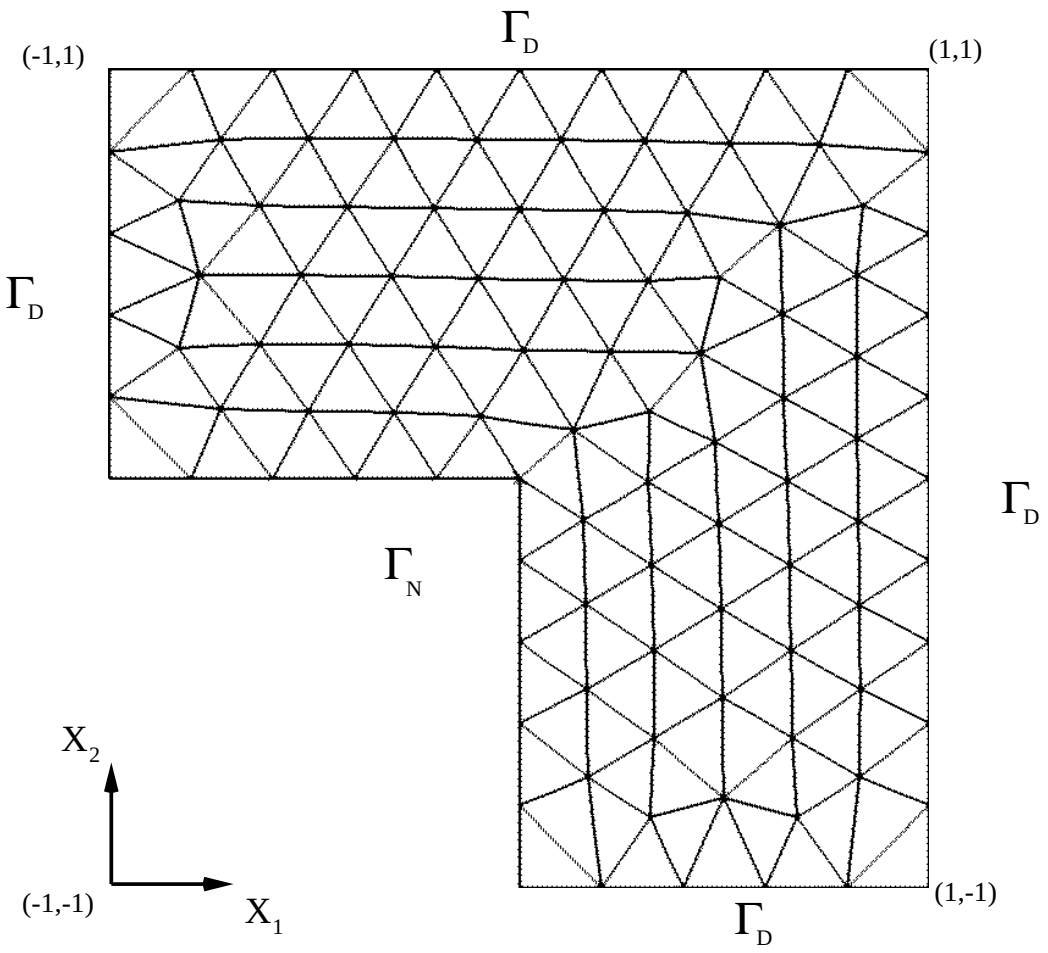

Figure 4.11: Example 2: L-shaped domain with triangulation.

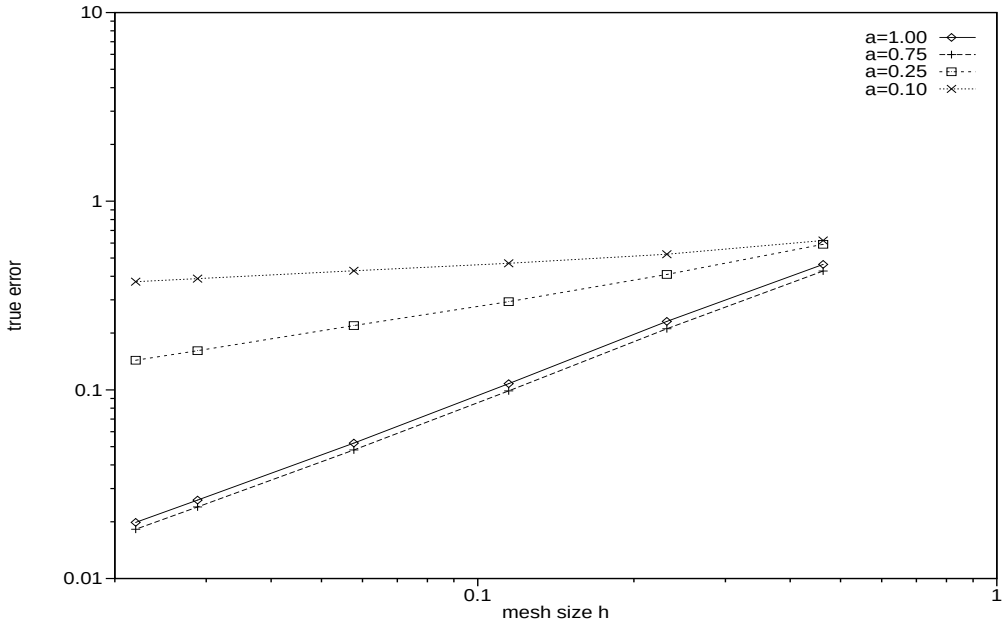

Figure 4.12: Example 2: The true error in the $H^1(\Omega)$-norm for various powers a indicating the smoothness of the solution.

with

$$\int_\Omega (fv + \frac{\partial u}{\partial x_1}\frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2}\frac{\partial v}{\partial x_2})\, dx = 0 \text{ for all } v \in H_0^1(\Omega)\,. \tag{4.8}$$

The function f is determined by the exact solution

$$u(x_1, x_2) = (x_1^2 + x_2^2)^a (x_1^2 - 1)(x_2^2 - 1) \tag{4.9}$$

for all $(x_1, x_2) \in \Omega$ ($a \in \mathbb{R}$). Depending on the value for the power a the solution u has a singularity at $(0, 0)$. The solution u belongs to the space $H_0^1(\Omega)$ if the power a is positive or equal to zero. The solution u belongs to $H^2(\Omega)$ if the power a is greater than $\frac{1}{2}$ or equal to zero. If $a < 0$ the resulting right hand side $f = -\Delta u$ does not belong to $H^0(\Omega)$ and therefore the variational problem (4.8) is not properly formulated. Triangulations of the domain were generated by the commercial mesh generator I-DEAS [44], see Figure 4.11.

In Figure 4.12 the true errors of a series of meshes with decreasing mesh size and almost constant mesh quality are shown. The abscissa gives the mean value of the element size. If the solution belongs to $H^2(\Omega)$, that is for $a = 0.75$ and $a = 1.00$, the convergence order is actually of order 1 but for $a = 0.25$ and $a = 0.10$ the convergence order declines since the solution is not smooth enough ($\notin H^2(\Omega)$).

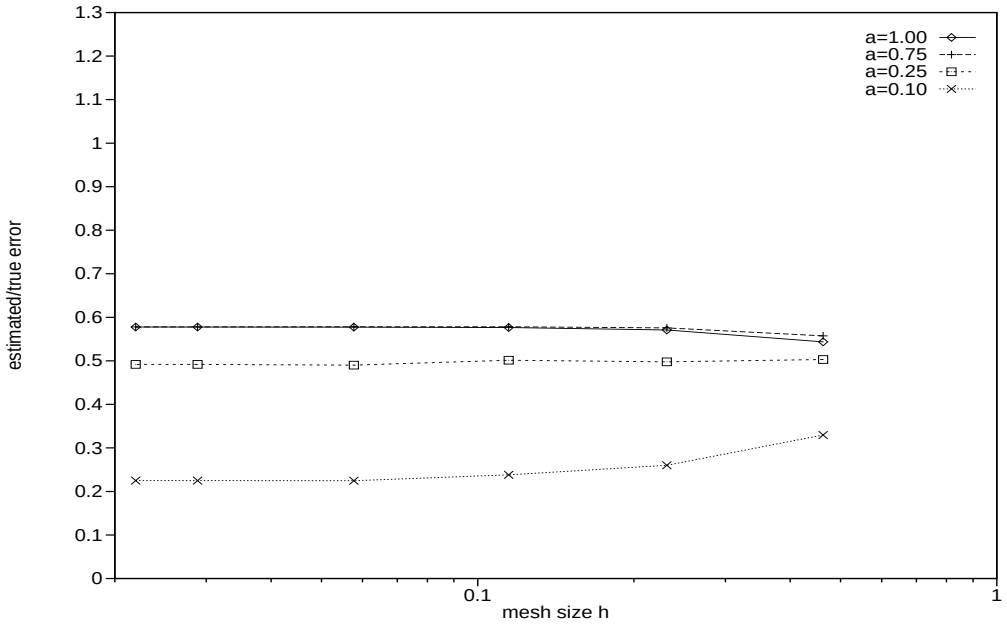

Figure 4.13: Example 2: The ratio of estimated and true error in the $H^1(\Omega)$-norm for various powers a indicating the smoothness of the solution.

Naturally the ratio of the estimated and true error for the projecting error estimate is the most interesting value in the test, see Figure 4.13. Analogously to Example 1 in Section 4.4 the ratio of the estimated and the true error seems to converge to a specific limit but the value of this limit depends on the smoothness of the solution u. For the case $a > \frac{1}{2}$ tested by $a = 1.00$ and $a = 0.75$ the approximations from the expansion V_{h+} by polynomials of order 2 have a higher convergence order to the solution u than the approximations u_h from the space V_h. Yet the convergence order will not be equal to 2 as the solution does not belong to the Sobolev space $H^3(\Omega)$ but it is greater than one. Therefore $(u_h, u_{h+})_{h>0}$ is saturated for the solution u with saturation bound 0. This is the reason why the ratio of true and estimated error converges to the limit ≈ 0.577 that appeared before in Example 1. The situation changes if the value of a is less than $\frac{1}{2}$. For $a = 0.25$ or $a = 0.1$ the saturation bound is not equal to zero since the addition of piecewise polynomials of order two to the approximation space V_h cannot improve the convergence order for $h \to 0$. The projecting error estimate becomes more inaccurate as predicted in Corollary 2. The actual value of the saturation bound cannot be determined by a practical calculation as it is very expensive to modify the VECFEM code to solve the discrete variational problem on V_{h+}.

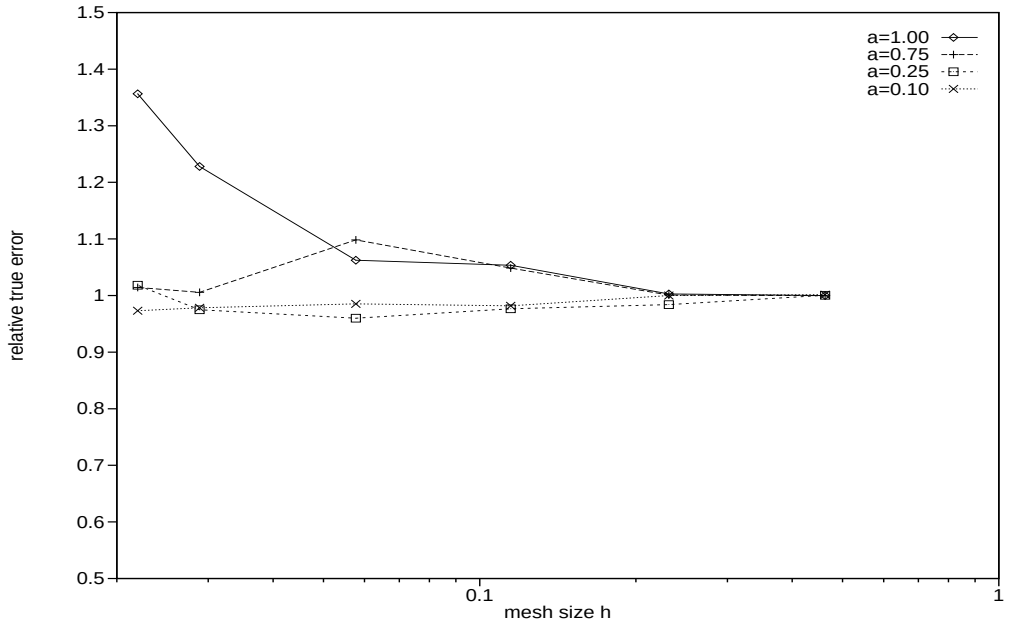

Figure 4.14: Example 2: True errors in the $H^1(\Omega)$-norm by the optimal stopping criterion relative to the true errors in the $H^1(\Omega)$-norm when using accuracy $TOL = 10^{-10}$.

In a second test the optimal stopping criterion (2.117) is investigated. The calculation is repeated with the optimal stopping criterion set to $\lambda = 0.075$. Figure 4.14 shows the ratio of the true errors when using the optimal stopping criterion and using the high accuracy of $TOL = 10^{-10}$. As seen in this figure the ratio is in the order of one. That means that the errors for the solutions computed with a high accuracy and with the optimal stopping criterion have nearly the same value. For $a = 1$ and a small mesh size the ratio increases up to 1.4 since the optimal stopping criterion prevents VECFEM to execute the next Newton-Raphson step. When applied in practice this deviation is acceptable. Further refinement of the mesh would admit the execution of this additional iteration step. This could not be tested since the needed number of elements exceeds the limit of the available I–DEAS installation.

Figure 4.15 shows the ratio of the CPU-time using the optimal stopping criterion and the high accuracy of $TOL = 10^{-10}$ on a Fujitsu VPP300. For both calculations the zero function is the initial guess for the Newton–Raphson iteration. Although the evaluation of the optimal stopping criterion requires the assemblage of a second right hand side in every iteration step the usage of this criterion saves more than 60% of the computing time. Naturally it is questionable whether an accuracy $TOL = 10^{-10}$ is reasonable. Moreover the

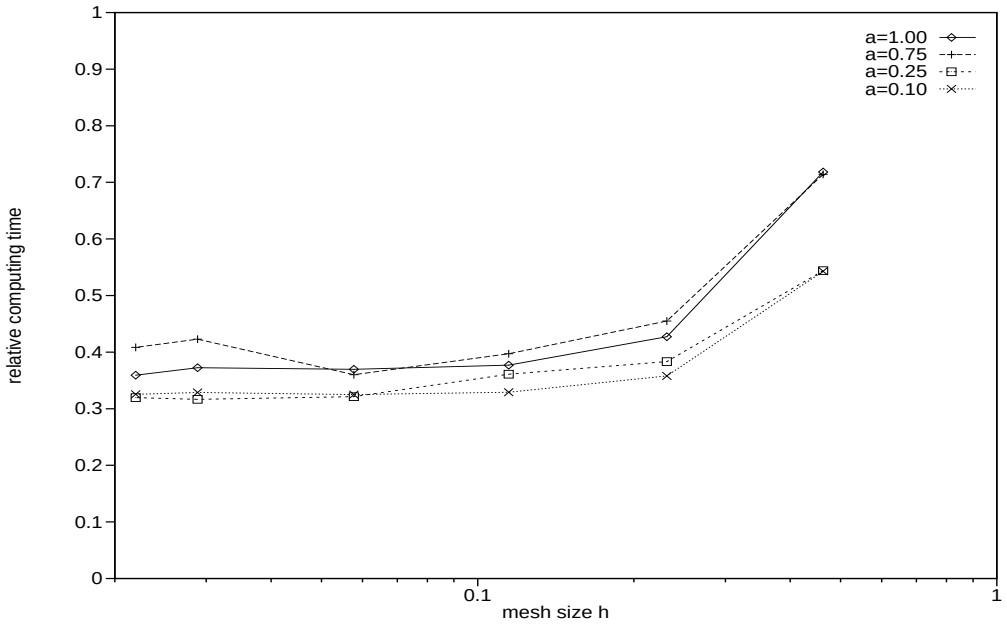

Figure 4.15: Example 2: Computing time when using the optimal stopping criterion relative to the computing time when using accuracy $TOL = 10^{-10}$.

implementation is not optimal and therefore the actual profit may be less.

4.6 Example 3: Structural Analysis

To illustrate that the projecting a–posteriori error estimate works also for systems of boundary value problems the equations of the linear elasticity are examined, e.g. see Dawe [29], Zienkiewicz [68]:

The displacement $u = (u_1, u_2, u_3)$ of a linear elastic body Ω under the action of internal and external forces is determined. The vector $\varepsilon(u)$ of the (linearized) strains is defined by

$$\begin{aligned} \varepsilon(u) &:= (\varepsilon_1(u), \varepsilon_2(u), \varepsilon_3(u), \gamma_{12}(u), \gamma_{23}(u), \gamma_{13}(u)) \\ &:= (\frac{\partial u_1}{\partial x_1}, \frac{\partial u_2}{\partial x_2}, \frac{\partial u_3}{\partial x_3}, \frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1}, \frac{\partial u_2}{\partial x_3} + \frac{\partial u_3}{\partial x_2}, \frac{\partial u_1}{\partial x_3} + \frac{\partial u_3}{\partial x_1}) \, . \end{aligned} \tag{4.10}$$

For linear elastic and isotropic material the stress vector

$$\sigma(u) := (\sigma_1(u), \sigma_2(u), \sigma_3(u), \tau_{12}(u), \tau_{23}(u), \tau_{13}(u)) \tag{4.11}$$

is calculated from the strain vector $\varepsilon(u)$ by Hooke's law:

$$\begin{array}{rcl} \sigma_1(u) &=& C_{11}\ \varepsilon_1(u) + C_{12}\ \varepsilon_2(u) + C_{12}\ \varepsilon_3(u) \\ \sigma_2(u) &=& C_{12}\ \varepsilon_1(u) + C_{11}\ \varepsilon_2(u) + C_{12}\ \varepsilon_3(u) \\ \sigma_3(u) &=& C_{12}\ \varepsilon_1(u) + C_{12}\ \varepsilon_2(u) + C_{11}\ \varepsilon_3(u) \\ \tau_{12}(u) &=& C_{44}\ \gamma_{12}(u) \\ \tau_{23}(u) &=& C_{44}\ \gamma_{23}(u) \\ \tau_{13}(u) &=& C_{44}\ \gamma_{13}(u)\ . \end{array} \tag{4.12}$$

The parameters C_{11}, C_{12} and C_{44} are material constants depending on the modules of elasticity and Poisson's ratio. In this example it is set $C_{11} = 1$, $C_{12} = \frac{1}{2}$ and $C_{44} = \frac{1}{4}$ corresponding to the non–dimensionalzed modelling of steal. The stress vector has to fulfil the equilibrium condition

$$\begin{array}{rcl} \dfrac{\partial\sigma_1(u)}{\partial x_1} + \dfrac{\partial\tau_{12}(u)}{\partial x_2} + \dfrac{\partial\tau_{13}(u)}{\partial x_3} - f_1 &=& 0 \\ \dfrac{\partial\tau_{12}(u)}{\partial x_1} + \dfrac{\partial\sigma_2(u)}{\partial x_2} + \dfrac{\partial\tau_{23}(u)}{\partial x_3} - f_2 &=& 0 \\ \dfrac{\partial\tau_{13}(u)}{\partial x_1} + \dfrac{\partial\tau_{23}(u)}{\partial x_2} + \dfrac{\partial\sigma_3(u)}{\partial x_3} - f_3 &=& 0\ . \end{array} \tag{4.13}$$

The function $f = (f_1, f_2, f_3)$ denotes the vector of internal forces (e.g. gravitation). Via the equations (4.10) and (4.12) the equilibrium condition is a system of three partial differential equations of order two for the sought displacement u. To make the solution of the equilibrium condition (4.13) unique boundary conditions have to be set. Boundary conditions for the stress introducing external surface loads are boundary conditions of the Neumann type. Restraint conditions prescribing values for the displacement are boundary conditions of the Dirichlet type.

The weak formulation of the boundary value problem arising from the equilibrium condition (4.13) and the boundary conditions is given in the following form: Set

$$V := \{(v_1, v_2, v_3) \in H^1(\Omega)^3 \mid v_1|_{\Gamma_1} = 0, v_2|_{\Gamma_2} = 0, v_3|_{\Gamma_3} = 0\}\ . \tag{4.14}$$

The sets $?_1, ?_2, ?_3 \subset \partial\Omega$ denote the locations of the restraint conditions for the displacement. The sought solution $u \in V$ is given by the variational problem

$$\begin{array}{rcl} \displaystyle\int_\Omega (& \sigma_1(u)\varepsilon_1(v) + \sigma_2(u)\varepsilon_2(v) + \sigma_3(u)\varepsilon_3(v) & + \\ & \tau_{12}(u)\gamma_{12}(v) + \tau_{13}(u)\gamma_{13}(v) + \tau_{23}(u)\gamma_{23}(v) & + \\ & f_1 v_1 + f_2 v_2 + f_3 v_3)\ dx & \\ + & \displaystyle\int_{\partial\Omega} (p_1 v_1 + p_2 v_2 + p_3 v_3)\ d? & = 0 \end{array} \tag{4.15}$$

for all $v \in V$. The function $p = (p_1, p_2, p_3) : \partial\Omega \to \mathbb{R}^3$ describes the surface loads. If $?_1 = ?_2 = ?_3$ has a positive surface measure it can be shown by Korn's inequality that the operator on $V \subset H^1(\Omega)^3$ involved in the variational problem (4.15) is well–posed, see Fichera [35].

The domain of the test problem is the tetrahedron with a unit triangle as base-surface and height 1.5. The tetrahedron has been rotated in a way that it is standing on the vertex with the most acute degree and this vertex is the origin, see Figure 4.16. Triangulations were generated by I–DEAS [44]. At the point $(0, 0, 0)$ the mesh size is the fourth part of the mesh size at the opposite face. The displacements u_1 and u_2 are prescribed to be zero at all vertices of the tetrahedron. The displacement u_3 is only prescribed at the origin. The internal force f and the surface load p are set by the given displacement

$$\begin{aligned} u_1 &= x_1(1.5 - x_3) \\ u_2 &= x_2(1.5 - x_3) \\ u_3 &= 1 - e^{x_3} \ . \end{aligned} \tag{4.16}$$

Figure 4.17 shows the convergence of the true errors to zero for decreasing mesh size. The ratio of estimated and true error shown in Figure 4.18 seems to converge to a value in the order of 0.65. Therefore the projective error estimate is equivalent to the true error for the solution of systems of boundary value problems as well. As one has already realized in the foregoing examples the projecting error estimate underestimates the true error. The corrective factor seems lightly to deviate from the known value 0.577.

4.7 Example 4: Navier-Stokes Equations

The velocity field $u := (u_1, u_2)$ of an incompressible Newtonian fluid in a domain Ω is the solution of the Navier–Stokes equations. In the non–dimensionalized formulation this is a system of three partial differential equations:

$$\begin{aligned} -\Delta u + R_e(u^T \cdot \nabla)u - \nabla p &= f \\ \nabla^T \cdot u &= 0 \end{aligned} \tag{4.17}$$

on the domain Ω. The unknown function $p : \Omega \to \mathbb{R}$ denotes the pressure. R_e is called the Reynolds number. The function $f = (f_1, f_2)$ describes an internal load working on the fluid. For both velocity components Dirichlet boundary conditions are set on the total boundary $\partial\Omega$ of the domain Ω. As the pressure is unique apart from a constant a norming condition for the

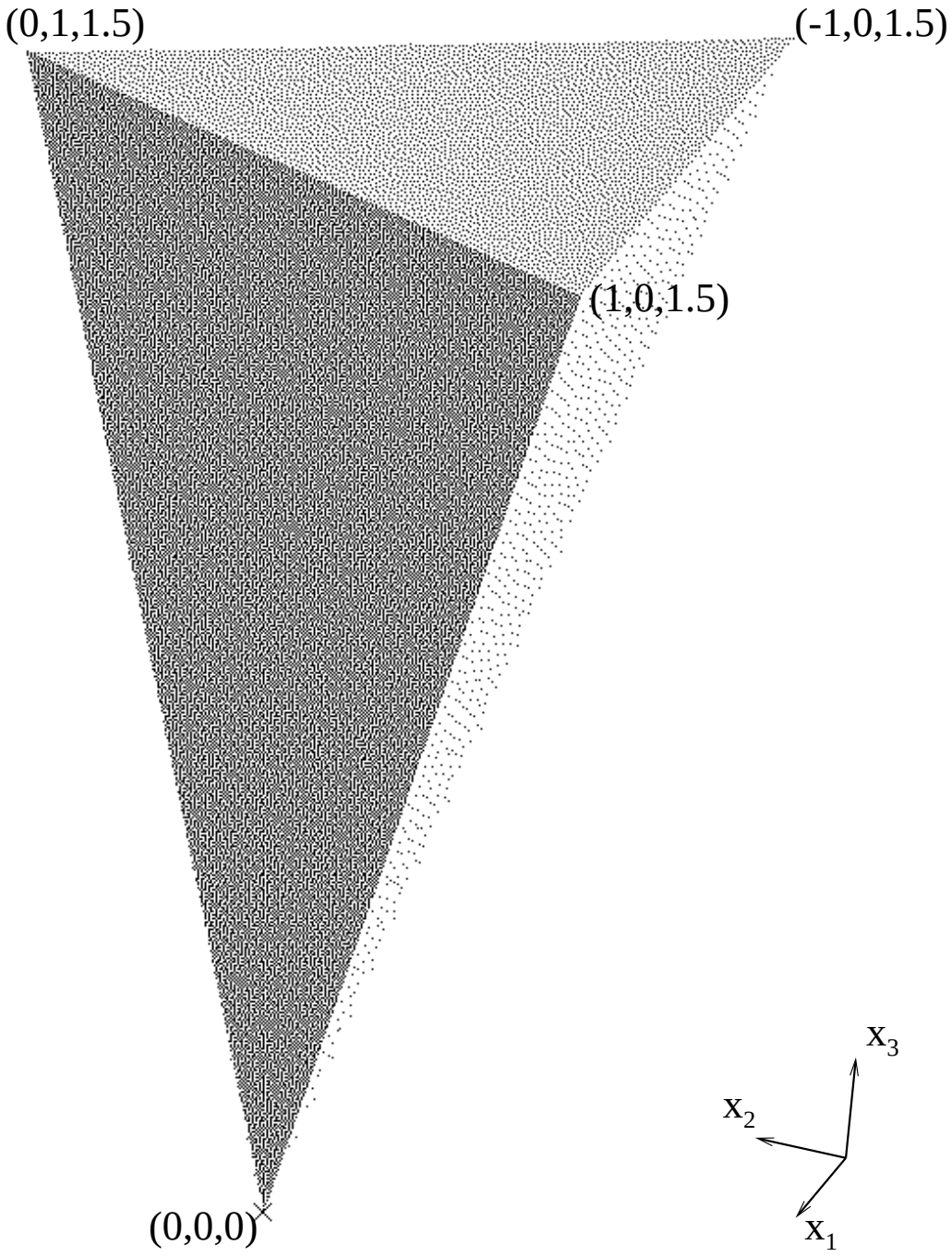

Figure 4.16: Example 3: Tetrahedron standing at one of its vertices.

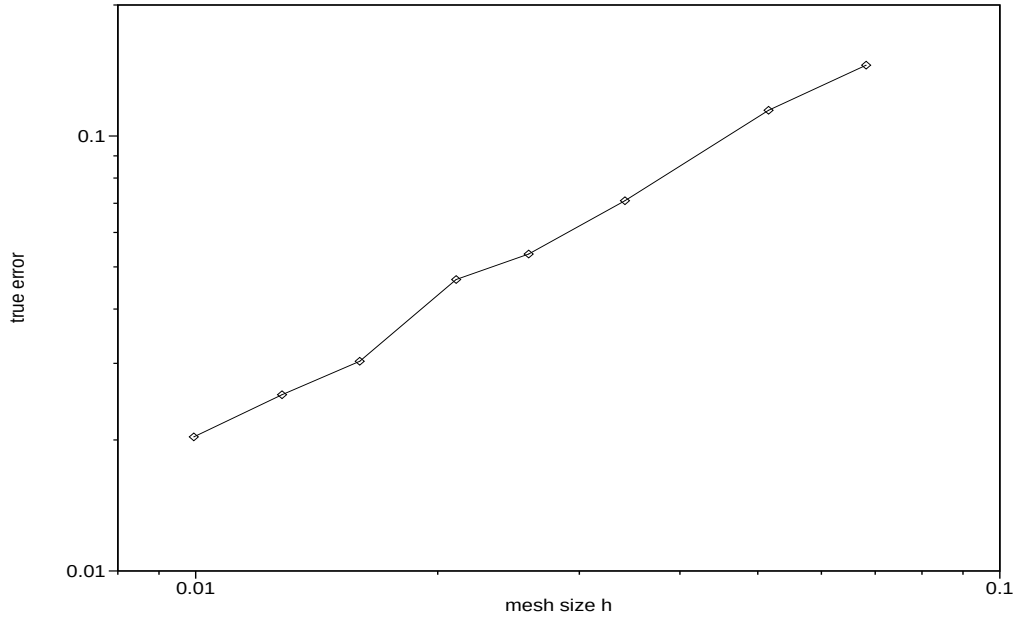

Figure 4.17: Example 3: The true error in the $H^1(\Omega)^3$-norm.

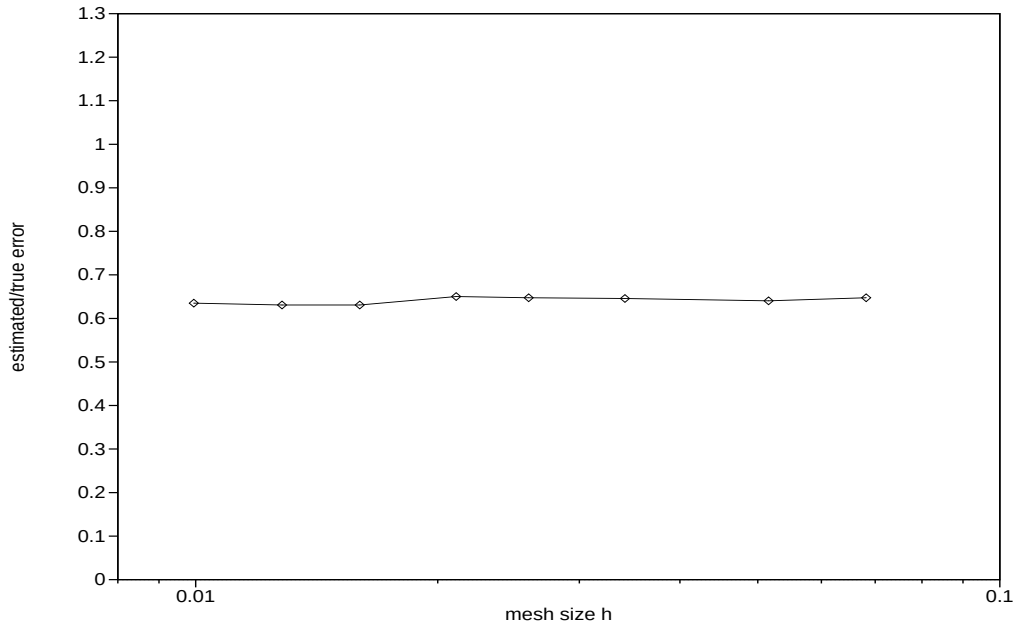

Figure 4.18: Example 3: The ratio of estimated and true error in the $H^1(\Omega)^3$-norm.

pressure has to be set, e.g.

$$\int_\Omega p \, dx = 0 \; . \tag{4.18}$$

Surveys on the numerical solution of the Navier–Stokes–equations by finite elements are given by Gunzenburger [42] and Chung [23] and a mathematical analysis is given by Girault [37].

The weak formulation of the Navier-Stokes equation (4.17) is given by the non–linear variational problem: find $(u, p) \in X \times Y$ with

$$\begin{aligned} \int_\Omega \; & \Big(\frac{\partial v_1}{\partial x_1}(\frac{\partial u_1}{\partial x_1} + p) + v_1 \, (R_e(u_1 \frac{\partial u_1}{\partial x_1} + u_2 \frac{\partial u_1}{\partial x_2}) - f_1) \quad + \\ & \frac{\partial v_2}{\partial x_2}(\frac{\partial u_2}{\partial x_2} + p) + v_2 \, (R_e(u_1 \frac{\partial u_2}{\partial x_1} + u_2 \frac{\partial u_2}{\partial x_2}) - f_2) \quad + \\ & (\frac{\partial v_1}{\partial x_2} + \frac{\partial v_2}{\partial x_1})(\frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1}) + q \, (\frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2}) \Big) \, dx \qquad = 0 \end{aligned} \tag{4.19}$$

for all $(v, q) \in X \times Y$. The vector space $X \times Y$ is defined by

$$Y := \{q \in H^0(\Omega) \mid \int_\Omega q \, dx = 0\} \tag{4.20}$$

and

$$X := \quad \{v \in H^1(\Omega)^2 \mid v_1|_{\partial\Omega} = v_2|_{\partial\Omega} = 0\} \; . \tag{4.21}$$

The variable q in the variational problem (4.19) is the Lagrangean multiplier to consider the continuity condition $\nabla^T \cdot u = 0$.

For Reynolds number $R_e = 0$ the problem becomes a linear variational problem called Stokes problem. Using the well-known analysis of saddle–point problems it can be shown that the operator involved in the variational problem of the Stokes problem is well-posed, see Brezzi [21]. Unfortunately the construction of suitable finite element approximation spaces is more difficult than for the previous examples. The selected spaces have to fulfil the Ladyzhenskaya-Babuška-Brezzi (or LBB) condition to ensure that the discrete operator is well-posed, see Brezzi [21]. Typically the approximation order of the pressure has to be one order less than the approximation order of the velocity. The general variational problem (4.19) has an unique solution for sufficiently small forces f and sufficiently small Reynolds numbers R_e only. In this case the properties of the Stokes problem are also valid. For larger forces or greater Reynolds numbers special solution techniques have to be used when solving the Navier–Stokes equations.

The test domain is the channel $[0, 3] \times [0, 1]$ of length 3 and height 1. The force f is selected in such a way that the exact solution of the Navier-Stokes

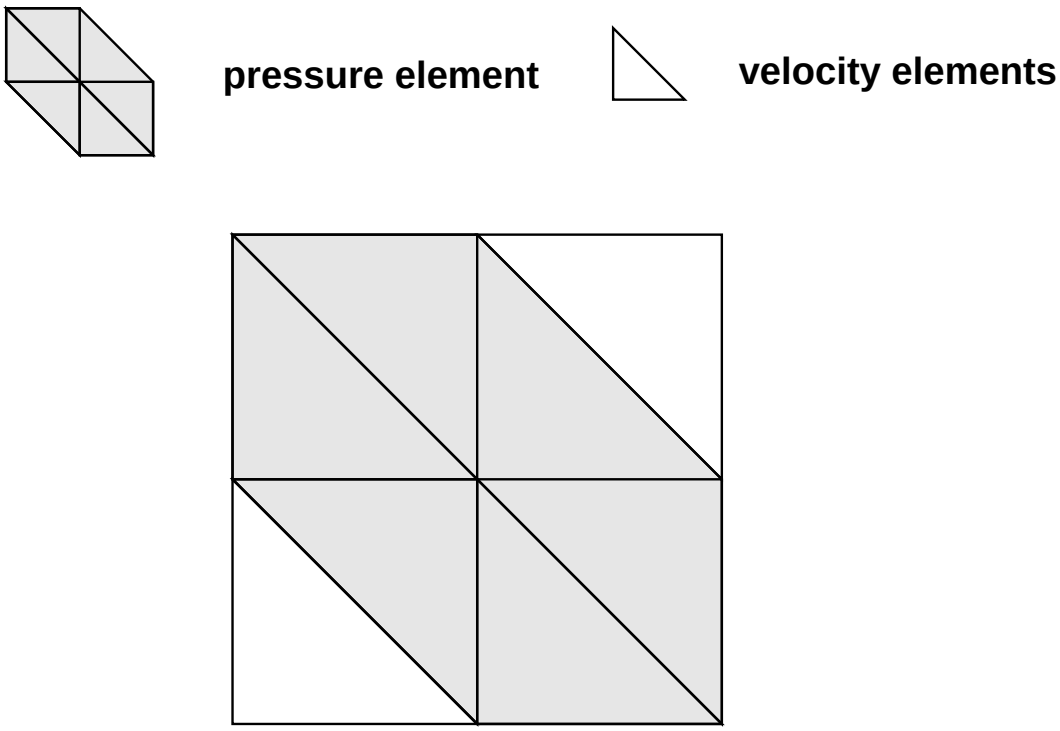

Figure 4.19: Example 4: Macro–element for pressure approximation.

equations is given by

$$\begin{aligned} u_1 &:= \frac{\partial \psi}{\partial x_2} \\ u_2 &:= -\frac{\partial \psi}{\partial x_1} \\ p &:= 0 \end{aligned} \tag{4.22}$$

where the stream function ψ is defined by

$$\psi(x_1, x_2) := (\frac{4}{3})^4 \left(x_1 \left(3 - x_1\right) x_2 \left(1 - x_2\right)\right)^2 . \tag{4.23}$$

From the definition it is evident that the selected velocity u fulfills the continuity condition $\nabla^T \cdot u = 0$ and the boundary condition $u_1|_{\partial\Omega} = u_2|_{\partial\Omega} = 0$. The treated problem does not describe a physical fluid (as the channel has only walls and there is no inlet or outlet) but it is a test problem.

For the approximation of the velocity polynomials of order one are used. The pressure is approximated by piecewise constant polynomials on macro–elements composed by six elements that are used for the velocity approximation, see Figure 4.19. In the discretization the norming condition (4.18)

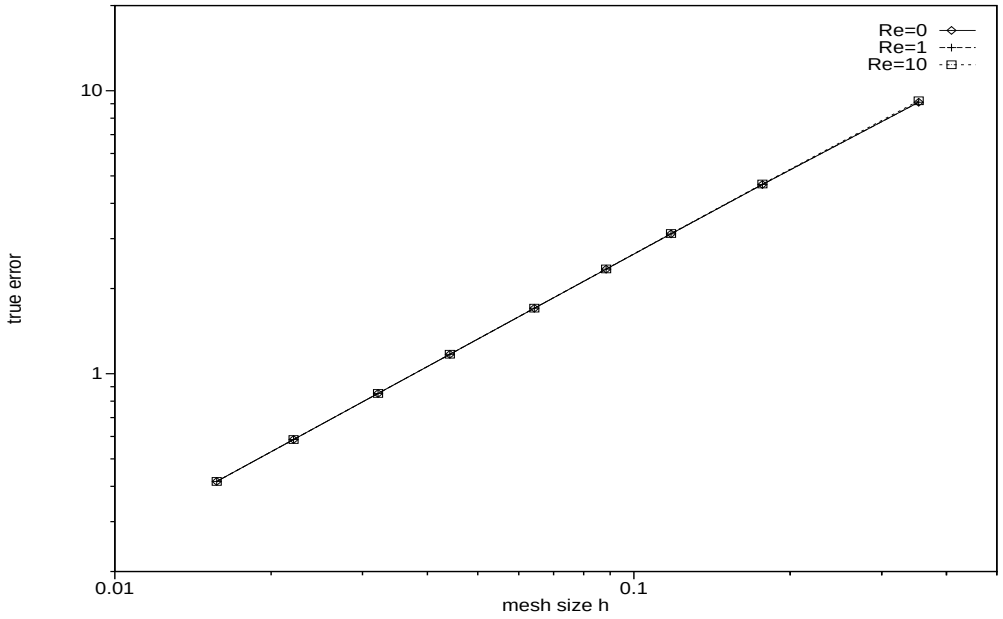

Figure 4.20: Example 4: The true error of the velocity in the $H^1(\Omega)^2$-norm for various Reynolds numbers.

for the pressure is replaced by a Dirichlet type condition at a single node in the interior of the domain. For the projecting error estimate the approximation space is expanded by Taylor-Hood elements (or $P_2 - P_1$-elements). They use quadratic polynomials for the velocity and linear polynomials for the pressure on the same element, see Cuvelier [1]. For smooth solutions (u, p) the approximation has the convergence order 1 and the approximation space for the error estimate produces a convergence order of at least order 2. Therefore the pair of approximations by macro–element approximations and by the extension with $P_2 - P_1$-elements is saturated for (u, p) with saturation bound 0 in the sense of Definition 2.

Figure 4.20 shows the dependence of the true error of the velocity components $u = (u_1, u_2)$ in the $H^1(\Omega)^2$-norm on the mesh size for various Reynolds numbers. The error of the pressure in the $H^0(\Omega)$-norm is shown by Figure 4.21. The true errors are independent of the Reynolds number (as the selected values for the Reynolds number are small). Reynolds numbers greater than 10 could not be tested as then the iterative methods in VECFEM did not converge. More interesting for the discussion is the behavior of the ratio of estimated and true errors for the velocity u which is shown in Figure 4.22 and for the pressure p which is shown in Figure 4.23. For the velocity components the $H^1(\Omega)^2$–norm and for the pressure the $H^1(\Omega)$–norm is used. Indepen-

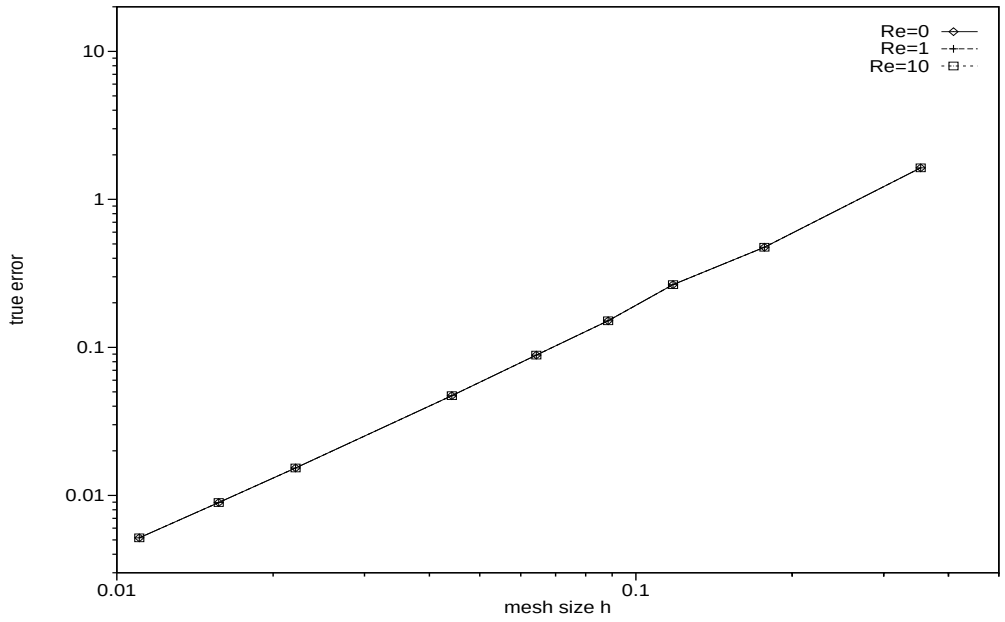

Figure 4.21: Example 4: The true error of the pressure in the $H^0(\Omega)$-norm for various Reynolds numbers.

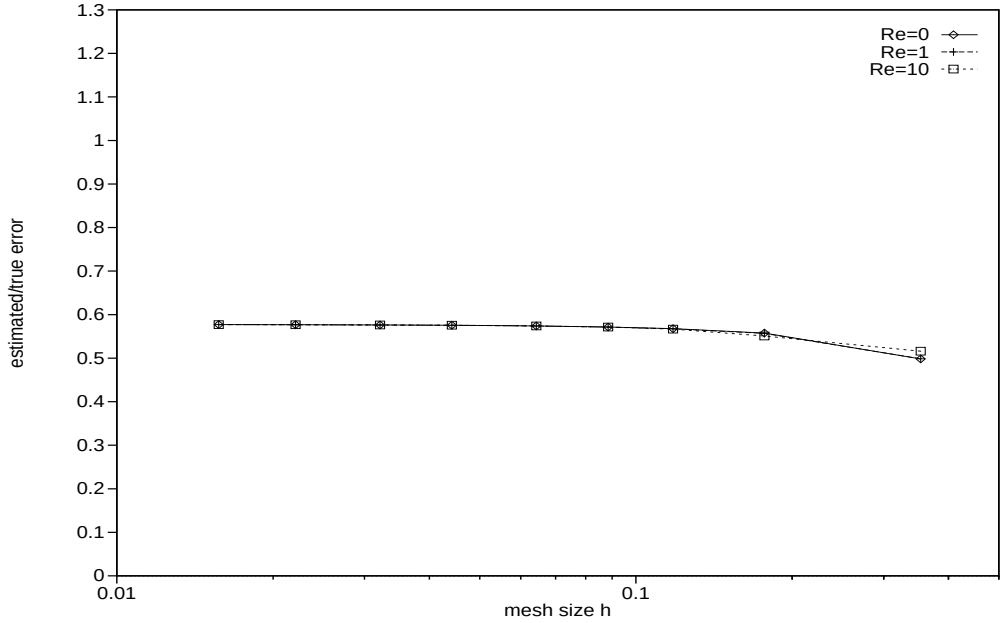

Figure 4.22: Example 4: The ratio of estimated and true error of the velocity in the $H^1(\Omega)^2$-norm for various Reynolds numbers.

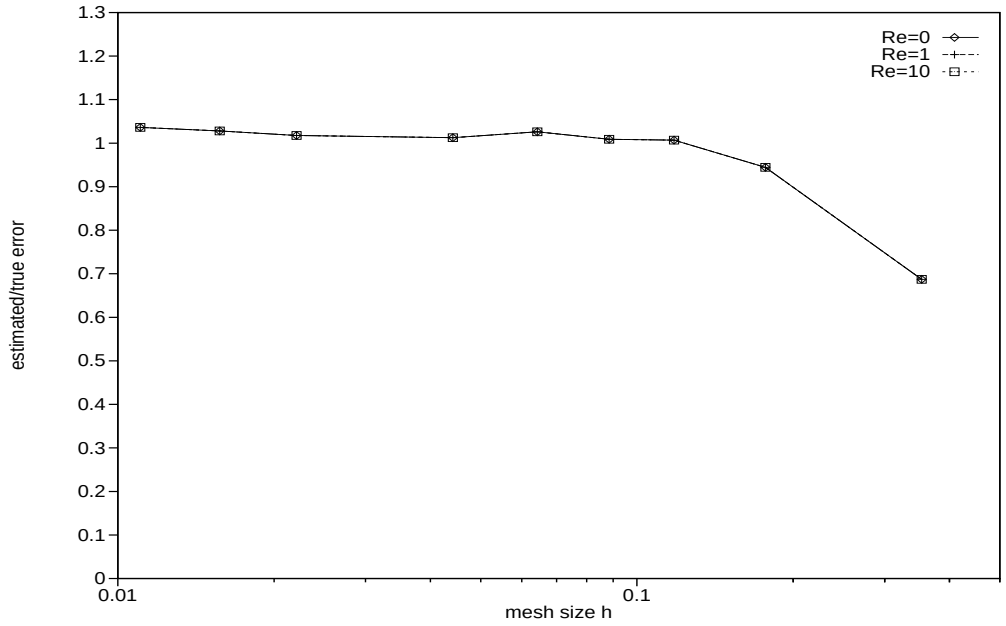

Figure 4.23: Example 4: The ratio of estimated and true error of the pressure in the $H^0(\Omega)$-norm for various Reynolds numbers.

dently of the Reynolds number the ratios for the velocity converge to a value in the order of 0.577 which is known from the previous two dimensional examples. The ratio of estimated and true error of the pressure converges to one. The reason for that is that another interpolation operator and another norm for the pressure like for the velocity components is used. However, this example shows that the projecting error estimate applied to saddle–point problems is also equivalent to the true error in the sense of Definition 3.

The optimal stopping criterion (2.117) is investigated. Figure 4.24 shows the ratio of the true error of the velocity u when using the optimal stopping criterion and a high accuracy $TOL = 10^{-8}$. Similar to Example 2 in Section 4.5 the ratio is in the order one. This demonstrates that the optimal stopping criterion works in an optimal manner. Figure 4.25 shows the ratio of the CPU time when calculating the solution with the optimal stopping criterion and the high accuracy on a Fujitsu VPP300. The saving of computing time is greater than 80%.

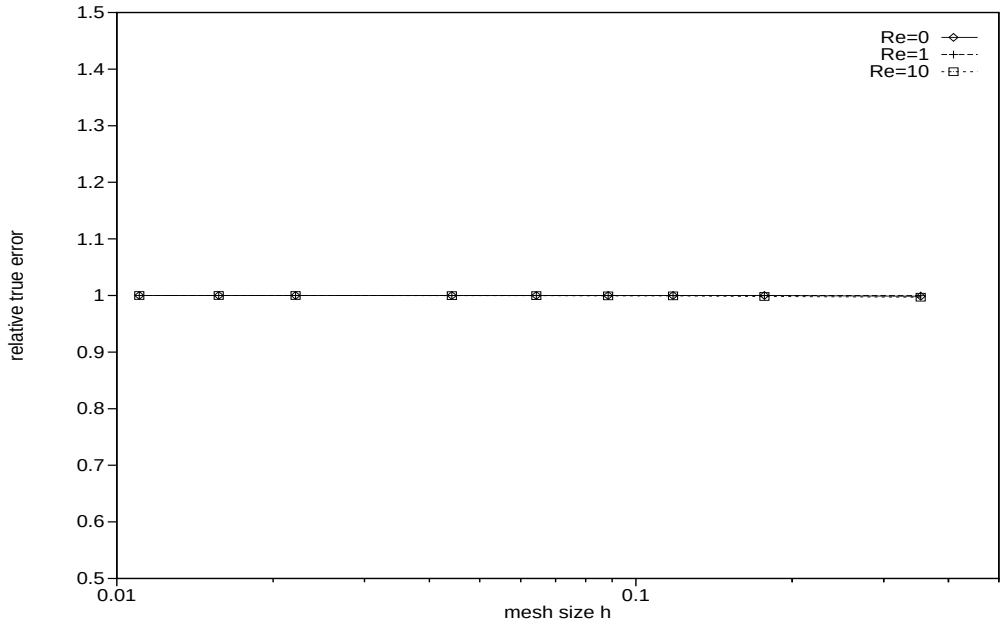

Figure 4.24: Example 4: True error of the velocity u in the $H^1(\Omega)^2$-norm by the optimal stopping criterion relative to the true error in the $H^1(\Omega)^2$-norm when using accuracy $TOL = 10^{-8}$.

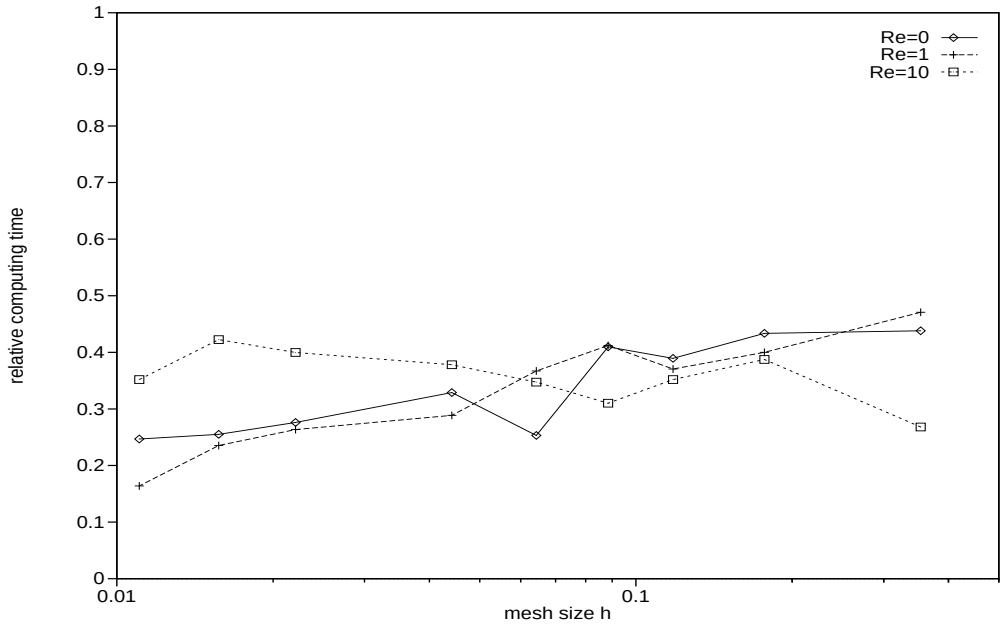

Figure 4.25: Example 4: Computing time when using the optimal stopping criterion relative to the computing time when using accuracy $TOL = 10^{-8}$.

Chapter 5

Conclusions

A general theory of a–posteriori error estimates for variational problems has been presented. The theory can be applied to non–linear well–posed variational problems. The basic idea is that a better solution approximation can be calculated by expanding the original approximation space V_h to a larger space $V_{h+} = V_h \oplus V_h^c$. The analysis includes the hierarchical error estimate which solves the error equation in V_h^c and the inflating error estimate which solves the error equation in the total expansion V_{h+}. In addition to these well–known error estimate techniques a new a–posteriori error estimate was derived from this general framework. It is called projecting error estimate since an approximation of the error is computed from the space V_h. Theorem 4 shows that all three error estimates are equivalent to the true error, i.e. they have exactly the same convergence order for decreasing mesh size like the exact error.

Moreover in Theorem 3 a stopping criterion for any iterative procedure to solve the non–linear discrete variational problem has been suggested. Balancing the discretization error and the stopping error ensures that the convergence order of the returned approximation towards the sought solution is optimal. The presented examples have shown that more than 60% of the arising computing costs can be saved by using this stopping criterion when the total error of the return approximation has to be minimal for the given FEM mesh and quadrature scheme. Under this condition the discrete variational problem can only be solved with a high accuracy as a–priori the size of the discretization error is unknown.

The projecting a–posteriori error estimate has been applied to the finite element method basing on the addition of higher order polynomials to the

original approximation space. Theorem 14 proves that the projecting error estimate for the FEM is equivalent to the true error in the sense of Definition 3. The proof has been given for a non–linear model problem but the results are valid for other elliptic variational problems and can be extend to saddle–point problems. The projecting error estimate considers the interpolation error as well as the error from numerical integration. Reusing the stiffness matrix which is available from the non–linear solver the calculation of the projecting error estimate is very cost–effective. It is always well–defined independently of the variational problem dealt with, no matter if it is linear or non–linear. This property in addition to the fact, that there are no specific conditions to the used FEM meshes, are the advantages of the projecting error estimate compared to other estimate techniques. As no deepened knowledge on the variational problem and no adapting are required the projecting error estimate is perfectly suitable for the application in black–box solver software for partial differential equations like VECFEM.

The presented two and three dimensional examples have confirmed the theoretical results for a FEM approximation by piecewise linear polynomials estimated by piecewise quadratic polynomials. The variety of the examples has shown that the projecting error estimate works for the analyized model problem as well as for other elliptic and saddle–point problems. It turned out that for most problems the true error is underestimated with a factor in the order of $0.577 \approx \frac{\sqrt{3}}{3}$ (in particular for two dimensional problems, for saddle–point problems the factor holds for the components approximated from $H^1(\Omega)$). As the theoretically obtained bounds for the ratio of estimated and true error (see inequality (3.165)) contain unknown constants it is not possible to verify this factor. Under- and overestimating of the true errors can be observed by other a–posteriori error estimate techniques, see Babuška [6, 8], Bank [16, 53], Duran [32], Rodriquez [54]. The extensive investigation of some known a–posteriori error estimates by Babuška [12] considering many problem parameters shows that underestimating by factor 0.577 is in the usual range of other error estimates.

The bounds for the ratio of estimated and true error for the projecting error estimate given in Theorem 3/Corollary 2 depend on the smoothness of the solution, the mesh quality and the condition number of the involved operators. But the examples have shown that only the smoothness of the solution influences the quality in a significant manner. However, for linear FEM approximations the results of the projecting error estimate based on piecewise quadratic polynomials and corrected by the factor 1.5 give a safe, reliable and robust estimate of the true error with a wide range of applications.

Bibliography

[1] C. Cuvelier; A-Segal and A.A. van Steenhoven. *Finite Element Methods and Navier-Stokes-Equations*. D. Reidel, Dordecht, 1986.

[2] D.A. Adams. *Sobolev Spaces*. Academic Press, New York, 1975.

[3] M. Ainsworth, J. Z. Zhu, A. W. Craig, and O. C. Zienkiewicz. Analysis of the Zienkiewicz-Zhu a–posterior error estimator in the finite element method. *Int. J. Numer. Methods Eng.*, 28:2161–2174, 1989.

[4] I. Babuška. The finite element method with Lagrangian multiplier. *Numer. Math.*, 20:179–192, 1972.

[5] I. Babuška. Courant element: Before and after. In M. Krizek, P. Neittaanmaeki, and R. Stenberg, editors, *Finite Element Methods: Fifity Years of the Courant Element*, pages 37–51. Marcel Dekker, New York, 1994.

[6] I. Babuška, R. Duran, and R. Rodriguez. Analysis of the efficiency of an posteriori error estimator for linear triangular finite elements. *SIAM J. Numer. Anal.*, 29(4):947–964, 1992.

[7] I. Babuška and A. Miller. Feedback finite element method with a–posteriori error estimation: Part I. the finite element method and some basic properties of the a–posteriori error estimator. *Comput. Methods Appl. Mech Engrg.*, 61:1–40, 1887.

[8] I. Babuška, L. Plank, and R. Rodriguez. Quality assessment of a–posteriori error estimators. *Fin. Elem. in Anal. Desgn.*, 11:285–306, 1992.

[9] I. Babuška and W. C. Reinboldt. A–posteriori error estimates for the finite element method. *Int. J. Numer. Methods Engrg.*, 12:1597–1615, 1978.

[10] I. Babuška and W. C. Reinboldt. Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.*, 15(4):736–754, 1978.

[11] I. Babuška and R. Rodriguez. The problem of the selection of an a–posteriori error estimator based on smoothening techniques. *Internat J. Numer. Methods Engrg.*, 36:539–567, 1993.

[12] I. Babuška, T. Strouboulis, and C.S. Upadhyay. A model study of the quality of a–posteriori error estimators for linear elliptic problems. Error estimation in the interior of patchwise uniform grids of triangles. *Comput. Methods in Appl. Mech. and Engrg.*, 114:307–378, 1994.

[13] I. Babuška, O.C. Zienkiewicz, J. Gago, and E. R. de A. Oliveira. *Accuracy Estimates and Adaptive Refinement in Finite Element Computations*. John Wiley, New York, 1986.

[14] R. E. Bank. Analysis of a local a–posteriori error estimate for elliptic equations. In I. Babuška, O.C. Zienkiewicz, J. Gago, and E. R. de A. Oliveira, editors, *Accuracy Estimates and Adaptive Refinement in Finite Element Computations*, New York, 1986. John Wiley.

[15] R. E. Bank and R. K. Smith. A–posteriori error estimates based on hierarchical bases. *SIAM J. Num. Anal.*, 30:921–935, 1993.

[16] R. E. Bank and A. Weiser. Some a–posteriori estimators for elliptic differential equations. *Math. Comp.*, 44:283–301, 1985.

[17] K.-J. Bathe. *Finite Element Procedures in Engineering Analysis*. Inc. Englewood Cliffs. Prentice Hall, New Jersey, 1982.

[18] F. Bornemann, B. Erdmann, and R. Korngruber. A–posteriori error estimates for elliptic problems in two and three space dimensions. Preprint SC 93-29, ZIB Berlin, 1993.

[19] J.H. Bramble and S.R. Hilbert. Estimation of linear functionals on Sobolev spaces with applications to Fourier transforms and spline interpolations. *SIAM J. Numer. Anal.*, 7:113–124, 1970.

[20] H. Brezis. *Operateurs Maximaux Monotones*. North–Holland, Amsterdam, 1973.

[21] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. Springer-Verlag, New York, 1991.

[22] B. W. Char, K. O. Geddes, G. H. Gonnet, B. L. Leong, M. B. Monagan, and S. M. Watt. *Maple V.* Waterloo Maple Software, Waterloo, 1992.

[23] T. J. Chung. *Finite Element Analysis in Fluid Dynamics.* McGraw-Hill, 1978.

[24] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems.* North Holland, Amsterdam, 1978.

[25] P. G. Ciarlet and J. L. Lions. *Handbook of Numerical Analysis*, volume 1. North Holland, Amsterdam, 1990.

[26] P. G. Ciarlet and P.-A. Raviart. The combined effect of curved boundaries and numerical integration in isoparametric finite element methods. In A.K. Aziz, editor, *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, pages 409–474. Academic Press, New York, 1972.

[27] P.G. Ciarlet and P.-A. Raviart. General Lagrange and Hermit interpolation in R^n with applications to the finite element method. *Arch. Rational Mesh. Engrg.*, 46:177–199, 1972.

[28] P. J. Davis and P. Rabinowitz. *Methods of Numerical Integration.* Academic, New York, 1975.

[29] D.J. Dawe. *Matrix and finite element displacement analysis of structures.* Clarendon Press, Oxford, 1984.

[30] L. Demkowicz, Ph. Devloo, and J.T. Oden. On a h-type mesh refinement strategy based on a minimization of interpolation error. *Comp. Methods Appl. Eng.*, 53:67–89, 1985.

[31] P. Deufelhard, P. Leinen, and H. Yserentant. Concepts of an adaptive hierarchical finite element code. Preprint SC 88–5, ZIB Berlin, 1988.

[32] R. Duran, M. Muschietti, and R. Rodriguez. On the asymptotic exactness of error estimators for linear triangle finite elements. *Numer. Math.*, 59:107–127, 1991.

[33] R. Duran, M. Muschietti, and R. Rodriguez. Asymptotically exact error estimators for rectangular finite elements. *SIAM J. Numeri. Anal.*, 29:78–88, 1992.

[34] V. Eijkhout and P. Vassilevski. The role of the strengthened Cauchy-Buniakowskii-Schwarz inequality in multilevel methods. *SIAM Review*, 33:405–419, 1991.

[35] G. Fichera. Existence theorems in elasticity. In S. Fluegge, editor, *Handbuch der Physik Band VIa/2*, pages 347–390. Springer Verlag, Berlin, 1972.

[36] S. Fluegge. *Handbuch der Physik Band VIa/2.* Springer Verlag, Berlin, 1972.

[37] V. Girault and P.-A. Raviart. *Finite Element Methods for the Navier-Stokes Equations.* Springer-Verlag, Berlin, 1986.

[38] L. Grosz, C. Roll, and W. Schönauer. VECFEM for mixed finite elements. Internal report 50/93, University of Karlsruhe, Computing Center, Postfach 6980, 76128 Karlsruhe, Germany, 1993.

[39] L. Grosz, C. Roll, and W. Schönauer. A black-box solver for the solution of general nonlinear functional equations by mixed FEM. In M. Křížek, P. Neittaanmäki, and R. Stenberg, editors, *Finite Element Methods, Fifty Years of the Courant Element*, pages 225–234. M. Dekker, 1994.

[40] L. Grosz and W. Schönauer. The nonlinear finite element solver VECFEM 3: The numerical algorithms. Internal report 65/96, University of Karlsruhe, Computing Center, Postfach 6980, 76128 Karlsruhe, Germany, 1996.

[41] A. Guessab. Cubature formulae which are exact on space P, intermediate between P_k and Q_k. *Numer. Math.*, 49:561–576, 1986.

[42] M. Gunzburger. *Finite Element Methods for viscous incompressible Flows.* Academic Press, Boston, 1989.

[43] H. Heuser. *Funktionalanalysis.* Teubner Verlag, Stuttgart, 1986.

[44] *I-DEAS, Solid Modeling, User's Guide.* SDRC, 2000 Eastman Drive, Milford, Ohio 45150, USA, 1990.

[45] C. Johnson. *Numerical Solutions of Partial Differential Equations by the Finite Element Method.* Cambridge University Press, Cambridge, New York, 1987.

[46] G. Kunert. Ein Residuenfehlerschätzer für anisotrope Tetraedernetze und Dreiecksnetze in der Finite–Elemente–Methode. Spc 95–10, TU Chemnitz–Zwickau, 1995.

[47] J.-L. Liu and W. C. Reinboldt. An a–posteriori error estimator for indefinite boundary value problems. Preprint, University of Pittsburgh, 1991.

[48] R.A. Nicolaides. On a class of finite elements generated by Lagrange interpolation. *SIAM J. Numer. Anal.*, 9:435–445, 1972.

[49] R.A. Nicolaides. On a class of finite elements generated by Lagrange interpolation II. *SIAM J. Numer. Anal.*, 10:182–189, 1972.

[50] *PDA Engineering: PATRAN II Release Notes Version 3.1.* Santa Ana, 1986.

[51] P.Grisvard. *Elliptic Problems in Non-Smooth Domains.* Pitman, Marshfield, Mass., 1985.

[52] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations.* Springer-Verlag, Berlin, Heidelberg, New York, 1994.

[53] B. D. Welfert R. E. Bank. A–posteriori estimators for the Stokes equations: a comparsion. *Comput. Methods in Appl. Mech. and Engrg.*, 28:591–623, 1991.

[54] R. Rodriquez. Some remarks on Zienkiewicz–Zhu estimator. *Numer. Meth. for PDE*, 10:625–635, 1994.

[55] W. Schönauer. *Scientific Computing on Vector Computers.* North-Holland, Amsterdam, New York, Oxford, Tokyo, 1987.

[56] W. Schönauer, K. Raith, and G. Glotz. The principle of the difference of difference quotients as a key to the selfadaptive solution of nonlinear partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 28:327–359, 1981.

[57] H. R. Schwarz. *Method of Finite Elements.* Academic Press, London, 1988.

[58] J. Stoer and R. Bulirsch. *Einführung in die numerische Mathematik I, II.* Springer-Verlag, Berlin, Heidelberg, New York, 1973.

[59] G. Strang. Approximation in the finite element method. *Numer. Math.*, 19:81–98, 1972.

[60] G. Strang and G. J. Fix. *An Analysis of the Finite Element Method.* Prentice-Hall, Englewood,NJ, 1973.

[61] R. Verfuerth. A simple error estimator for the Stokes equation. *Numer. Math.*, 55:309–325, 1989.

[62] R. Verfuerth. A–posteriori error estimates for nonlinear problems. *Math. Comput.*, 62(206):445–475, 1994.

[63] R. Verfuerth. The equivalence of a–posteriori error estimators. In W. Hackbusch and G. Wittum, editors, *Fast Solvers for Flow Problems*, pages 273–283, Wiesbaden, 1995. Vieweg.

[64] R. Verfuerth. *A Review of A–Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques.* Wiley-Teubner, New York, 1996.

[65] R. Weiss, H. Häfner, and W. Schönauer. LINSOL (LINear SOLver)—description and user's guide for the parallelized version. Internal report 61/95, University of Karlsruhe, Computing Center, Postfach 6980, 76128 Karlsruhe, Germany, 1995.

[66] H. Yserentant. On the multi-level splitting of finite element spaces. *Numer. Math.*, 49:379–412, 1986.

[67] J. Z. Zhu and O. C. Zienkiewicz. Superconvergence recovery technique and a–posteriori error estimators. *Int. J. Numer. Methods Eng.*, 30:1321–1339, 1990.

[68] O. C. Zienkiewicz. *The Finite Element Method in Engineering Science.* McGraw-Hill, London, second edition, 1971.

[69] O. C. Zienkiewicz and A. Craig. Adaptive refinement, error estimates, multigrid solution, and hierarchic finite element method concepts. In I. Babuška, O.C. Zienkiewicz, J. Gago, and E. R. de A. Oliveira, editors, *Accuracy Estimates and Adaptive Refinement in Finite Element Computations*, New York, 1986. John Wiley.

[70] O. C. Zienkiewicz and J. Z. Zhu. A simple error estimator and adaptive procedure for practical engeneering analysis. *Int. J. Numer. Methods Eng.*, 24:337–357, 1987.

[71] O. C. Zienkiewicz and J. Z. Zhu. The three R's of engineering analysis and error estimation and adaptivity. *Comp. Methods App. Mech. Eng.*, 82:95–113, 1990.

[72] O.C. Zienkiewicz and J. Z. Zhu. The superconvergent patch recovery and a–posteriori error estimates, part I. *Int. J. Numer. Methods Eng.*, 33:1331–1364, 1992.

[73] O.C. Zienkiewicz and J. Z. Zhu. The superconvergent patch recovery and a–posteriori error estimates, part II. *Int. J. Numer. Methods Eng.*, 33:1365–1382, 1992.

[74] M. Zlamal. On the finite element method. *Numer. Math.*, 12:394–409, 1968.

Appendix A

List of Notations

Sorted by appearance in the thesis.

$S(x, \delta)$ (ball with radius δ and centre x) ..... (3.4), page 46

$cl(K)$, $int(K)$, ∂K (closure, interior, boundary of K) ..... page 46

h_K (radius of smallest ball in K) ..... (3.5), page 46

ρ_K (radius of biggest ball containing K) ..... (3.6), page 46

Ψ (affine transformation) ..... (3.9), page 47

α, $|\alpha|$ (multi–index) ..... (3.11), page 48

$D^\alpha v$ (α-th partial derivative of v) ..... (3.12), page 48

$W^{m,q}(\Omega)$ (integrable q-th power of derivatives up to order m) (3.13), page 48

$|.|_{m,q,\Omega}$ (semi–norm in $W^{m,q}(\Omega)$) ..... (3.15), page 48

$\|.\|_{m,q,\Omega}$ (norm in $W^{m,q}(\Omega)$) ..... (3.16), page 48

$H^m(\Omega)$ (Hilbert space $W^{m,2}(\Omega)$) ..... (3.16), page 48

$|.|_{m,\Omega} := |.|_{m,2,\Omega}$ (semi–norm in $H^m(\Omega)$) ..... (3.16), page 48

$\|.\|_{m,\Omega} := \|.\|_{m,2,\Omega}$ (norm in $H^m(\Omega)$) ..... (3.17), page 48

$C^m(\Omega)$ (m–times continuously differentiable) ..... page 49

$W^{m,q}(\mathcal{T})^d$ (product space of $W^{m,q}(\Omega)$) ..... (3.24), page 50

T^0 (n-simplex) ..... (3.28), page 51

P_k (polynomials of order k) ..... (3.31), page 51

$X^{0,k} := \{x_i^{0,k}\}_{i=1,d_k}$ (local degrees of freedom) ..... (3.32), page 52

$\mathcal{I}^k$ (local interpolation operator) ..... (3.33), page 52

Q^l (quadrature scheme on T^0 being exact of order l) ..... (3.35), page 53

E^l (linear functional on $C^0(T^0)$ vanishing on P_l) ..... (3.37), page 54

$u_;$ (vector of u and its first spatial derivatives) ..... (3.40), page 55

Appendix B

Lists of Definitions, Theorems and Figures

List of Definitions

List of Theorems

List of Corollaries

List of Lemmata

List of Figures