

# A posteriori error estimation for anisotropic tetrahedral and triangular finite element meshes

Von der Fakultät für Mathematik der Technischen Universität Chemnitz  
genehmigte

**D i s s e r t a t i o n**

zur Erlangung des akademischen Grades

Doctor rerum naturalium

(Dr. rer. nat.)

vorgelegt

von Dipl.-Math. Gerd Kunert  
geboren am 19. Februar 1970 in Freiberg

eingereicht am 23. Juli 1998

Gutachter: Prof. Dr. Bernd Heinrich  
Prof. Dr. Rüdiger Verfürth  
Prof. Dr. Dietmar Kröner

Tag der Verteidigung: 8. Januar 1999

## Preface

Many physical problems lead to boundary value problems for partial differential equations, which can be solved with the finite element method. In order to construct adaptive solution algorithms or to measure the error one aims at reliable *a posteriori* error estimators. Many such estimators are known, as well as their theoretical foundation.

Some boundary value problems yield so-called *anisotropic* solutions (e.g. with boundary layers). Then anisotropic finite element meshes can be advantageous. However, the common error estimators for isotropic meshes fail when applied to anisotropic meshes, or they were not investigated yet.

For *rectangular* or *cuboidal* anisotropic meshes a modified error estimator had already been derived. In this paper error estimators for anisotropic *tetrahedral* or *triangular* meshes are considered. Such meshes offer a greater geometrical flexibility.

For the Poisson equation we introduce a residual error estimator, an estimator based on a local problem, several Zienkiewicz-Zhu estimators, and an  $L_2$  error estimator, respectively. A corresponding mathematical theory is given. For a singularly perturbed reaction-diffusion equation a residual error estimator is derived as well. The numerical examples demonstrate that reliable and efficient error estimation is possible on anisotropic meshes.

The analysis basically relies on two important tools, namely anisotropic interpolation error estimates and the so-called bubble functions. Moreover, the correspondence of an anisotropic mesh with an anisotropic solution plays a vital role.

**Keywords:** finite elements, anisotropic mesh, tetrahedral mesh, triangular mesh, Poisson equation, singularly perturbed reaction-diffusion equation, anisotropic *a posteriori* error estimate, anisotropic interpolation error estimate

**AMS(MOS):** 65N30, 65N15, 35B25

# Contents

List of Symbols . . . . .	5
<b>1 Introduction</b> . . . . .	<b>7</b>
1.1 Introduction to anisotropic finite elements . . . . .	7
1.2 Aim of this work and outline . . . . .	11
<b>2 Preliminaries</b> . . . . .	<b>15</b>
2.1 Notation and basic properties . . . . .	15
2.1.1 Notation of the tetrahedron . . . . .	15
2.1.2 Transformations . . . . .	17
2.1.3 The directional derivative $\tilde{D}_i$ . . . . .	19
2.1.4 Notation of the summation . . . . .	20
2.1.5 Auxiliary subdomains . . . . .	21
2.2 Requirements on the mesh . . . . .	21
2.3 Basic tools . . . . .	22
2.3.1 Anisotropic trace inequalities . . . . .	22
2.3.2 Inverse inequalities for finite element functions . . . . .	24
2.3.3 Bubble functions . . . . .	25
<b>3 The Poisson equation</b> . . . . .	<b>29</b>
3.1 Analytical Background . . . . .	29
3.2 Residual error estimator . . . . .	32
3.2.1 Relation between anisotropic mesh and anisotropic function . . . . .	32
3.2.2 Anisotropic interpolation estimates . . . . .	35
3.2.3 Anisotropic residual error estimator . . . . .	45
3.2.4 Discussion of the matching function . . . . .	51
3.3 A local Dirichlet problem error estimator . . . . .	53
3.3.1 Introduction and definition . . . . .	53
3.3.2 Preliminary result . . . . .	56
3.3.3 Equivalence and bounds of the estimator . . . . .	61
3.3.4 Condition number of the finite element matrix . . . . .	65
3.3.5 Comments and Remarks . . . . .	67
3.4 Zienkiewicz-Zhu like error estimators . . . . .	69
3.4.1 Introduction . . . . .	69
3.4.2 Cuboidal or rectangular mesh . . . . .	69
3.4.3 Tetrahedral or triangular meshes . . . . .	70
3.4.4 Conclusions . . . . .	72
3.5 $L_2$ error estimator . . . . .	73
3.5.1 Special $L_2$ bubble functions and their inverse inequalities . . . . .	73
3.5.2 The matching function for $L_2$ error estimation . . . . .	77

3.5.3	Anisotropic interpolation estimates . . . . .	78
3.5.4	Anisotropic $L_2$ error estimator . . . . .	80
<b>4</b>	<b>A reaction-diffusion equation</b>	<b>83</b>
4.1	Analytical Background . . . . .	83
4.2	Residual error estimator . . . . .	85
4.2.1	Special face bubble functions . . . . .	85
4.2.2	Anisotropic interpolation estimates . . . . .	86
4.2.3	Anisotropic residual error estimator . . . . .	87
<b>5</b>	<b>Numerical examples</b>	<b>91</b>
5.1	Scope of and introduction to the numerical experiments . . . . .	91
5.1.1	General Remarks . . . . .	91
5.1.2	The matching functions $m_1$ and $m_2$ . . . . .	92
5.1.3	Interpolation error estimates . . . . .	93
5.1.4	Finite element error estimates . . . . .	93
5.1.5	Computational effort . . . . .	95
5.1.6	Aim of the experiments . . . . .	95
5.2	Numerical examples . . . . .	97
5.2.1	Example 1 (2D; anisotropic solution and adapted mesh) . . . . .	97
5.2.2	Example 2 (2D; anisotropic solution and misadapted mesh) . . . . .	105
5.2.3	Example 3 (2D; isotropic solution and anisotropic mesh) . . . . .	109
5.2.4	Example 4 (3D; anisotropic solution and adapted mesh) . . . . .	111
5.3	Conclusions . . . . .	117
	<b>Summary</b>	<b>119</b>
	<b>Bibliography</b>	<b>121</b>

# List of symbols

The list below comprises important notation accompanied by a brief explanation and (where possible) the page number of its definition or first occurrence. The notation and all proofs are given for the three-dimensional case ( $d = 3$ ). The treatment of the easier two-dimensional case ( $d = 2$ ) is straight-forward. Note that then some meanings are different (e.g.  $T$  being a triangle instead of a tetrahedron).

Symbol		page
$\Omega \subset \mathbb{R}^d$	$d$ -dimensional computational domain, $d = 2, 3$	
$C^k(\omega)$	space of $k$ times continuously differentiable functions	
$L_2(\omega)$	space of square integrable functions	
$H^k(\omega)$	Sobolev space of functions whose $k^{\text{th}}$ derivative is in $L_2(\omega)$	
$H_o^k(\omega)$	Sobolev space of functions of $H^k(\omega)$ satisfying the corresponding homogeneous Dirichlet boundary conditions	
$L_\infty(\omega), \ \cdot\ _{\infty,\omega}$	space of essentially bounded functions with norm	
$\mathbb{P}^k(\omega)$	space of polynomials of order $k$ or less.	
$(\cdot, \cdot), (\cdot, \cdot)_\omega$	$L_2(\Omega)$ scalar product or $L_2(\omega)$ scalar product	
$\ \cdot\ $	$L_2$ norm over $\Omega$	
$\ \cdot\ _\omega, \ \cdot\ _E$	$L_2$ norm over a domain $\omega$ or a face $E$	
$\ \cdot\ _{\mathbb{R}^d}$	Euclidean norm of a vector	
$\ \cdot\ _{\mathbb{R}^{d \times d}}$	spectral norm of a matrix	
$\mathbf{e}_i$	unitary vectors of $\mathbb{R}^d$	
$\Gamma = \partial\Omega$	boundary of $\Omega$	
$\Gamma_D, \Gamma_N, \Gamma_R$	Dirichlet, Neumann and Robin boundary part of $\partial\Omega$	29
$\alpha$	parameter of the Robin boundary conditions	29
$\mathcal{T}_h$	triangulation of $\Omega$	21
$V_h, V_{o,h}$	finite element spaces over $\mathcal{T}_h$	30
$T$	tetrahedron (or triangle) of the triangulation $\mathcal{T}_h$	
$ T $	volume $\text{meas}_d(T)$ of $T$	
$\mathbf{p}_i$	special vectors of $T$ , $i = 1 \dots d$	16
$h_i = h_{i,T}$	length $ \mathbf{p}_i $ of vector $\mathbf{p}_i$	16
$h_{\min,T}$	$:= \min\{h_{i,T}\} = h_{d,T}$	16
$h_i(\mathbf{x}), h_{\min}(\mathbf{x})$	global function that has value $h_{i,T}$ or $h_{\min,T}$ over $T$	16
$E, E_T$	arbitrary face of $T$ (or edge of a triangle $T$ )	16
$ E $	$\text{meas}_{d-1}(E)$	16
$h_{E,T}$	length of the height over $E$ in a tetrahedron $T$	16
$h_E$	$(h_{E,T_1} + h_{E,T_2})/2$ for $E = T_1 \cap T_2$	22
$h_{\min,E}$	$(h_{\min,T_1} + h_{\min,T_2})/2$ for $E = T_1 \cap T_2$	22
$\bar{T}$	standard tetrahedron (see definition)	17
$\hat{T}$	reference tetrahedron	17
$F_A$	affine linear mapping from standard tetrahedron $\bar{T}$ onto $T$	17
$F_C$	affine linear mapping from reference tetrahedron $\hat{T}$ onto $T$	17
$A_T, C_T \in \mathbb{R}^{d \times d}$	transformation matrices of the maps $F_A$ and $F_C$	17
$C(\mathbf{x})$	global matrix function that coincides with $C_T$ over $T$	17
$H_T \in \mathbb{R}^{d \times d}$	$\text{diag}(h_1, h_2, h_3)$	17

Symbol		page
$\tilde{D}_i$	(unitary) directional derivative in the direction $\mathbf{p}_i$	19
$a, a_j$	node (or vertex) of $\mathcal{T}_h$	20
$\mathcal{N}, \mathcal{N}_I, \mathcal{N}_T, \dots$	sets of nodes	20
$\omega_T, \omega_E$	auxiliary local subdomains	21
$M_j$	macro element of the node $a_j$	35
$\lambda_{T,i}$	barycentric coordinates of $T$	25
$b_T$	element bubble function (related to $T$ )	25
	special $L_2$ element bubble function	73
$b_E$	face bubble function (related to $E$ )	25
	special $L_2$ face bubble function	74
$F_{ext}$	extension operator $F_{ext} : \mathbb{P}^0(E) \rightarrow \mathbb{P}^0(T)$	27
$a(\cdot, \cdot)$	bilinear form	29, 83
$\ \cdot\ $	energy norm $\ \cdot\ ^2 = a(\cdot, \cdot)$ , problem dependent	30, 83
$\varphi_j$	linear basis function of node $a_j$ ,	38
	$\varphi_j(a_i) = \delta_i^j$ (Kronecker symbol)	
$R_o$	Clément interpolation operator $H_o^1(\Omega) \mapsto V_{o,h}$	41
$m_1(v, \mathcal{T}_h)$	matching function for $v \in H^1(\Omega)$	33
$m_2(v, \mathcal{T}_h)$	matching function for $v \in H^2(\Omega)$	78
$\nabla^R(u_h)$	recovered (averaged) gradient	52
$m_1^R(u - u_h, \mathcal{T}_h)$	approximation of $m_1(u - u_h, \mathcal{T}_h)$	52
$P_1, P_2, P_3$	mappings into finite dimensional spaces	45
$r_T(v_h)$	element residual (problem dependent)	45, 87
$r_E(v_h)$	face residual defined individually for interior faces (gradient jump), Neumann or Robin boundary faces, resp.	45
$n_E$	any of the two unitary normal vectors of $E$	45
$\partial/\partial n$	directional derivative with respect to the outer normal unit vector $n$	47
$\eta_{R,T}, \eta_R, \eta_{R,\omega_T}$	residual error estimator (Poisson equation, energy norm) for $T, \Omega$ and $\omega_T$ ,	46, 61
$\zeta_T, \zeta$	local and global approximation term	46
$V_T$	ansatz and test space of the local problem	55
$\eta_{D,T}, \eta_D, \eta_{D,\omega_T}$	local Dirichlet problem error estimator (energy norm) for $T, \Omega$ and $\omega_T$	55, 61
$\eta_{Z,T}, \eta_{Z_{1/2/3}}$	Zienkiewicz-Zhu like error estimators	70, 71, 72
$D^2 v$	matrix of the second partial derivatives of $v$	73
$\eta_{R,L_2,T}, \eta_{R,L_2}$	$L_2$ residual error estimator for $T$ and $\Omega$	80
$\zeta_{L_2,T}, \zeta_{L_2}$	local and global $L_2$ approximation term	80
$\varepsilon$	parameter of the singularly perturbed equation	83
$\alpha_T$	$\min\{1, \varepsilon^{-1/2} \cdot h_{min,T}\}$	86
$\eta_{\varepsilon,R,T}$	residual error estimator (energy norm)	88

# Chapter 1

## Introduction

### 1.1 Introduction to error estimation and anisotropic finite elements

Computer simulation plays nowadays an important role in many industries and fields like automobile and aircraft construction, semiconductor device simulation, weather and pollution forecasts. In many cases the modeling leads to boundary value problems for partial differential equations. When dealing with a real-world problem, its exact solution  $u$  can virtually never be found analytically. Thus one tries to obtain an approximate solution  $u_h$  by discretizing the original problem. Popular discretization methods include the finite difference method, the finite element method, the finite volume method, and the boundary element method. Our work is devoted exclusively to the *finite element method (FEM)*. The literature about this method is vast, and we cite the standard textbook by Ciarlet [26] as well as [27] for basic consideration and some overview.

Variational analysis of a second-order boundary value problem leads to the *variational (or weak) formulation* whose solution is denoted by  $u$ . Then the finite element method can be employed to solve this problem numerically. The *finite element formulation* is obtained by applying the Galerkin procedure to the weak formulation. This discretization employs a family  $\mathcal{F} = \{\mathcal{T}_h\}$  of meshes (also referred to as triangulations or grids) in conjunction with so-called ansatz and test functions. The approximate (or finite element) solution is denoted by  $u_h$ . For simplicity we restrict ourselves to bounded, polyhedral, three-dimensional or two-dimensional domains  $\Omega \subset \mathbb{R}^d$ , i.e.  $d = 3$  or  $2$ .

Our particular interest focuses on the approximation error  $u - u_h$  of the discretization method, and on how this error can be measured in some suitable norm. Since the exact but unknown solution  $u$  is involved, the error can not be measured exactly but only estimated by lower and upper bounds. Such error estimation can be distinguished in *a priori* and *a posteriori* error estimation.

*A priori* error estimates use the exact solution  $u$  and/or input data. They allow the prediction of the decrease of the error norm as the discretization becomes better (for example, when a discretization parameter  $h \rightarrow 0$ ). Introductory details about this so-called asymptotic behaviour can be found in [26]. The knowledge of  $u_h$  is not necessary. In turn, the size of the error norm can not be estimated for a single, actual mesh (i.e. discretization).

In contrast to this, *a posteriori* error estimates require and employ the approximate solution  $u_h$ . By means of  $u_h$  and the given data one aims to estimate  $\|u - u_h\|_*$  for the current mesh, but predictions of the asymptotic behaviour of the error are not possible.

When investigating real-world problems one is primarily interested in an accurate solution computed with as little effort as possible. Hence *a priori* error estimates are less suited, and *a posteriori* error estimators are required. The latter ones basically serve two purposes. Firstly, when an approximate solution  $u_h$  is computed, one usually wants to know how accurate  $u_h$  is. An estimation of the error norm  $\|u - u_h\|_*$  can answer this question, provided the error norm is chosen suitably. The importance of this purpose is readily understood when considering failure predictions in structural analysis, for example.

Secondly, *a posteriori* error estimators are indispensable ingredients of adaptive solution algorithms. Such algorithms commonly consist of the following steps (cf. [58] for example).

0. Start with an initial mesh  $\mathcal{T}_0$ .
1. Solve the corresponding discrete system.
2. Compute the local *a posteriori* error estimator for each element  $T$  of the mesh.
3. When the estimated global error is small enough then stop.  
Otherwise refine the mesh (based on the local error estimate), and re-iterate.

Our work focuses solely on *a posteriori* error estimation. Since the pioneering work by Babuška and Rheinboldt [13, 14] many more error estimators have been designed, and the literature on this topic is vast. We resist to give an exhaustive survey but present the following key references instead. For further reading we recommend the overview work of Verfürth [58], for example, and the literature cited therein. Before going into detail, recall the abstract framework of error estimation (see e.g. [58]). It is well-known that the error norm (here in the space  $H_o^1(\Omega)$ ) is, up to multiplicative constants, bounded from above and below by the norm of the residual in the dual space (here  $[H_o^1(\Omega)]^* = H^{-1}(\Omega)$ ). Since this dual norm is difficult to evaluate, most *a posteriori* error estimators try to approximate it by quantities which can be more easily computed from the given data.

The history of *a posteriori* error estimation starts with Babuška and Rheinboldt who presented a so-called ‘residual error estimator’ [13] and a ‘local problem error estimator’ [14]. Bank and Weiser [16] derived different local problem error estimators. Zienkiewicz and his coworkers proposed error estimators based on a recovered gradient [44, 66] (nowadays often called ‘Zienkiewicz-Zhu error estimators’). Later an improved estimator has been designed by means of the ‘superconvergent patch recovery’ [67]. Johnson, Eriksson et al. contributed to the methodology of error estimation (cf. [29] for example), and to error estimation for various differential equations. Verfürth [58] derived local lower bounds of the error, unified the theory of several error estimators, and considered other aspects of error estimation (see e.g. [23, 59]). Further and quite general ideas to derive error estimators will now be briefly discussed.

Firstly, over the last years the so-called *complementary energy principle* has been increasingly investigated and utilized. Provided that certain approximation effects are neglected, it yields an upper error bound without unknown constants. This can be viewed as an advantage over some other error estimators. The roots of the complementary energy principle go back to Synge [55] and Aubin and Burchard [12]. More recently the principle has been applied by [1, 2, 15, 33, 37, 62], to name but a few.

Secondly, over the past years attention has been drawn towards *superconvergence*. This property has been exploited to derive error estimators, see e.g. [36, 67]. Additionally it is possible to analyse model problems on highly regular meshes. It should be stressed that superconvergence will only occur in special circumstances. Roughly speaking, it requires

meshes which are sufficiently ‘regular’ (in some sense). Nowadays these assumptions are increasingly understood. We refer to Wahlbin [61], the proceedings [36] and the references cited therein.

It seems worth mentioning that Becker and Rannacher [17, 18] pursue a novel, interesting approach. They do not consider primarily an error *norm* but some *functional* of the error. Estimates are obtained via a dual problem. It remains to be seen whether this method can gain popularity.

Error estimation for elliptic problems is today fairly well understood for standard discretizations. Up to now, the field of research in error estimation has become widespread. Some important topics include estimators for parabolic and hyperbolic problems, the investigation of general nonlinear problems (e.g. the Navier-Stokes equation), the incorporation of error estimation in adaptive algorithms and the analysis thereof, and the extension of error estimation to modern varieties of the finite element method like the *hp*-version FEM or anisotropic FEM. A comprehensive overview of a particular field of research can be obtained via electronic databases.

When the quality of error estimation is to be assessed, one often encounters the terms *reliable*, *efficient* and *asymptotically exact*. To explain these terms, define the so-called *effectivity index* to be the ratio of the estimated error and the true error (in some norm). Primarily one is interested in estimators that reliably bound the error, i.e. the error is guaranteed to be smaller than some estimated value. Such estimators are called *reliable*. In other words, the error is bounded from above, and the effectivity index is bounded from below. Secondly, an estimator is said to be *efficient* if it bounds the error from below. This corresponds to an effectivity index bounded from above. Efficient local estimators are desirable in order to reduce the global error by refining elements with large local error contributions. In our exposition we will refer to the lower and upper bound of the error, respectively.

Lastly, an estimator is said to be *asymptotically exact* if the effectivity index tends to one as the discretization becomes finer. According to [1, 15, 58], asymptotic exactness is a very fragile property that depends on or requires, amongst other things, very regular meshes and a fairly smooth solution. It can not be expected to hold for general types of meshes used in practical engineering computations.

The topic of our work is a special class of problems which can be solved very efficiently by a non-classical finite element method. Some boundary value problems yield a solution which exhibits little variation in one direction but much change in an other direction. Such solutions are called *anisotropic*. Examples include functions which are almost constant or linear in one direction, and/or which have a boundary layer or an interior layer. An equivalent description is that an anisotropic function shows an almost one-dimensional (or lower-dimensional) behaviour. By this we mean that the function varies significantly only perpendicularly to a certain manifold. Typical problems with anisotropic solutions include the reaction-diffusion-convection equation (see e.g. [8, 32, 38, 64]), the Poisson equation in a three-dimensional domain with an edge of an interior angle larger than  $\pi$  (see e.g. [10] and the numerical example 4 of chapter 5), and other problems arising from fluid dynamics or weather simulation, for example. Functions which are not anisotropic are called *isotropic* – clearly, this distinction is not a strict mathematical partitioning of the set of functions but rather a matter of degree.

One feature of the classical finite element method is that the ratio of the diameters of the circumscribed and inscribed spheres of a finite element (e.g. rectangle, tetrahedron, or cube) is bounded. Such meshes are referred to as *isotropic meshes*. But when an

anisotropic solution as mentioned above occurs it is sensible to violate this condition and to employ highly stretched elements instead. From a heuristic point of view, as well as from anisotropic interpolation analysis (e.g. [6]), it is natural to use a small mesh size in the direction of the rapid variation of the solution, and a larger mesh size in the direction of little variation (i.e. along the manifold of the anisotropy). We also say that the mesh is then anisotropically aligned with the solution, and we refer to it as *anisotropic mesh*. In this way one hopes to capture the important features of the solution with much less elements than when using an isotropic mesh. Numerical evidence confirms that problems with anisotropic solutions can indeed be solved more efficiently on anisotropic meshes (i.e. with less degrees of freedom, less computational effort, or less memory to achieve the same accuracy). If the anisotropy of a solution occurs along a curved manifold then the anisotropic mesh (i.e. a stretched element) has to follow that curved manifold.

At first anisotropic finite elements have been applied by engineers, but soon scientists were to investigate and analyse these methods, see e.g. Beinert and Kröner [19], Fröhlich, Lang and Roitzsch [30], Kornhuber and Roitzsch [34], Nochetto [41], Peraire et al. [42], Rachowicz [43], Rick, Greza and Koschel [45], Siebert [50], Vilsmeier, Hänel et al. [60], Zienkiewicz and Wu [65]. Stein and Ohnibus [54] consider error estimators for elasticity problems. Their estimators which are also applied to anisotropic meshes are mainly based on an engineering point of view, but the mathematical theory is incomplete.

Anisotropic finite element methods with emphasis on *a priori* error estimation have been considered for example by Apel and Dobrowolski [7], Apel and Lube [8], Apel and Nicaise [10], Miller, O’Riordan and Shishkin [39], Roos [47], Zhou and Rannacher [64]. But although anisotropic finite elements are used, their theoretical foundation is weaker than for isotropic elements.

Most of the aforementioned papers show that anisotropic meshes are not applied out of the blue. Usually such meshes are embedded in an adaptive anisotropic solution strategy. For example, one could start with an isotropic mesh and refine it gradually until the desired anisotropic mesh is obtained. Further possibilities include an improvement or an adaption of an existing anisotropic mesh (e.g. for time dependent problems), or the generation of a new mesh. Summarizing, an adaptive anisotropic strategy involves basically the following steps.

0. Start with an initial mesh  $\mathcal{T}_0$ .
1. Solve the corresponding discrete system.
2. Compute the local *a posteriori* error estimator for each element  $T$  of the mesh.
3. When the estimated global error is small enough then stop.  
Otherwise obtain information for a new, better mesh. This includes:
  - Detect regions of anisotropic behaviour of the solution.
  - Determine a (quasi) optimal aspect ratio and stretching direction of the finite elements.
  - Determine the element size.
4. Based on this information, construct a new mesh or perform a mesh refinement, and re-iterate.

Typical applications of this strategy (to mostly real-world problems) can be studied in the work of [42, 45, 60, 65], to name but a few. In what follows we will discuss each step of the aforementioned procedure and the difficulties that evolve from them.

In step 1 the finite element system is to be solved. Over the last years it turned out that the performance of modern solvers (like multigrid and multilevel methods) can heavily deteriorate on anisotropic meshes. First improvements are achieved by utilizing results of similar problems (see e.g. [31]). Current research is increasingly devoted to the construction of robust solvers and preconditioners for anisotropic meshes, and some numerical results are already available. Here we do not consider such questions.

The estimation of the error (step 2) is more critical. Consider, for example, the papers [42, 43, 45, 60] which all propose an adaptive strategy (mostly in the context of flow simulation). It becomes apparent that every step but the error estimation is addressed and answered. In practice it is common to omit the *error estimation*, and to employ only certain *mesh refinement criteria* which are derived by heuristic considerations. The only explicit and analytically founded *a posteriori* error estimators for anisotropic meshes are, to our knowledge, due to [35, 50].

Similarly, information about the anisotropic solution is often drawn from heuristic arguments (step 3). This includes the analysis of the second order partial derivatives [25, 42, 45, 51, 65], of the level lines [34] or of the gradient (or gradient jump) of some values [19, 43, 50].

The last step, namely the remeshing and/or the anisotropic mesh generation, has become a field of intensified research, and it is increasingly understood. It is done by mesh refinement and adjustment (see, e.g. [19, 21, 30, 34, 43, 45, 52]), or by a new mesh generation ([20, 24, 25, 42, 45, 60]), or by a mixture of both. The generation of a new mesh is, for example, often based on an advancing front algorithm, on Delaunay triangulation, or on inserting or deleting of nodes (and edges and triangles). Mesh improvement operators (like edge swapping or vertex relocation) are often applied to enhance the quality of meshes. In order to obtain anisotropic elements, the use of an anisotropic local metric (also referred to as virtual transformation) is quite popular. For example, such a metric can be derived from the Hessian of the function under investigation (see e.g. [20, 25, 42, 45, 51]). A background mesh can be advantageous to provide some data such as the metric. A certain overview of several of the aforementioned methods as well as a discussion of their advantages and drawbacks is presented in [20]. In [51], some basic aspects of optimal anisotropic meshes (with respect to the error) are investigated. A review of different papers in the field of computational fluid dynamics with anisotropic meshes is given as well. Finally, it is easily comprehensible that the mesh generation step requires much programming effort and a sophisticated data structure which also has to suit the error estimation process.

## 1.2 Aim of this work and outline

Our work focuses on *a posteriori* error estimation for anisotropic meshes, and attention is paid in particular to the mathematical theory. As we have seen, the analytical foundation of anisotropic estimators is at the very beginning. To our knowledge, the only other analytically founded estimator is due to Siebert [50] who considers the Poisson equation and utilizes *cuboidal, rectangular* or *prismatic grids (meshes)*. He derives *one* error estimator for anisotropic meshes. More precisely, the well-known residual error estimator for isotropic meshes is modified to cope with the anisotropic meshes.

In our work (as in [35]) we concentrate on *tetrahedral* and *triangular anisotropic grids* which offer a greater geometrical flexibility. Understandably this requires more effort than for cuboidal grids since tetrahedra do not have three natural directions.

Apart from the choice of the elements, the aim of this work is threefold. We investigate *several error estimators* (residual error estimators, local problem error estimators, Zienkiewicz-Zhu like error estimators), *different norms* (energy norm,  $H^1$  seminorm,  $L_2$  norm), and *two types of differential equations*.

The Poisson equation of chapter 3 serves as a basic model problem to investigate the error estimators on anisotropic meshes. The singularly perturbed reaction-diffusion equation (cf. chapter 4) reveals further properties due to the anisotropy but also features that are related to the governing equation. Furthermore, this example shall show (or at least indicate) that our anisotropic theory can be applied to (almost) real-life problems.

We remark here that isotropic meshes can be considered as special anisotropic meshes (i.e. with bounded aspect ratio). Thus our whole theory can be applied to isotropic meshes as well. As it is to be expected, all the corresponding isotropic results are reproduced, and no artificial restrictions are imposed.

The underlying finite element method utilizes piecewise linear ansatz and test functions (higher order ansatz functions will be briefly discussed in Remark 3.12). The remaining paper is organized as follows.

In **chapter 2** we first introduce some notation. In contrast to cuboidal elements (meshes) we now do not dispose of three natural directions (axes). This feature gives the desired geometrical flexibility but aggravates the analysis at the same time.

The well-known *transformation technique* turns out to be particularly advantageous among all the notation and basic tools of that chapter. We discovered that matters can be facilitated by utilizing two different transformations. A finite element (tetrahedron)  $T$  can be mapped onto the ‘unitary tetrahedron’  $\bar{T}$ , or onto the (what we call) ‘reference tetrahedron’  $\hat{T}$ . The corresponding transformation matrices are vital in the further analysis.

The *mesh requirements* of our theory are quite general. We may stress here that no maximum angle condition is necessary for the error estimation.

Then some *basic tools* are introduced. These are two anisotropic trace inequalities, the bubble functions (which are analogous to their isotropic counterparts, cf. [58]), and so-called inverse inequalities for bubble functions and for finite element functions.

**Chapter 3** is devoted to the *Poisson equation* as a model problem for our investigations. Here we demonstrate the analysis of error estimation on anisotropic meshes. First some basic analytical results are stated. In the following Sections we present a residual error estimator and a local problem error estimator (both for anisotropic meshes of course). They are suitable for several kinds of boundary conditions, and they measure the error in the energy norm. The Zienkiewicz-Zhu like error estimators introduced next estimate the error in the  $H^1$  seminorm. Finally an error estimator for the  $L_2$  norm is derived. From a theoretical point of view the residual error estimator of Section 3.2 is the most important one since proofs for some other estimators are based on it. Hence we introduce there several important tools and motivate and discuss them in detail.

*Residual error estimators* (for isotropic meshes) and their analysis have been known for a long time (e.g. [13]). Residual error estimators for anisotropic cuboidal or prismatic meshes have been proposed by Siebert [50]. For anisotropic tetrahedral meshes we want to employ similar principles. Thus we start by deriving interpolation error estimates.

Different boundary conditions are taken into account. The main results are lower and upper bounds of the error (Theorem 3.4). As a by-product, *Robin boundary conditions* can now be treated for both anisotropic and isotropic meshes (as far as we know, the latter results is new as well). Furthermore the anisotropic analysis exhibits a new feature which does not occur for isotropic meshes – now it is important how good the anisotropic mesh corresponds to a given anisotropic function (in a certain sense to be defined). This correspondence influences both the interpolation error estimates and the finite element error estimate. Because of this importance (not only for error estimation but also for mesh generation/refinement) the matter is discussed extensively in Sections 3.2.1 and 3.2.4.

Motivated by *local problem error estimators* on isotropic meshes (cf. [14, 16, 58]), we extend such an estimator to anisotropic meshes. In Section 3.3 the anisotropic version is introduced and shown to be equivalent to the anisotropic residual error estimator. By this means lower and upper error bounds are derived which are the central results for this estimator (Theorem 3.7). We prove that the local problem to be solved is well-conditioned. Remarks on Robin boundary conditions and higher order ansatz functions conclude that Section. We may note here that the analysis of this error estimator turns out to be considerably more technical than in the isotropic case (see e.g. Lemma 3.5). New ideas and techniques (e.g. compactness arguments) are introduced.

Error estimators based on an averaged (or recovered) gradient [66, 67] are quite popular, amongst other things because they are comparatively cheap to compute. So we investigate whether they can be applied to anisotropic meshes, and we motivate and introduce three so-called anisotropic *Zienkiewicz-Zhu like error estimators*. A proof is established which shows the equivalence of one of our estimators with a modified residual error estimator. Note, however, that the whole theory of Zienkiewicz-Zhu like error estimators is less well developed for anisotropic meshes.

An  $L_2$  error estimator for non-uniform isotropic meshes has been derived by Eriksson and Johnson [29]. Ainsworth and Oden [1] present a slightly different analysis, and Verfürth [58] additionally proves a lower error bound. In Section 3.5 we utilize some of these ideas and propose an  $L_2$  error estimator that is suitable for anisotropic meshes. As it is common in  $L_2$  error estimation, duality arguments play an important role. The analysis requires similar tools and techniques as for the residual error estimator (in the energy norm). These techniques have to be modified to suit the needs of  $L_2$  error estimation. To this end, we introduce special anisotropic  $L_2$  bubble functions (which are different from the ones of [58]), and prove the corresponding inverse inequalities. Together with further anisotropic interpolation error estimates the lower and upper bounds of the finite element error are derived (Theorem 3.14). These bounds render our  $L_2$  error estimator reliable and efficient.

The Poisson problem is well suited to study the effects, requirements, and difficulties of *a posteriori* error estimation on anisotropic meshes. Real-world problems, however, are more complicated. Some important problem classes which frequently yield anisotropic solutions include diffusion-convection-reaction equations, and flow simulations (see e.g. [42, 45, 60]). In **chapter 4** we consider the *singularly perturbed reaction-diffusion equation*  $-\varepsilon\Delta u + u = f$  as a model problem. It displays certain typical features of the aforementioned problems, for example boundary layers which can be discretized advantageously with anisotropic meshes. Yet a rigorous analysis is still possible, although the knowledge of *a posteriori* error estimators (for isotropic meshes) has been unsatisfactory for a long time. For most error estimators the upper and lower bounds of the error are not asymptotically equivalent, i.e. they differ by a factor that increases, for example, as

$\varepsilon \rightarrow 0$ . The only two error estimators with asymptotically equivalent upper and lower bound on the error are, to our knowledge, due to Angermann [3] (for a special norm) and to Verfürth [59] (for the energy norm).

Based on the latter paper, we are aiming to estimate the error for anisotropic meshes. Some new tools are required, including special bubble functions, inverse inequalities and interpolation error estimates. Fortunately one can partly utilize the methodology of the Poisson equation and the results which were used there. An anisotropic *residual error estimator* is proposed, and lower and upper bounds in the energy norm are proven. This main result for the singularly perturbed problem is presented in Theorem 4.4.

**Chapter 5** is devoted to the *numerical experiments*. Many of the results that we obtained are now numerically investigated. The emphasis of the experiments clearly lies on error estimation but further aspects (e.g. interpolation error estimates) are considered as well. We mainly pursue two aims. Firstly we want to find out whether a certain inequality only holds in an asymptotic sense (i.e. for many degrees of freedom), or if it yields useful results also for specific examples. Secondly we can learn about the size, range, and distribution of inequality constants. This can indicate about the quality/sharpness of certain estimates.

We consider several examples, and consider mesh sequences as well as single, particular meshes to broaden our insight into the numerical behaviour of the estimates. The experiments cover two-dimensional and three-dimensional examples. The computations were carried out on serial and on parallel machines. As a by-product this shows that the computation of the error estimators can be parallelized without much effort. In Section 5.3 conclusions of all numerical examples are recapitulated.

Finally we summarize the results obtained. Furthermore we present some open problems which we think to be of importance.

# Chapter 2

## Preliminaries

### 2.1 Notation and basic properties

Let  $\Omega \subset \mathbb{R}^d$ ,  $d = 2, 3$ , be an open, bounded, polygonal or polyhedral domain over which the differential equation is posed. All considerations are made for the three-dimensional case. The application to the easier two-dimensional case is readily possible. Norms of functions are  $L_2$  norms unless otherwise stated. A norm without subscript denotes  $\|\cdot\| = \|\cdot\|_{L_2(\Omega)}$ , i.e. the  $L_2$  norm over the whole domain  $\Omega$ . All vector norms are Euclidean norms, and norms of matrices are spectral norms.

Constants are denoted by  $c$ . They are generic constants, i.e. always independent of the underlying triangulation or of the function in question, and they may have different values at different occurrences. We write

$$\begin{aligned} x \lesssim y &\iff x \leq c \cdot y \\ x \gtrsim y &\iff x \geq c \cdot y, & c > 0 \\ x \sim y &\iff \underline{c} \cdot x \leq y \leq \bar{c} \cdot x, & \underline{c} > 0 \end{aligned} .$$

The notation with an explicit constant  $c$  is used only when a dependence on some other values is expressed (e.g.  $c_\Omega$ ) or when further details are thus revealed. Also, the sharper notation  $\leq$  is used instead of  $\lesssim$  wherever possible.

#### 2.1.1 Notation of the tetrahedron

Assume that a triangulation  $\mathcal{T}_h$  (also called a mesh or a grid) is given which satisfies the usual conformity conditions (see Ciarlet [26], Chapter 2). Let  $T$  be an arbitrary tetrahedron thereof. For this tetrahedron the following notation is introduced. The four vertices of  $T$  are denoted by  $P_0, \dots, P_3$  according to these three conditions:<sup>1</sup>

- Let  $P_0P_1$  be the longest edge of  $T$ .
- There exist two triangles that contain the edge  $P_0P_1$ . The one with largest area is denoted by  $\triangle P_0P_1P_2$ .
- Let  $P_0P_2$  be the shortest edge of  $\triangle P_0P_1P_2$ . This determines which vertex is  $P_0$  and  $P_1$ , respectively. Let  $P_3$  be the remaining vertex.

---

<sup>1</sup>Later on, we use the same notation  $P_1, P_2, P_3$  almost exclusively for certain mappings, but the meaning will always be clear from the context.

This notation is not uniquely determined if, for instance,  $T$  has two edges which are simultaneously the longest ones. However, it turns out that then either choice of the notation fits into the theory. Additionally we define three vectors:

- $\mathbf{p}_1 := \overrightarrow{P_0 P_1}$ .
- Let  $\mathbf{p}_2$  be that vector in the plane of  $P_0 P_1 P_2$  that points to  $P_2$  and that is perpendicular to  $\mathbf{p}_1$ .
- Let  $\mathbf{p}_3$  be that vector to  $P_3$  that is perpendicular to  $\triangle P_0 P_1 P_2$ .

Hence  $\mathbf{p}_1 \dots \mathbf{p}_3$  are mutually orthogonal. Figure 2.1 visualizes this notation. In the two dimensional case one only has the triangle  $P_0 P_1 P_2$  which is denoted in exactly the same way.

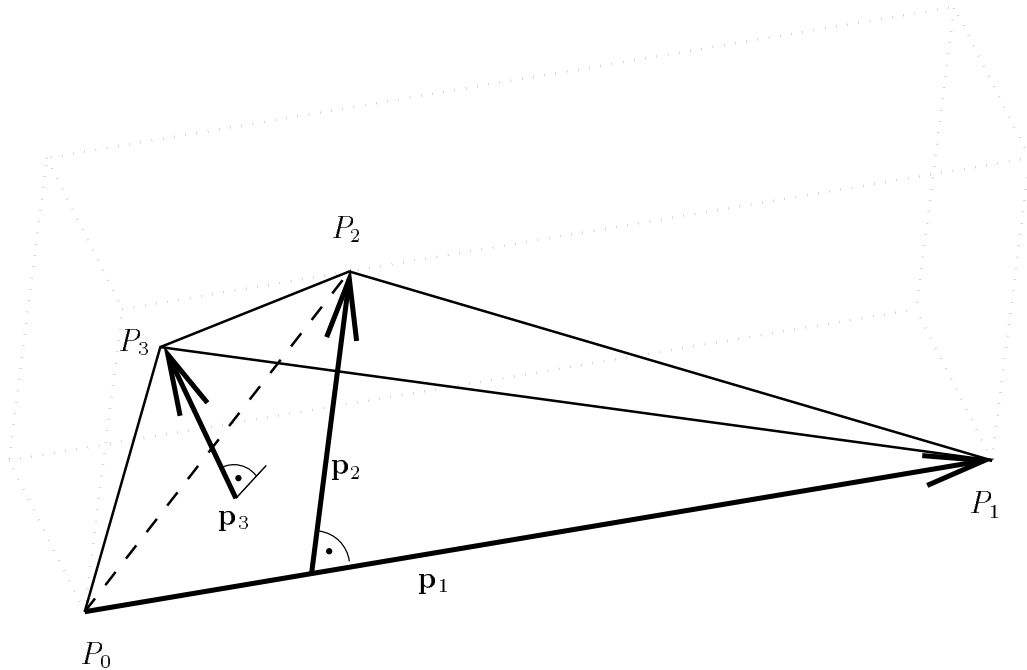


Figure 2.1: Notation of tetrahedron  $T$

The length of the vectors  $\mathbf{p}_i$  is denoted by

$$h_i = h_{i,T} := |\mathbf{p}_i| \quad i = 1, 2, 3 \quad .$$

Because of the definition of the  $P_i$  we conclude immediately  $h_1 > h_2 \geq h_3$ . We further define

$$h_{\min,T} := \min_{i=1 \dots d} \{h_{i,T}\} = h_{d,T}$$

(in  $\mathbb{R}^3$  thus  $h_{\min,T} = h_{3,T}$  holds). Furthermore piecewise constant functions  $h_i(\mathbf{x})$  and  $h_{\min}(\mathbf{x})$  are defined for almost all  $\mathbf{x} \in \Omega$  according to

$$h_i(\mathbf{x}) := h_{i,T} \quad h_{\min}(\mathbf{x}) := h_{\min,T} \quad \text{for } \mathbf{x} \in T, i = 1 \dots d. \quad (2.1)$$

The boundary of a tetrahedron  $T$  consists of four faces (i.e. triangles). Such a face is denoted by  $E$ , and its  $(d-1)$  dimensional measure is expressed by  $|E| := \text{meas}_{d-1}(E)$ . The length of the height over such a face  $E$  will be denoted by

$$h_{E,T} := 3|T|/|E| \quad .$$

### 2.1.2 Standard and reference tetrahedron, transformations and their properties

Let  $T$  be an arbitrary but fixed tetrahedron. Mainly we will employ two affine linear mappings  $F_A$  and  $F_C$  which will be defined as follows.

Let  $\vec{P}_0$  be the (column) vector from the origin of the coordinate system to  $P_0$ , and let  $\vec{P}_0P_i$  be the (column) vectors from  $P_0$  to  $P_i$ ,  $i = 1, 2, 3$ . We define the matrices  $A_T, C_T \in \mathbb{R}^{3 \times 3}$  by

$$A_T := \begin{pmatrix} \vec{P}_0P_1, \vec{P}_0P_2, \vec{P}_0P_3 \end{pmatrix} \quad \text{and} \quad C_T := \begin{pmatrix} \mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3 \end{pmatrix} \quad . \quad (2.2)$$

Sometimes we want to refer to the matrix  $C_T$  not only on an actual tetrahedron  $T$  but on a larger domain. Thus we introduce a matrix (or more precisely a matrix function)  $C(\mathbf{x})$  which is defined globally for almost all  $\mathbf{x} \in \Omega$  and which coincides with  $C_T$  on the (interior of a) tetrahedron  $T$ :

$$C(\mathbf{x}) := C_T \quad \text{for } \mathbf{x} \in T \quad . \quad (2.3)$$

Additionally a matrix  $H_T$  is defined by

$$H_T := \text{diag}(h_1, h_2, h_3) \quad .$$

With the help of the matrices  $A_T$  and  $C_T$  we now define the affine linear transformations

$$F_A(\boldsymbol{\mu}) := A_T \cdot \boldsymbol{\mu} + \vec{P}_0 \quad \text{and} \quad F_C(\boldsymbol{\mu}) := C_T \cdot \boldsymbol{\mu} + \vec{P}_0$$

with  $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)^T$ . These mappings are closely related to the following tetrahedra.

**Definition 2.1 (Standard tetrahedron and reference tetrahedron)** *The standard or unitary tetrahedron  $\bar{T}$  is defined by its vertices  $\bar{P}_0 = (0, 0, 0)^T$  and  $\bar{P}_i = \mathbf{e}_i^T$ ,  $i = 1 \dots d$ . Enumerate the faces  $\bar{E}_i$  of  $\bar{T}$  such that*

$$\bar{E}_i := \bar{T} \cap \{x_i = 0\} \quad , i = 1 \dots d \quad \text{and} \quad \bar{E}_0 := \bar{T} \cap \{|\mathbf{x}|_1 = 1\} \quad ,$$

*i.e. face  $\bar{E}_i$  is opposite the vertex  $\bar{P}_i$ .*

*The reference tetrahedron  $\hat{T}$  is defined implicitly by the mapping  $F_C$ , i.e.  $\hat{T} := F_C^{-1}(T)$ .*

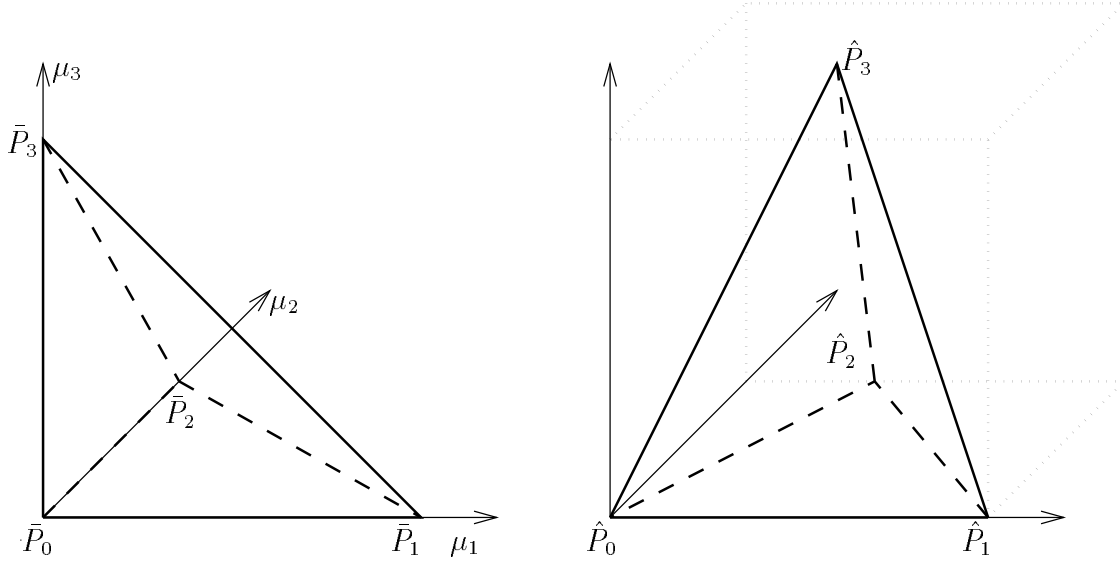
The definition readily implies

$$F_A : \bar{T} \mapsto T \quad \text{and} \quad F_C : \hat{T} \mapsto T \quad .$$

The vertices of  $\hat{T}$  are  $\hat{P}_0 = (0, 0, 0)^T$ ,  $\hat{P}_1 = (1, 0, 0)^T$ ,  $\hat{P}_2 = (\hat{x}_2, 1, 0)^T$  and  $\hat{P}_3 = (\hat{x}_3, \hat{y}_3, 1)^T$  because of the definition of  $F_C$ . The conditions on the  $P_i$  yield immediately  $0 < \hat{x}_2 \leq 1/2$ ,  $0 < \hat{x}_3 < 1$  and  $-1 < \hat{y}_3 < 1$ . Figures 2.1 and 2.2 may illustrate this definition (the circumscribed rectangular prisms shall facilitate the visualization).

The two transformations play a vital role in many proofs, but they serve different purposes. Basically, an inequality on the standard tetrahedron  $\bar{T}$  depends on  $\bar{T}$  but not on  $T$ . Such an inequality can only be transformed via  $F_A$ . In contrast, the transformation via  $F_C$  is better suited to the anisotropic elements, as it turns out. Then one has to prove the inequality under consideration on the reference tetrahedron  $\hat{T}$ . This inequality now *does* depend on  $\hat{T}$  and  $T$ , and a more sophisticated analysis is required.

Variables that are related to the standard tetrahedron  $\bar{T}$  and the reference tetrahedron  $\hat{T}$  are referred to with a bar and a hat, respectively (e.g.  $\bar{\nabla}$ ,  $\hat{v}$ ).

Figure 2.2: Standard tetrahedron  $\bar{T}$  and reference tetrahedron  $\hat{T}$ 

The determinants of both mappings are

$$|\det(A_T)| = |\det(C_T)| = h_1 \cdot h_2 \cdot h_3 = 6 \cdot |T| \quad .$$

The transformed derivatives satisfy

$$\bar{\nabla} \bar{v} = A_T^T \nabla v \quad \text{and} \quad \hat{\nabla} \hat{v} = C_T^T \nabla v \quad .$$

In order to bound the norms of some transformation matrix we state the following simple Lemma (see also [26]).

**Lemma 2.1 (Bound of the norm of a transformation matrix)** *Let  $A$  be a linear transformation that maps the (closed) domain  $\hat{G} \subset \mathbb{R}^d$  onto  $G$ . The spectral norm of the corresponding transformation matrix satisfies*

$$\|A\mathbf{x}\|_{\mathbb{R}^d} / \|\mathbf{x}\|_{\mathbb{R}^d} \leq \|A\|_{\mathbb{R}^d \times d} \leq d(G) / \varrho(\hat{G}) \quad \forall \mathbf{x} \in \mathbb{R}^d, \mathbf{x} \neq \mathbf{0}$$

with  $d(G) := \max_{\mathbf{y}, \mathbf{z} \in G} \|\mathbf{y} - \mathbf{z}\|_{\mathbb{R}^d}$  and  $\varrho(\hat{G}) := \text{diameter of the largest sphere } S \subset \hat{G}$ .

**Lemma 2.2 (Norms of some matrices)** *The following relations hold.*

$$\|A_T^T C_T^{-T}\|_{\mathbb{R}^{3 \times 3}} = \|C_T^{-1} A_T\|_{\mathbb{R}^{3 \times 3}} \sim 1 \quad (2.4)$$

$$\|C_T^T A_T^{-T}\|_{\mathbb{R}^{3 \times 3}} = \|A_T^{-1} C_T\|_{\mathbb{R}^{3 \times 3}} \sim 1 \quad (2.5)$$

$$\|C_T H_T^{-1}\|_{\mathbb{R}^{3 \times 3}} = \|H_T C_T^{-1}\|_{\mathbb{R}^{3 \times 3}} = 1 \quad (2.6)$$

$$\|H_T^{-1}\|_{\mathbb{R}^{3 \times 3}} = \|C_T^{-1}\|_{\mathbb{R}^{3 \times 3}} = h_{\min, T}^{-1} \quad (2.7)$$

$$\|A_T^{-1}\|_{\mathbb{R}^{3 \times 3}} \sim h_{\min, T}^{-1} \quad (2.8)$$

$$\|A_T\|_{\mathbb{R}^{3 \times 3}} \sim \|C_T\|_{\mathbb{R}^{3 \times 3}} \sim h_{1, T} \quad (2.9)$$

**Proof:** Let  $T - \vec{P}_0$  be the tetrahedron  $T$  shifted by  $-\vec{P}_0$ . The mappings  $A_T$ ,  $C_T^{-1}$  and  $C_T^{-1}A_T$  act as follows:

$$\bar{T} \xrightarrow{A_T} (T - \vec{P}_0) \xrightarrow{C_T^{-1}} \hat{T} \quad \text{and thus} \quad \bar{T} \xrightarrow{C_T^{-1}A_T} \hat{T},$$

i.e.  $C_T^{-1}A_T$  maps the standard tetrahedron  $\bar{T}$  onto the reference tetrahedron  $\hat{T}$ . The Lemma from above implies immediately  $\|C_T^{-1}A_T\|_{\mathbb{R}^{3 \times 3}} \lesssim 1$ . From  $C_T^{-1}A_T : \bar{P}_0\bar{P}_1 \mapsto \hat{P}_0\hat{P}_1$  one has  $1 \leq \|C_T^{-1}A_T\|_{\mathbb{R}^{3 \times 3}}$  which proves (2.4) and analogously (2.5).

Because of  $C_T^T \cdot C_T = H_T^2$  from (2.2) we conclude  $(H_T C_T^{-1})^T \cdot H_T C_T^{-1} = I$  and  $(C_T H_T^{-1})^T \cdot C_T H_T^{-1} = I$ . Thus (2.6) is derived. Note that  $\|C_T^{-1}H_T\|_{\mathbb{R}^{3 \times 3}} \neq 1$ .

The equality  $\|H_T^{-1}\|_{\mathbb{R}^{3 \times 3}} = h_{\min, T}^{-1}$  is obvious. Equality (2.7) follows immediately from  $\|C_T^{-1}\|_{\mathbb{R}^{3 \times 3}}^2 = \lambda_{\max}(C_T^{-1}C_T^{-T}) = \lambda_{\max}(H_T^{-2})$ .

The inequalities  $\|C_T^{-1}\|_{\mathbb{R}^{3 \times 3}} = \|C_T^{-1}A_T \cdot A_T^{-1}\|_{\mathbb{R}^{3 \times 3}} \leq \|C_T^{-1}A_T\|_{\mathbb{R}^{3 \times 3}} \cdot \|A_T^{-1}\|_{\mathbb{R}^{3 \times 3}}$  and  $\|A_T^{-1}\|_{\mathbb{R}^{3 \times 3}} = \|A_T^{-1}C_T \cdot C_T^{-1}\|_{\mathbb{R}^{3 \times 3}} \leq \|A_T^{-1}C_T\|_{\mathbb{R}^{3 \times 3}} \cdot \|C_T^{-1}\|_{\mathbb{R}^{3 \times 3}}$  and (2.4)–(2.7) eventually imply (2.8).

Finally, relation (2.9) follows immediately from the definition (2.2) of the matrices, and the length of the vectors contained therein.  $\blacksquare$

Finally, a norm  $\|\cdot\|_T$  over an actual tetrahedron  $T$  is often transformed into a norm over the standard tetrahedron  $\bar{T}$  or the reference tetrahedron  $\hat{T}$ . The following relations hold. Let  $v \in L_2(T)$  and  $T \subset \mathbb{R}^3$ . For a mapping  $F_A(\boldsymbol{\mu}) = A_T \cdot \boldsymbol{\mu} + \vec{P}_0$  one obtains

$$\begin{aligned} \int_T v^2(\mathbf{x})d\mathbf{x} &= \int_{\bar{T}} \bar{v}^2(\boldsymbol{\mu}) \cdot |\det A_T| d\boldsymbol{\mu} = 6|T| \cdot \int_{\bar{T}} \bar{v}^2(\boldsymbol{\mu})d\boldsymbol{\mu} \\ \text{or} \quad \|v\|_T &= \sqrt{6|T|} \cdot \|\bar{v}\|_{\bar{T}} \\ \text{and similarly} \quad \|v\|_T &= \sqrt{6|T|} \cdot \|\hat{v}\|_{\hat{T}} \\ \text{and} \quad \|v\|_E &= \sqrt{|E|/|\bar{E}|} \cdot \|\bar{v}\|_{\bar{E}} \quad E \subset \partial T \quad . \end{aligned}$$

### 2.1.3 The directional derivative $\tilde{D}_i$

In order to motivate the derivatives  $\tilde{D}_i$  consider rectangular or cuboidal finite elements (cf. [50]). There are three (or two) natural directions that correspond to the coordinate axes. The partial derivatives that correspond to these axes too are thus sufficient for an error analysis.

In contrast to this a tetrahedron or a triangle does not possess these natural directions. However the (normalized) directions  $\mathbf{p}_1, \mathbf{p}_2$ , and  $\mathbf{p}_3$  that correspond to  $C_T$  will prove to be useful. This leads to the following definition.

**Definition 2.2 (Directional derivative)** *Let  $v$  be a function in  $H^1(T)$ . The directional derivative  $\tilde{D}_{i,T}$  is defined by*

$$\begin{pmatrix} \tilde{D}_{1,T} v \\ \tilde{D}_{2,T} v \\ \tilde{D}_{3,T} v \end{pmatrix} := H_T^{-1} C_T^T \cdot \nabla v \quad , v \in H^1(T). \quad (2.10)$$

Here this derivative  $\tilde{D}_{i,T}$  is defined for a fixed tetrahedron  $T$ . Hence we introduce a derivative  $\tilde{D}_i$  which is defined globally for almost all  $\mathbf{x} \in \Omega$ , and which coincides with  $\tilde{D}_{i,T}$

on a tetrahedron  $T$ :

$$\tilde{D}_i v(\mathbf{x}) := \tilde{D}_{i,T} v(\mathbf{x}) \quad \text{for } \mathbf{x} \in T \quad .$$

Note that this derivative  $\tilde{D}_i$  depends on the triangulation  $\mathcal{T}_h$ , and it is defined separately over each tetrahedron  $T$ .

When considering each component in the definition above, the directional derivative is equivalent to

$$\tilde{D}_{i,T} v = h_i^{-1} \cdot (\mathbf{p}_i, \nabla v) \quad i = 1 \dots d$$

i.e.  $\tilde{D}_{i,T}$  is the (unitary) directional derivative along the direction  $\mathbf{p}_i$ .

The orthogonality of the vectors  $\mathbf{p}_i$  and the definition  $h_i = |\mathbf{p}_i|$  implies that  $H_T^{-1} C_T^T$  is an orthogonal matrix. Thus

$$\sum_{i=1}^d (\tilde{D}_i v)^2 = |\nabla v|^2 \quad \text{or} \quad \|H_T^{-1} C_T^T \nabla v\|_T = \|\nabla v\|_T \quad (2.11)$$

$$\text{and} \quad \sum_{i=1}^d h_i^2 (\tilde{D}_i v)^2 = |C_T^T \nabla v|^2 \quad \text{or} \quad \sum_{i=1}^d h_i^2 \|\tilde{D}_{i,T} v\|_T^2 = \|C^T \nabla v\|_T^2. \quad (2.12)$$

The last equations indicate that derivatives can be written either component-wise in terms of  $\tilde{D}_i$ , or they can be written in the compact form of  $C_T^T \nabla$ .

In this work all results and proofs are given in this compact form since  $C_T^T \nabla$  on the actual tetrahedron  $T$  is related (via  $F_C$ ) directly to  $\hat{\nabla}$  on the reference tetrahedron  $\hat{T}$ . Main results however are also given in terms of  $\tilde{D}_i$  for two reasons. Firstly this might facilitate the understanding of the underlying principles, and secondly an extension to rectangular or cuboidal finite elements is then readily possible.

With the help of (2.11) and (2.12) any term involving derivatives can be expressed easily in either form.

## 2.1.4 Notation of the summation

In the derivation of error estimators many sums appear. For an unambiguous yet clear and short notation the following conventions will be used.

*Tetrahedra* (or triangles in 2D) are denoted by  $T, T', T''$  or  $T_i$ . Sums are written in the form  $\sum_{T \in \mathcal{T}_h}$  or  $\sum_{T \cap \Gamma_D \neq \emptyset}$ , for example.

*Faces* of a tetrahedron (or an edge of a triangle in 2D) are denoted by  $E$  or  $E_i$ . The sum over all inner faces of  $\Omega$ , or the sum over all non-Dirichlet faces of  $T$  are, for example, denoted by  $\sum_{E \subset \Omega \setminus \Gamma}$  or  $\sum_{E \subset \partial T \setminus \Gamma_D}$ , respectively.

*Nodes* (or vertices) of a tetrahedron (or a triangle in 2D) are denoted by  $a$  or  $a_j$ . For a clear reference we define the following sets.

- $\mathcal{N}$  set of all nodes of  $\mathcal{T}_h$
- $\mathcal{N}_I$  set of all interior nodes of  $\mathcal{T}_h$
- $\mathcal{N}_T$  set of all nodes of a tetrahedron  $T$
- $\mathcal{N}_E$  set of all nodes of a face  $E$
- $\mathcal{N}_D, \mathcal{N}_R$  set of all nodes contained in  $\Gamma_D$  or  $\bar{\Gamma}_R$ , respectively
- $\mathcal{N}_{R,\alpha}$  special subset of  $\mathcal{N}_R$ ; defined by (3.9)

Sums are then written in the form  $\sum_{a \in \mathcal{N}_I}$  or  $\sum_{a_j \in \mathcal{N}_{R,\alpha} \setminus \mathcal{N}_D}$ , respectively.

### 2.1.5 Auxiliary subdomains

Two auxiliary subdomains that occur in many estimates are defined now. Let  $T \in \mathcal{T}_h$  be an arbitrary tetrahedron. Let  $\omega_T$  be that domain that is formed by  $T$  and all (at most four) adjacent tetrahedra that have a common face with  $T$ :

$$\omega_T := \bigcup_{T' \cap T = E} T' \quad .$$

Note that  $\omega_T$  consists of less than five tetrahedra if  $T$  has a boundary face.

Let  $E$  be an inner face (triangle) of  $\mathcal{T}_h$ , i.e. there are two tetrahedra  $T_1$  and  $T_2$  having the common face  $E$ . Let the domain  $\omega_E := T_1 \cup T_2$ . If  $E$  is a boundary face set  $\omega_E := T$  with  $T \supset E$ .

Figure 2.3 depicts both domains for the two-dimensional and the three-dimensional case.

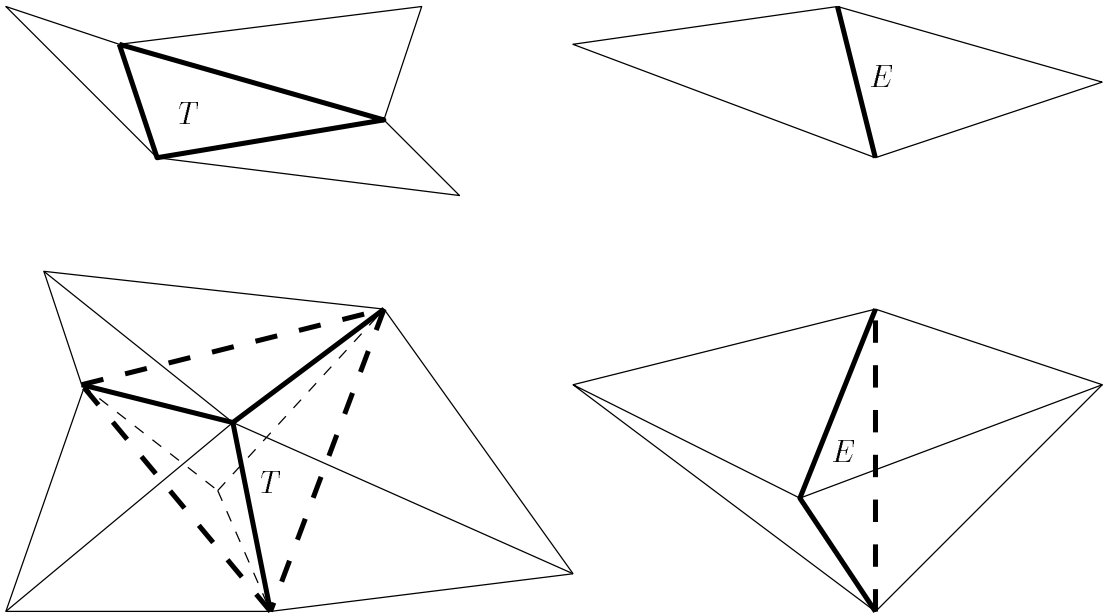


Figure 2.3: Auxiliary domains  $\omega_T$  (left) and  $\omega_E$  (right) for  $\Omega \subset \mathbb{R}^2$  and  $\Omega \subset \mathbb{R}^3$

## 2.2 Requirements on the mesh

Let  $a_1, \dots, a_N$  be the nodes of the triangulation  $\mathcal{T}_h$ . In addition to the usual conformity conditions of the mesh (see Ciarlet [26], Chapter 2) we demand the following assumptions.

1. The number of tetrahedra that contain the node  $a_j$  is bounded uniformly.
2. The dimensions of adjacent tetrahedra must not change rapidly, i.e.

$$h_{i,T'} \sim h_{i,T} \quad \forall T, T' \text{ with } T \cap T' \neq \emptyset, i = 1 \dots d \quad . \quad (2.13)$$

**Remark 2.1** Note that we do *not* assume a maximum angle condition for our error estimation.  $\square$

**Remark 2.2** Assume that  $T$  and  $T'$  are adjacent tetrahedra. If in any inequality the terms  $h_{i,T}$  or  $h_{i,T'}$  occur then assumption (2.13) implies that both terms can be exchanged mutually. The inequality constants are then still independent of  $T$  or  $\mathcal{T}_h$ . This feature is exploited to some extend.

Assumption (2.13) implies in particular that we can use a term  $h_{i,T} = h_i$  for describing the dimension  $h_i$  of a local subdomain like  $\omega_T$  or  $\omega_E$ .

Finally, consider a face  $E = T_1 \cap T_2$ . The length of the heights  $h_{E,T_1} \sim h_{E,T_2}$  are related to the tetrahedra  $T_1$  and  $T_2$ , respectively. However, in some places (for example, if one has a sum over all interior faces like (3.14)) it is advantageous to have an equivalent term  $h_E$  which is related only to  $E$ , i.e. where it is not necessary to state explicitly whether  $h_{E,T_1}$  or  $h_{E,T_2}$  is chosen. For an unambiguous notation we thus define

$$h_E := (h_{E,T_1} + h_{E,T_2})/2 \quad ,$$

with obvious modifications if  $E$  is a boundary face. Analogously define<sup>2</sup>

$$h_{min,E} := (h_{min,T_1} + h_{min,T_2})/2 \quad . \quad \square$$

## 2.3 Basic tools

In this Section some basic tools and inequalities are listed.

### 2.3.1 Anisotropic trace inequalities

The first trace inequality is readily obtained by standard scaling techniques.

**Lemma 2.3 (First trace inequality)** *Let  $T$  be an arbitrary tetrahedron and  $E$  be a face of it. For  $v \in H^1(T)$  the trace inequality*

$$\|v\|_E^2 \lesssim h_E^{-1} (\|v\|_T^2 + \|C_T^T \nabla v\|_T^2) \quad (2.14)$$

*holds. The component-wise form is*

$$\|v\|_E^2 \lesssim h_E^{-1} \cdot \left( \|v\|_T^2 + \sum_{i=1}^d h_{i,T}^2 \|\tilde{D}_i v\|_T^2 \right) \quad .$$

**Proof:** Consider the transformation  $F_A$ , the standard tetrahedron  $\bar{T} := F_A^{-1}(T)$ , the face  $\bar{E} := F_A^{-1}(E)$  of  $\bar{T}$ , and the function  $\bar{v} := v \circ F_A \in H^1(\bar{T})$ . The trace Theorem gives

$$\|\bar{v}\|_{\bar{E}}^2 \lesssim \|\bar{v}\|_{H^1(\bar{T})}^2 = \|\bar{v}\|_{\bar{T}}^2 + \|\bar{\nabla} \bar{v}\|_{\bar{T}}^2 \quad .$$

The transformation into the actual tetrahedron (via  $F_A$ ) yields

$$|E|^{-1} \cdot \|v\|_E^2 \lesssim |T|^{-1} (\|v\|_T^2 + \|A_T^T \nabla v\|_T^2) \quad .$$

From (2.4)

$$\|A_T^T \nabla v\|_T = \|A_T^T C_T^{-T} \cdot C_T^T \nabla v\|_T \leq \|A_T^T C_T^{-T}\|_{\mathbb{R}^{3 \times 3}} \cdot \|C_T^T \nabla v\|_T \lesssim \|C_T^T \nabla v\|_T$$

<sup>2</sup>Note that  $h_{min,E}$  is *not* the minimal size of the two-dimensional face  $E$ , as the notation could probably suggest.

can be derived. Utilizing  $6|T| = |E| \cdot h_E$  results in the trace inequality (2.14).  $\blacksquare$

The second, improved trace inequality in the isotropic version (i.e. on the standard tetrahedron) is, to our knowledge, due to Verfürth [59]. We state this inequality, repeat the proof for self-containment, and transform the inequality to our anisotropic context.

**Lemma 2.4 (Second trace inequality)** *Let  $T$  be an arbitrary tetrahedron and  $E$  be a face of it. For  $v \in H^1(T)$  the trace inequality*

$$\|v\|_E^2 \lesssim h_E^{-1} \cdot \|v\|_T \cdot (\|v\|_T + \|C_T^T \nabla v\|_T) \quad (2.15)$$

holds. The component-wise form is

$$\|v\|_E^2 \lesssim h_E^{-1} \cdot \|v\|_T \cdot \left( \|v\|_T + \sum_{i=1}^d h_{i,T} \|\tilde{D}_i v\|_T \right) .$$

**Proof:** Again standard scaling arguments will be used. Therefore consider first the standard tetrahedron  $\bar{T}$ .

Let  $\bar{v} \in H^1(\bar{T})$  vanish on  $\bar{E}_0$ . Then

$$\|\bar{v}\|_{\bar{E}_k}^2 \leq 2 \cdot \|\bar{v}\|_{\bar{T}} \cdot \|\partial \bar{v} / \partial x_k\|_{\bar{T}}$$

holds for  $k = 1 \dots d$ . To derive this consider a fixed index  $k$ . Since  $\bar{v}$  vanishes on  $\bar{E}_0$  we obtain for all  $\mathbf{y} \in \bar{E}_k$

$$|\bar{v}(\mathbf{y})|^2 = |\bar{v}(\mathbf{y})|^2 - |\bar{v}(\mathbf{y} + (1 - |\mathbf{y}|_1)e_k)|^2 = -2 \int_0^{1-|\mathbf{y}|_1} \bar{v}(\mathbf{y} + te_k) \cdot \frac{\partial}{\partial x_k} \bar{v}(\mathbf{y} + te_k) dt$$

since  $\mathbf{y} + (1 - |\mathbf{y}|_1)e_k \in \bar{E}_0$ . Integrating over  $\bar{E}_k$ , invoking Fubini's Theorem and the Cauchy-Schwarz inequality establishes the desired estimate.

Consider now a function  $v \in H^1(\bar{T})$  that vanishes on an arbitrary face  $\bar{E}_i, 0 \leq i \leq d$ . Let  $\bar{E}$  be a face of  $\bar{T}$ . Then

$$\|v\|_{\bar{E}}^2 \lesssim \|v\|_{\bar{T}} \cdot \|\nabla v\|_{\bar{T}} .$$

To prove this assume  $\bar{E} \neq \bar{E}_i$  since otherwise the inequality is trivial. We employ an affine linear mapping  $F_i$  which satisfies

$$\begin{aligned} F_i &: \mathbf{x}(\boldsymbol{\mu}) = F \cdot \boldsymbol{\mu} + \boldsymbol{\mu}_0 && \text{with } F \in \mathbb{R}^{d \times d} \\ F_i &: \bar{T} \mapsto \bar{T} && \text{and } \bar{E}_0 \mapsto \bar{E}_i . \end{aligned}$$

Assume that the face  $\bar{E}_k$  is mapped onto  $\bar{E}$ , with  $k \neq 0$ . The function  $\bar{v} := v \circ F_i$  vanishes on  $\bar{E}_0$  and thus the previous inequality implies

$$\|\bar{v}\|_{\bar{E}_k}^2 \leq 2 \cdot \|\bar{v}\|_{\bar{T}} \cdot \|\partial \bar{v} / \partial x_k\|_{\bar{T}} .$$

Lemma 2.1 yields readily  $\|F\|_{\mathbb{R}^{d \times d}} \lesssim 1$ , and  $|\bar{E}|/|\bar{E}_k| \lesssim 1$  is obvious. The transformation back to  $v$  results in the desired inequality

$$\|v\|_{\bar{E}}^2 \lesssim \|v\|_{\bar{T}} \cdot \|e_k^T F^T \nabla v\|_{\bar{T}} \leq \|v\|_{\bar{T}} \cdot \|e_k^T F^T\|_{\mathbb{R}^d} \cdot \|\nabla v\|_{\bar{T}} \lesssim \|v\|_{\bar{T}} \cdot \|\nabla v\|_{\bar{T}} .$$

Consider finally an arbitrary function  $v \in H^1(\bar{T})$ . Let  $\bar{E}$  be any of the faces of  $\bar{T}$ , and enumerate the vertices of  $\bar{T}$  such that the vertices of  $\bar{E}$  are numbered first. Denote by  $\lambda_1 \cdots \lambda_{d+1}$  the barycentric coordinates of  $\bar{T}$ . Then  $\lambda_1 + \cdots + \lambda_d = 1$  on  $\bar{E}$ , and thus

$$\|v\|_{\bar{E}} \leq \sum_{i=1}^d \|\lambda_i \cdot v\|_{\bar{E}} \lesssim \sum_{i=1}^d \|\lambda_i \cdot v\|_{\bar{T}}^{1/2} \cdot \|\nabla(\lambda_i \cdot v)\|_{\bar{T}}^{1/2}$$

since  $\lambda_i \cdot v$  vanishes on  $\bar{E}_i$ . The chain rule, the Cauchy-Schwarz inequality, the actual representation of  $\lambda_i$ , and  $|\lambda_i| \leq 1$  imply

$$\|\nabla(\lambda_i \cdot v)\|_{\bar{T}}^2 = \sum_{j=1}^d \left\| v \cdot \frac{\partial \lambda_i}{\partial x_j} + \lambda_i \cdot \frac{\partial v}{\partial x_j} \right\|_{\bar{T}}^2 \leq 4 \cdot \|v\|_{\bar{T}}^2 + 2 \cdot \|\nabla v\|_{\bar{T}}^2$$

yielding 
$$\|v\|_{\bar{E}}^2 \lesssim \|v\|_{\bar{T}} \cdot (\|v\|_{\bar{T}} + \|\nabla v\|_{\bar{T}}) \quad .$$

This constitutes the trace inequality on the standard tetrahedron  $\bar{T}$ . The transformation onto the actual tetrahedron  $T$  is completely analogous to the proof of the first trace inequality and therefore omitted here.  $\blacksquare$

### 2.3.2 Inverse inequalities for finite element functions

In several proofs the following inverse inequalities are required. Note that, strictly speaking, (2.16) and (2.17) constitute norm equivalences (over finite dimensional spaces) whereas (2.18) does not.

**Lemma 2.5** *Let  $T$  be an arbitrary tetrahedron,  $E$  a face of it, and  $v_h \in V_h(T) = \mathbb{P}^1(T)$  a finite element function over  $T$ . Then*

$$\|v_h\|_{\infty, T} \sim |T|^{-1/2} \cdot \|v_h\|_T \quad (2.16)$$

$$\|v_h\|_{\infty, E} \sim |E|^{-1/2} \cdot \|v_h\|_E \quad (2.17)$$

hold. Furthermore, let  $w_h \in \mathbb{P}^4(T)$ . Then

$$\begin{aligned} \|C_T^T \nabla w_h\|_T &\lesssim \|w_h\|_T \\ \text{or} \quad \|\tilde{D}_i w_h\|_T &\lesssim h_{i,T}^{-1} \cdot \|w_h\|_T \quad i = 1 \dots d \end{aligned} \quad (2.18)$$

in its component-wise form.

**Proof:** The proofs are again based on the transformation technique. The norm equivalence

$$\|\bar{v}_h\|_{\infty, \bar{T}} \sim \|\bar{v}_h\|_{\bar{T}} \quad \forall \bar{v}_h \in V_h(\bar{T}) = \mathbb{P}^1(\bar{T})$$

holds on the standard tetrahedron  $\bar{T}$  since both norms act on the finite dimensional space  $\mathbb{P}^1(\bar{T})$ . Via

$$\|v_h\|_{\infty, T}^2 = \|\bar{v}_h\|_{\infty, \bar{T}}^2 \sim \|\bar{v}_h\|_{\bar{T}}^2 = |\det A_T|^{-1} \cdot \|v_h\|_T^2$$

one readily obtains (2.16). For a face  $E$  and (2.17) proceed similarly.

The third inequality is derived analogously. Starting with

$$\|\bar{\nabla} \bar{w}_h\|_{\bar{T}} \lesssim \|\bar{w}_h\|_{\bar{T}} \quad \forall \bar{w}_h \in \mathbb{P}^4(\bar{T}) \quad ,$$

the transformation via  $F_A$  gives for an arbitrary  $w_h \in \mathbb{P}^4(\bar{T})$

$$\|A_T^T \nabla w_h\|_T \lesssim \|w_h\|_T \quad .$$

From (2.5) we obtain

$$\|C_T^T \nabla w_h\|_T = \|C_T^T A_T^{-T} \cdot A_T^T \nabla w_h\|_T \leq \|C_T^T A_T^{-T}\|_{\mathbb{R}^{3 \times 3}} \cdot \|A_T^T \nabla w_h\|_T \lesssim \|A_T^T \nabla w_h\|_T$$

and thus the desired result.  $\blacksquare$

### 2.3.3 Bubble functions

Bubble functions and the so-called inverse inequalities related to them play a vital role in our finite element error analysis. Of course different bubble functions can (and have to) be employed for different classes of problems and norms involved. Nevertheless we define here the probably most versatile and commonly used bubble functions. The corresponding inverse inequalities are given and proved.

Other bubble functions that are utilized for an  $L_2$  estimate alone are introduced in the appropriate Section 3.5.

Let  $T \in \mathcal{T}_h$  be an arbitrary tetrahedron, and denote by  $\lambda_{T,1}, \dots, \lambda_{T,4}$  its barycentric coordinates. The *element bubble function*  $b_T \in \mathbb{P}^4(T)$  is defined by

$$b_T := 256 \lambda_{T,1} \cdot \lambda_{T,2} \cdot \lambda_{T,3} \cdot \lambda_{T,4} \quad \text{on } T \quad . \quad (2.19)$$

Let  $E$  be an inner face (triangle) of  $\mathcal{T}_h$ , and let  $T_1$  and  $T_2$  be the two tetrahedra that contain  $E$ . Enumerate the vertices of  $T_1$  and  $T_2$  such that the vertices of  $E$  are numbered first. The *face bubble function*  $b_E$  is then defined by

$$b_E := 27 \lambda_{T_i,1} \cdot \lambda_{T_i,2} \cdot \lambda_{T_i,3} \quad \text{on } T_i, i = 1, 2 \quad . \quad (2.20)$$

This definition is extended in the obvious way for boundary faces  $E \subset \Gamma$ , i.e.  $b_E$  is then defined only on one tetrahedron. For simplicity assume that  $b_T$  and  $b_E$  are extended by zero outside their original domain of definition. Note that  $b_E$  is piecewise cubic on  $\omega_E$ . Both bubble functions satisfy

$$0 \leq b_T(\mathbf{x}), b_E(\mathbf{x}) \leq 1 \quad , \quad \max b_T = \max b_E = 1 \quad .$$

The following examples of the corresponding two-dimensional bubble functions give some impression of their shape. Note that  $b_T \in C^\infty(T)$  and  $b_E \in C^0(\omega_E) \cap H^1(\omega_E)$  but  $b_E \notin H^2(\omega_E)$ .

The norm of the gradient of both bubble functions is bounded as follows.

**Lemma 2.6 (Gradient of both bubble functions)** *For all faces  $E$  of  $T$ , and all  $v_h \in \mathbb{P}^1(T)$  one has*

$$\|\nabla b_T\|_T \sim h_{\min,T}^{-1} \cdot |T|^{1/2} \quad (2.21)$$

$$\|\nabla(v_h \cdot b_E)\|_T \sim h_{\min,T}^{-1} \cdot \|v_h\|_T \quad . \quad (2.22)$$

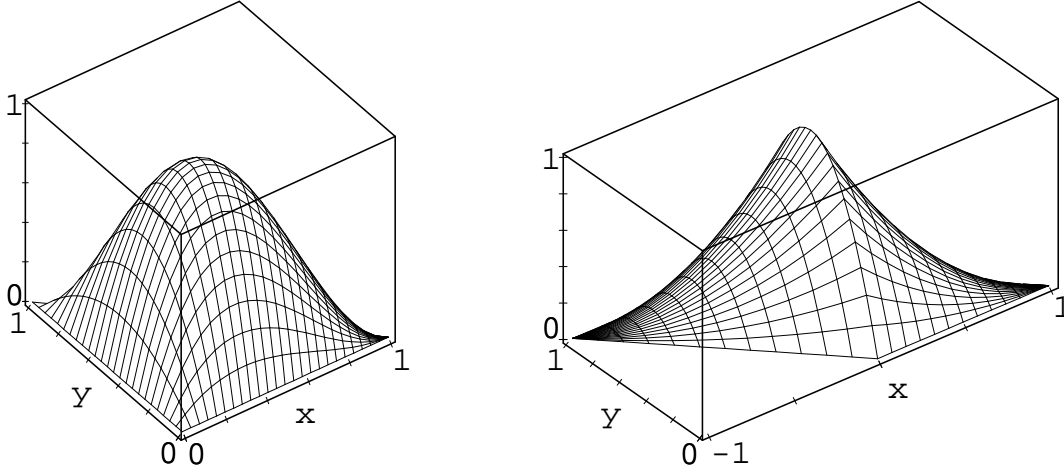


Figure 2.4: Element bubble function  $b_T$  and face bubble function  $b_E$  (in  $\mathbb{R}^2$ )

**Proof:** Standard scaling arguments and the relation  $C_T^T \cdot C_T = H_T^2$  imply

$$\begin{aligned} \|\nabla b_T\|_T^2 &= 6|T| \cdot \|C_T^{-T} \hat{\nabla} \hat{b}_T\|_{\hat{T}}^2 \\ &= 6|T| \cdot \int_{\hat{T}} \hat{\nabla}^T \hat{b}_T \cdot C_T^{-1} C_T^{-T} \cdot \hat{\nabla} \hat{b}_T = 6|T| \cdot \int_{\hat{T}} \hat{\nabla}^T \hat{b}_T \cdot H_T^{-2} \cdot \hat{\nabla} \hat{b}_T \\ &> 6|T| \cdot h_{min,T}^{-2} \cdot \int_{\hat{T}} \left( \frac{\partial \hat{b}_T}{\partial \hat{z}} \right)^2. \end{aligned}$$

The reference tetrahedron  $\hat{T}$  is uniquely determined by its vertices  $(0, 0, 0)^T$ ,  $(1, 0, 0)^T$ ,  $(\hat{x}_2, 1, 0)^T$ , and  $(\hat{x}_3, \hat{y}_3, 1)^T$ , with  $0 < \hat{x}_2 \leq 1/2$ ,  $0 < \hat{x}_3 < 1$  and  $|\hat{y}_3| < 1$  (cf. Section 2.1.2). Define the compact set

$$S_1 := \left\{ (\hat{x}_2, \hat{x}_3, \hat{y}_3) \quad : \quad 0 \leq \hat{x}_2 \leq 1/2, 0 \leq \hat{x}_3 \leq 1, |\hat{y}_3| \leq 1 \right\}$$

which covers all possible tetrahedra  $\hat{T}$  (and some more). Obviously  $\|\partial \hat{b}_T / \partial \hat{z}\|_{\hat{T}}$  varies continuously over  $S_1$  and thus attains its minimum. This is positive since  $\partial \hat{b}_T / \partial \hat{z}$  cannot vanish everywhere on  $\hat{T}$ . Therefore

$$\|\nabla b_T\|_T^2 \gtrsim h_{min,T}^{-2} \cdot |T|.$$

In order to bound the gradient norm from above, utilize (2.18) and (2.11) to obtain

$$\|\nabla b_T\|_T = \|H_T^{-1} C_T^T \nabla b_T\|_T \lesssim h_{min,T}^{-1} \cdot \|C_T^T \nabla b_T\|_T \lesssim h_{min,T}^{-1} \cdot \|b_T\|_T.$$

This proves the equivalence since  $0 \leq b_T \leq 1$ .

For (2.22) proceed analogously. Suppose that  $v_h$  has the nodal values  $v_1 \cdots v_4$  such that

$$v_h = \sum_{i=1}^4 v_i \cdot \lambda_{T,i},$$

with  $\lambda_{T,i}$  being the usual barycentric coordinates of  $T$ . The case  $v_h \equiv 0$  is trivial so assume without loss of generality  $\sum_{i=1}^4 v_i^2 = 1$ . By analogy with above

$$\|\nabla(v_h \cdot b_E)\|_T^2 \geq 6|T| \cdot h_{min,T}^{-2} \cdot \int_{\hat{T}} \left( \sum_{i=1}^4 v_i \cdot \frac{\partial(\lambda_{T,i} \cdot \hat{b}_T)}{\partial \hat{z}} \right)^2.$$

Define now the compact set

$$S_2 := S_1 \times \left\{ (v_1, v_2, v_3, v_4)^T \in \mathbb{R}^4 : \sum_{i=1}^4 v_i^2 = 1 \right\}$$

and observe that the minimum of the integral from above is positive over  $S_2$  which results in  $\|\nabla(v_h \cdot b_E)\|_T^2 \gtrsim h_{min,T}^{-2} \cdot |T|$ . On the other hand one has

$$\|\nabla(v_h \cdot b_E)\|_T \lesssim h_{min,T}^{-1} \cdot \|v_h \cdot b_E\|_T \leq h_{min,T}^{-1} \cdot \|v_h\|_T \sim h_{min,T}^{-1} \cdot |T|^{1/2}$$

by analogy with above. This finishes the proof of (2.22).  $\blacksquare$

Let  $E$  be an interior or boundary face of  $\mathcal{T}_h$ . Then an extension operator  $F_{ext} : \mathbb{P}^1(E) \rightarrow C^0(\omega_E)$  will be required.

**Definition 2.3 (Linear extension operator)** *Let  $T \supset E$  be a tetrahedron with barycentric coordinates such that  $\lambda_{T,1} \cdots \lambda_{T,3}$  are related to the nodes of  $E$ , and  $\lambda_{T,4}$  is related to the remaining node of  $T$ . Let the corresponding nodal values of  $\varphi \in \mathbb{P}^1(E)$  be  $\varphi_1, \varphi_2, \varphi_3$ . Define the linear extension operator on  $T$  by*

$$F_{ext}(\varphi) := \sum_{i=1}^3 \varphi_i \cdot \lambda_{T,i} + \frac{\varphi_1 + \varphi_2 + \varphi_3}{3} \cdot \lambda_{T,4} \in \mathbb{P}^1(T) \quad . \quad (2.23)$$

If  $E$  is an interior face (and thus  $\omega_E = T_1 \cup T_2$ ), define  $F_{ext}$  on each  $T_i$  separately.

The extension operator satisfies  $F_{ext}(\varphi)|_E = \varphi$ . Furthermore,  $F_{ext}(\varphi)$  is a piecewise linear, continuous function over  $\omega_E$ . Later on,  $F_{ext}$  is applied to face residuals  $r_E$ . Only the following cases occur.

$$\begin{array}{lll} E \subset \Omega \setminus \Gamma : & r_E \in \mathbb{P}^0(E) & \implies F_{ext}(r_E) \in \mathbb{P}^0(\omega_E) \\ E \subset \Gamma_N : & r_E \in \mathbb{P}^0(E) & \implies F_{ext}(r_E) \in \mathbb{P}^0(T) \\ E \subset \Gamma_R : & r_E \in \mathbb{P}^1(E) & \implies F_{ext}(r_E) \in \mathbb{P}^1(T) \quad . \end{array}$$

The next relations will later be useful to bound the element residual  $r_T \in \mathbb{P}^0(T)$  and the face residuals  $r_E$ . When  $E$  is an interior face or an Neumann boundary face then  $r_E \in \mathbb{P}^0(E)$ . For  $E$  being a Robin boundary face  $r_E \in \mathbb{P}^1(E)$  holds.

In its original form, these relations are also known as *inverse inequalities*. However, with the help of the previous Lemma one can actually prove not only an inequality but an equivalence relation instead.

**Lemma 2.7 (Equivalences/Inverse inequalities for bubble functions)**

Assume that  $\varphi_T \in \mathbb{P}^0(T)$  and  $\varphi_E \in \mathbb{P}^1(E)$ . Then

$$\|b_T^{1/2} \cdot \varphi_T\|_T \sim \|\varphi_T\|_T \quad (2.24)$$

$$\|\nabla(b_T \cdot \varphi_T)\|_T \sim h_{min,T}^{-1} \cdot \|\varphi_T\|_T \quad (2.25)$$

$$\|b_E^{1/2} \cdot \varphi_E\|_E \sim \|\varphi_E\|_E \quad (2.26)$$

$$\|F_{ext}(\varphi_E) \cdot b_E\|_T \sim h_E^{1/2} \cdot \|\varphi_E\|_E \quad \text{for } E \subset T \quad (2.27)$$

$$\|\nabla(F_{ext}(\varphi_E) \cdot b_E)\|_T \sim h_E^{1/2} \cdot h_{min,T}^{-1} \cdot \|\varphi_E\|_E \quad \text{for } E \subset T \quad (2.28)$$

**Proof:** For all relations the transformation technique is applied.

Obviously  $\|b_{\bar{T}}^{1/2} \cdot \|_{\bar{T}}$  and  $\| \cdot \|_{\bar{T}}$  are equivalent norms on the finite dimensional space  $\mathbb{P}^0(\bar{T})$ . The transformation from the unitary tetrahedron  $\bar{T}$  to the actual tetrahedron  $T$  leads directly to (2.24). Relation (2.26) is proven in exactly the same way. The equivalence (2.25) follows immediately from (2.21) of Lemma 2.6.

The relation

$$\|F_{ext}(\varphi_E) \cdot b_E\|_T^2 = 6|T| \cdot \|F_{ext}(\bar{\varphi}_E) \cdot b_E\|_{\bar{T}}^2 \sim 6|T| \cdot \|\bar{\varphi}_E\|_{\bar{E}}^2 = h_E \cdot \|\varphi_E\|_E^2$$

holds for all  $\varphi_E \in \mathbb{P}^1(E)$  since  $\| \cdot \|_{\bar{E}}$  and  $\|b_{\bar{E}} \cdot F_{ext}(\cdot)\|_{\bar{T}}$  are equivalent norms over a finite dimensional space  $\mathbb{P}^1(\bar{E})$  of polynomials. Thus (2.27) is proven. Utilizing (2.22) and the techniques from above finally yields (2.28). ■

**Remark 2.3** Bubble functions  $b_T$  or  $b_E$  which are transformed via  $F_A^{-1}$  become the corresponding bubble functions on the standard tetrahedron  $\bar{T}$ , respectively, i.e.

$$b_{\bar{T}} = \bar{b}_T := b_T \circ F_A \quad \text{and} \quad b_{\bar{E}} = \bar{b}_E := b_E \circ F_A \quad .$$

A similar relation holds for the transformation  $F_C$ . □

# Chapter 3

## The Poisson equation

### 3.1 Analytical Background

Let  $\Omega \subset \mathbb{R}^d$ ,  $d = 2, 3$ , be an open, bounded, polygonal or polyhedral domain. Assume that its boundary  $\Gamma = \partial\Omega \in C^{0,1} \cap PC^2$  consists of three mutually disjoint portions,

$$\Gamma = \Gamma_D \cup \Gamma_N \cup \Gamma_R \quad ,$$

where  $\Gamma_D, \Gamma_N$  and  $\Gamma_R$  consist of a finite number of mutually disjoint parts of positive measure, or they are empty, respectively. These portions are related to the Dirichlet boundary conditions (also known as b.c. of the first type), to the Neumann boundary conditions (also known as b.c. of the second type), and to the Robin boundary conditions (also known as b.c. of the third type, exchange boundary conditions, or Newton b.c.), respectively.

The *classical formulation* of the Poisson problem then reads as follows: Find a solution  $u$  of

$$\left. \begin{aligned} -\Delta u &= f && \text{in } \Omega \\ u &= 0 && \text{on } \Gamma_D \\ \partial u / \partial n &= g_N && \text{on } \Gamma_N \\ \partial u / \partial n &= \alpha \cdot (g_R - u) && \text{on } \Gamma_R \end{aligned} \right\} \quad (3.1)$$

Under suitable smoothness assumptions on the data (i.e.  $f, g_N, g_R, \alpha$  and  $\Omega$ ) this problem yields a unique *classical solution*  $u \in C^2(\Omega) \cap C^1(\Omega \cup \Gamma_N \cup \Gamma_R) \cap C(\bar{\Omega})$ . This classical formulation, however, often turns out to describe the phenomena of real-world problems insufficiently, see e.g. [22]. For example, there may be no classical solution (because of its high smoothness requirements) but a sensible physical solution (which is less smooth than required).

Therefore, the so-called *variational* or *weak formulation* is more appropriate:

$$\left. \begin{aligned} \text{Find } u \in H_o^1(\Omega) : & \quad a(u, v) = \langle \mathfrak{f}, v \rangle \quad \forall v \in H_o^1(\Omega) \\ \text{with} & \quad a(u, v) := \int_{\Omega} \nabla^T u \cdot \nabla v + \int_{\Gamma_R} \alpha \cdot u \cdot v \\ & \quad \langle \mathfrak{f}, v \rangle := \int_{\Omega} f \cdot v + \int_{\Gamma_N} g_N \cdot v + \int_{\Gamma_R} \alpha \cdot g_R \cdot v \quad . \end{aligned} \right\} \quad (3.2)$$

The corresponding *weak solution*  $u$  is sought in the better suited and larger space  $H_o^1(\Omega)$ , which is the usual Sobolev space of functions from  $H^1(\Omega)$  whose trace on the Dirichlet part  $\Gamma_D$  of the boundary vanishes:

$$H_o^1(\Omega) := \{v \in H^1(\Omega) : u|_{\Gamma_D} = 0\} \quad .$$

The *energy norm* for this variational formulation is defined by

$$\|v\|^2 := a(v, v) = \|\nabla v\|_{\Omega}^2 + \|\alpha^{1/2} \cdot v\|_{\Gamma_R}^2 \quad . \quad (3.3)$$

Note that this coincides with the  $H^1$  seminorm when no  $\Gamma_R$  boundary part is present.

Throughout this paper we demand

- $f \in L_2(\Omega)$ ,
- $g_N \in L_2(\Gamma_N)$ ,  $g_R \in L_2(\Gamma_R)$ ,
- $\alpha \in L_{\infty}(\Gamma_R)$  with  $\alpha(\mathbf{x}) \geq \alpha_0 > 0$  a.e. on  $\Gamma_R$ ,
- $\text{meas}_{d-1}\Gamma_D + \text{meas}_{d-1}\Gamma_R > 0$ .

Then the assumption of the Lax-Milgram Lemma (cf. [22, 26]) are satisfied, namely

- $\mathfrak{f} \in [H_o^1(\Omega)]^* = H^{-1}(\Omega)$
- $a(\cdot, \cdot)$  is elliptic, i.e.  $\exists \mu_1 > 0 : a(v, v) \geq \mu_1 \cdot \|v\|_{H_o^1(\Omega)}^2 \quad \forall v \in H_o^1(\Omega)$
- $a(\cdot, \cdot)$  is bounded, i.e.  $|a(v, w)| \leq \mu_2 \cdot \|v\|_{H_o^1(\Omega)} \cdot \|w\|_{H_o^1(\Omega)} \quad \forall v, w \in H_o^1(\Omega)$  ,

cf. [22], and [63, Section 1.2] for the treatment of  $\Gamma_R$ . The Lax-Milgram Lemma answers the question of the existence and uniqueness of a weak solution to the positive. We remark that for convex domains  $\Omega$  and  $\Gamma \equiv \Gamma_D$  one has  $u \in H^2(\Omega)$ , but otherwise one obtains in general only  $u \in H^{1+\lambda}(\Omega)$  with some  $\lambda \in (0, 1)$ .

The domain  $\Omega$  is supposed to be polyhedral. Assume further a triangulation  $\mathcal{T}_h$  that consists of tetrahedra (3D) or triangles (2D), and which covers  $\Omega$  completely. We assume that each boundary face (or more precisely, its interior) belongs completely to either  $\Gamma_D$ ,  $\Gamma_N$  or  $\Gamma_R$ , respectively. For our error estimation we additionally demand that the function  $\alpha(\mathbf{x}) \in L_{\infty}(\Gamma_R)$  is piecewise constant over each Robin face  $E$ , i.e.

$$\alpha(\mathbf{x}) \Big|_E \in \mathbb{P}^0(E) \quad \forall E \in \Gamma_R \quad .$$

Let now  $V_h \subset C(\bar{\Omega})$  be the space of piecewise linear, continuous functions over  $\mathcal{T}_h$ . Let  $V_{o,h} \subset V_h$  be the subspace of those functions of  $V_h$  with homogeneous Dirichlet boundary conditions, i.e.  $V_{o,h} := V_h \cap H_o^1(\Omega)$ . The *approximate* or *finite element solution*  $u_h$  is obtained via

$$\text{Find } u_h \in V_{o,h} : \quad a(u_h, v_h) = \langle \mathfrak{f}, v_h \rangle \quad \forall v_h \in V_{o,h} \quad . \quad (3.4)$$

The Poisson equation is one of the simplest boundary value problems, and error estimators for *isotropic meshes* are long known and well established. Therefore this problem has been chosen as model problem to investigate how error estimators perform (or have to be modified) if one encounters an *anisotropic solution* or utilizes an *anisotropic mesh*.

Let us specify the framework of this chapter. We try to bound the error  $u - u_h$  in

- the energy norm  $\|v\|_{\Omega}$
- the  $L_2$  norm.

Furthermore different error estimators are investigated. The residual error estimator (Section 3.2) and the local Dirichlet problem error estimator (Section 3.3) estimate the error in the energy norm. An anisotropic Zienkiewicz-Zhu like error estimator which aims at the energy norm is derived as well. An equivalence proof is given for anisotropic meshes of tensor product type. Our results are briefly discussed in Section 3.4. The  $L_2$  error estimator (Section 3.5) is self-explanatory.

Last but not least it should be mentioned here that certain *interpolation* error estimates play a vital role in our analysis. The kind of the interpolation estimate is apparently strongly related to the finite element error estimate to be obtained. Yet the accuracy of our interpolation error estimate depends on how good the anisotropic mesh corresponds to the anisotropic solution (or more precisely, to the error  $u - u_h$ ). This view is supported by heuristic arguments — it seems sensible that the tetrahedra are stretched along that direction where the (directional) derivative of the solution varies little (i.e. when the solution shows an almost lower-dimensional behaviour). Sections 3.2 and 3.5.2, 3.5.3 are devoted to this topic.

## 3.2 Residual error estimator

Residual error estimators have been known for a long time, and they were probably the first estimators ever to be analysed [13]. Since then much work has been devoted to this type of error estimator for various problem classes. Verfürth [56, 58] derived lower bounds of the error.

Residual error estimators which are suitable for anisotropic meshes have been first investigated into by Siebert [50]. There cuboidal and prismatic meshes were considered.

The estimator presented here works on tetrahedral and triangular meshes which are more difficult to deal with. This is mainly due to the fact that rectangular prisms have three natural directions (i.e. along their edges), but tetrahedra do not have these. Hence a more sophisticated analysis as well as modified tools become necessary. Moreover, Siebert's estimator is improved by using a different factor for the norm of the gradient jump (cf. Remark 3.6 on page 51). Thus we can omit an additional condition that Siebert requires.

For clarity, let us recall the methodology of the proofs for the residual error estimator. On *isotropic meshes*, an upper bound of the error is usually derived by means of interpolation error estimates. Bubble functions and their inverse inequalities are utilized to prove a lower bound of the error (cf. [58]).

On *anisotropic meshes*, the situation becomes more complex and technical. The structure of the proofs is similar to [50, 58], for example, although the details vary significantly. To start with, the anisotropic mesh and an anisotropic solution should correspond in some way to yield reliable estimates. This correspondence is important for both, interpolation error estimates and finite element error estimates. Therefore we motivate and define a so-called *matching function*, and discuss it extensively in Section 3.2.1. Note that this matching function has no isotropic counterpart. Anisotropic interpolation estimates play a vital part in order to derive an upper bound of the error. Hence the well-known Clément interpolation operator as well as the proofs of the interpolation estimates are modified to take the anisotropy into account. Section 3.2.2 is devoted to this topic. In Section 3.2.3 the anisotropic error estimator is defined. The main results which are presented in Theorem 3.4 give lower and upper bounds of the error. For a thorough understanding of these results we recommend to consider the previous Sections first, and 3.2.1 in particular. The upper bound of the error is readily obtained via the interpolation estimates. The lower error bound relies on the anisotropic bubble functions and the corresponding inverse inequalities. Although the bubble functions are defined as for isotropic finite elements, the analysis of the inverse inequalities now requires new techniques such as compactness arguments (cf. Section 2.3.3). Finally, Section 3.2.4 again addresses the matching function, its influence on the error estimation and its approximation.

We note here that our residual error estimator is suitable for Dirichlet, Neumann, and Robin boundary conditions, respectively. In particular Robin boundary conditions can now be treated. To our knowledge, this results is new also for isotropic meshes.

### 3.2.1 Relation between anisotropic mesh and anisotropic function

When investigating a residual error estimator for anisotropic meshes, we want to employ the same basic principles as for isotropic meshes. More precisely, a certain kind of interpolation error estimates is to be derived first. With its help, the finite element error is then bounded globally from above.

Proceeding this way, we naturally use different and more technical methods than for isotropic meshes. But even more important, the results of isotropic meshes can not be transferred identically to anisotropic meshes. A certain factor appears now both at the interpolation error estimates and the finite element error estimate. This factor is related to how good the chosen anisotropic mesh corresponds to the anisotropic function under consideration. Basically, the better this correspondence the smaller the factor (but always  $\geq 1$ ), and the better the estimate (in a meaning that is to be specified later on). The importance of an anisotropic mesh that corresponds to an anisotropic function can be described and interpreted in different ways.

- Siebert [50] defines a set  $H_{\mathcal{T}}^{1,2}(\Omega)$  of functions which are adapted to a given mesh. Results are given only for such functions.
- We have decided to introduce the tool of a *matching function*  $m_1(v, \mathcal{T}_h)$ . This matching function measures the correspondence (or the degree of the alignment) of an anisotropic mesh  $\mathcal{T}_h$  with an anisotropic function  $v$ . One advantage of our definition is that we can isolate the influence of the mesh alignment on various estimates by means of  $m_1$ . The worse the alignment, the larger  $m_1$  becomes, and the less accurate the estimates are.

Analytical and numerical examples show that our definition describes the problem reasonably, and that it admits useful results.

- Another point of view is as follows. On isotropic meshes, the lower and upper finite element error bound contain the same terms (up to multiplicative constants which are independent of the mesh and the function). With our definition below, isotropic meshes immediately yield  $m_1(\cdot, \mathcal{T}_h) \sim 1$ .

Apparently the situation for anisotropic meshes is different. For unconditional estimates (i.e. where constants are independent of mesh and function), one has to use different terms for the lower and the upper bound. This implies a different quality of both estimates. If the same terms are desired for both bounds then apparently one has to pay for this. Only meshes that are adapted to the function under consideration are then allowed. In view of our definition below, this corresponds to  $m_1(\cdot, \mathcal{T}_h) \sim 1$ .

Before going into more detail we present the definition of the matching function  $m_1(v, \mathcal{T}_h)$ .

**Definition 3.1 (Matching function  $m_1$ )** *Let  $v \in H^1(\Omega)$  be an arbitrary non-constant function, and  $\mathcal{F}$  be a family of triangulations of  $\Omega$ .*

*Define the matching function  $m_1(\cdot, \cdot) : H^1(\Omega) \times \mathcal{F} \mapsto \mathbb{R}$  by*

$$m_1(v, \mathcal{T}_h) := \left( \sum_{T \in \mathcal{T}_h} h_{min,T}^{-2} \cdot \|C_T^T \nabla v\|_T^2 \right)^{1/2} / \|\nabla v\| \quad . \quad (3.5)$$

Using (2.1), (2.3) and (2.11), (2.12), the compact and the component-wise form of  $m_1(\cdot, \cdot)$  read

$$\begin{aligned} m_1(v, \mathcal{T}_h) &= \|h_{min}^{-1}(\mathbf{x}) \cdot C^T(\mathbf{x}) \nabla v\| / \|\nabla v\| \\ &= \left( \sum_{T \in \mathcal{T}_h} \sum_{i=1}^d \frac{h_{i,T}^2}{h_{min,T}^2} \cdot \|\tilde{D}_i v\|_T^2 \right)^{1/2} / \left( \sum_{T \in \mathcal{T}_h} \sum_{i=1}^d \|\tilde{D}_i v\|_T^2 \right)^{1/2} \quad . \end{aligned}$$

Set  $h_T := h_{1,T}$  and let  $\varrho_T$  be the radius of the largest inscribed sphere of  $T$ . From

$$1 \leq m_1(v, \mathcal{T}_h) \lesssim \max_{T \in \mathcal{T}_h} h_T / \varrho_T \quad .$$

one observes that the matching function is a natural extension of the isotropic case.

The component-wise form allows a heuristic and verbal interpretation. For  $m_1$  to be small, the norm  $\|\tilde{D}_i v\|_T$  should be comparatively ‘small’ when  $h_{i,T} \gg h_{min,T}$ . In other words, along the corresponding direction  $\mathbf{p}_i$  there should be little variation of  $v$ . This coincides with the intuitive proposition that meshes are stretched along that direction where ‘nothing important’ happens (or, in other terms, where the function shows an almost lower-dimensional behaviour). Note, however, that this heuristic interpretation is only local whereas  $m_1$  is defined globally.

Of course a solution can display an anisotropic behaviour along a curved boundary (e.g. a curved interior layer). Such cases can be treated as well. Then an anisotropic mesh should properly match the anisotropic solution, i.e. the stretching direction of the anisotropic elements should follow the curved manifold.

The matching function  $m_1(\cdot, \mathcal{T}_h)$  enters the interpolation error estimates (cf. Theorem 3.3 on page 41) and the upper bound of the finite element error (cf. Theorem 3.4 on page 46) as a multiplicative factor. It is not immediately obvious whether  $m_1$  is introduced via an inadequate analysis, or if it is really necessary. Of course this question is investigated and yields the following results.

- Consider first the interpolation error estimates. The *analytical* example of Remark 3.3 on page 44 proves that  $m_1$  is indispensable. This means that the desired estimate without  $m_1$  does not hold (with constants independent of the mesh and the function). The *numerical* example 2 strongly indicates the same behaviour.
- For the upper bound of the finite element error the numerical example 2 strongly suggests that  $m_1$  is necessary there as well. (An analytical example does not seem feasible here).

These investigations imply that the matching function is not a mere technical assumption. In contrast,  $m_1$  exhibits the very nature of our anisotropic estimates, namely the necessity to provide an adapted mesh in order to obtain efficient estimates. In this sense  $m_1$  must not be viewed as a restriction but as a useful tool.

Finally we briefly note on the compatibility of our anisotropic results with their isotropic counterparts. Of course our estimates can be applied to isotropic meshes too. We know that an anisotropic mesh may or may not be aligned with (or corresponds to) an anisotropic function. In contrast, an isotropic mesh always corresponds to an isotropic function, and the definition from above readily implies a bounded matching function  $m_1$ . Hence  $m_1$  merges with the remaining constants of the estimation and thus remains invisible. In this way we obtain the well-known isotropic results as a special case of our analysis.

Summarizing, the influence of the matching function  $m_1$  is discussed in more detail as follows. The matching function plays a vital role in the interpolation error estimates of Section 3.2.2, and Remark 3.3 on page 44 proves the necessity of  $m_1$  analytically. Next, the upper bound (3.23) of the residual error estimator contains  $m_1$ . The equivalence of this residual error estimator with other estimators introduces  $m_1$  to further upper error bounds. Closely related, Section 3.2.4 investigates how  $m_1(u - u_h, \mathcal{T}_h)$  can be made small (i.e. how to construct meshes that are adapted, and which yield efficient estimates). Furthermore we present a way to estimate  $m_1(u - u_h, \mathcal{T}_h)$ . Finally the numerical experiments (and example 2 in particular) give numerical evidence of the necessity of  $m_1$ .

### 3.2.2 Anisotropic interpolation estimates

Interpolation estimates are a major tool in the error analysis performed here. Since the interpolation has to act on functions  $v \in H^1(\Omega)$  we cannot use the usual Lagrange interpolation. Therefore the interpolation operator introduced in this Section follows the lines of Clément [28] instead. All estimates, however, are derived for the use on anisotropic meshes. We note that other interpolation operators are possible, cf. [49, 5].

A local  $L_2$  projection, along with approximation estimates, will be presented first. Then the Clément interpolation operator is constructed. Finally it is modified in such a way that homogeneous Dirichlet boundary conditions will be preserved and that Robin boundary conditions are taken account of. The main interpolation error estimates are given in Theorem 3.3.

#### The local $L_2$ projection

Consider a node  $a_j$ . A so-called *macro element*  $M_j$  (or patch) of this node  $a_j$  is defined by

$$M_j := \bigcup_{T: a_j \in \mathcal{N}_T} T \quad ,$$

i.e.  $M_j$  consists of all tetrahedra containing  $a_j$ . For simplicity the subscript  $j$  will be omitted in the next Lemma and proof.

**Lemma 3.1** *Let  $a$  be a node of  $\mathcal{T}_h$  and  $M$  the corresponding macro element. Let the local  $L_2$  projection  $P : H^1(M) \mapsto \mathbb{P}^0(M)$  be defined by*

$$\int_M (v - Pv) \cdot \varphi = 0 \quad \forall \varphi \in \mathbb{P}^0(M) \quad .$$

*Then the relations*

$$\|v - Pv\|_M \leq \|v\|_M \quad (3.6)$$

$$\|v - Pv\|_M \lesssim \|C^T(\mathbf{x})\nabla v\|_M \quad (3.7)$$

$$\|C^T(\mathbf{x})\nabla(v - Pv)\|_M = \|C^T(\mathbf{x})\nabla v\|_M \quad (3.8)$$

*hold. The component-wise form of (3.8) is*

$$\|\tilde{D}_i(v - Pv)\|_M \lesssim h_{i,M}^{-1} \cdot \sum_{k=1}^d h_{k,M} \|\tilde{D}_k v\|_M \quad i = 1 \dots d \quad ,$$

*with  $h_{i,M}$  explained in Remark 2.2 on page 22.*

**Proof:** The first inequality is readily obtained using the projection orthogonality:

$$\|v - Pv\|_M^2 = \int_M (v - Pv)(v - Pv) = \int_M (v - Pv) \cdot v \leq \|v - Pv\|_M \cdot \|v\|_M$$

since  $Pv \in \mathbb{P}^0(M)$ .

The second inequality requires a closer investigation. A continuous mapping  $F_B$  that maps a reference domain  $\check{M} \in \mathcal{M}$  onto the macro element  $M$  will play an important role in the proof. Furthermore, the set  $\mathcal{M}$  of reference domains shall be finite. For a start we will construct the reference domains  $\check{M}$ .

Assume that the macro element  $M$  is the union of  $K$  tetrahedra  $T_1 \dots T_K$ . Let the nodes of  $M$  be  $a_1 \dots a_L$  (apart from node  $a$ ), where  $L$  is bounded because of the mesh requirements. Two macro elements  $M$  and  $M'$  are said to belong to the same *class* iff

- they consist of the same number of tetrahedra, i.e.  $K = K'$ ,
- the tetrahedra and the nodes can be numbered such that for all  $i = 1 \dots K$  the following holds: If the tetrahedron  $T_i$  has the nodes  $a, a_{j_1}, a_{j_2}, a_{j_3}$  then the tetrahedron  $T'_i$  has the nodes  $a', a'_{j_1}, a'_{j_2}, a'_{j_3}$ .

This condition implies that the triangulations of both macro elements are topologically equivalent. The number of such topologies is bounded since  $K$  is bounded. Therefore the number of classes of macro elements is bounded as well. For a fixed class an arbitrary macro element will be chosen whose node  $a$  coincides with the coordinate origin. This macro element is said to be the reference domain of this class. All reference domains form the (finite) set  $\mathcal{M}$ . Note that a condition on the size of the reference domains is not necessary.

Let now  $M$  be an arbitrary macro element and  $\check{M}$  be the corresponding reference domain. Because of the construction of the reference domain there exists a continuous, piecewise linear mapping  $F_B$  that satisfies

$$\begin{aligned} F_B & : \check{M} \mapsto M \\ F_B = F_i & : \mathbf{x}(\boldsymbol{\mu}) = B_i \boldsymbol{\mu} + \mathbf{a} \quad \text{on } \check{T}_i, \quad B_i \in \mathbb{R}^{d \times d}, \mathbf{a} \in \mathbb{R}^d \\ \text{with } F_i & : \check{T}_i \mapsto T_i \quad \text{affine linear, } i = 1 \dots K \quad . \end{aligned}$$

Temporarily  $\check{T}_i$  and  $T_i$  shall denote the  $i$ -th tetrahedron of  $\check{M}$  and  $M$ , respectively, and  $\mathbf{a}$  denotes the vector corresponding to node  $a$ . Variables that are related to the reference domain will be denoted by a  $\check{\phantom{x}}$  (small check).

The Poincaré inequality holds for the domain  $\check{M}$ . Its inequality constant can be chosen independently of  $\check{M}$  since the number of reference domains  $\check{M} \in \mathcal{M}$  is bounded. Thus for  $\check{u} \in H^1(\check{M})$

$$\int_{\check{M}} |\check{u}|^2 \lesssim \left| \int_{\check{M}} \check{u} \right|^2 + \int_{\check{M}} |\check{\nabla} \check{u}|^2 \quad .$$

For a function  $v \in H^1(M)$  define an averaging operator  $I : H^1(M) \rightarrow \mathbb{P}^0(M)$  by

$$Iv := |\check{M}|^{-1} \cdot \sum_{i=1}^K \int_{T_i} v \cdot |\det B_i|^{-1} = \text{const.}$$

Set  $\check{v} := v \circ F_B \in H^1(\check{M})$ . The definition of  $I$  gives

$$\int_{\check{M}} \check{I}v = |\check{M}| \cdot Iv = \sum_{i=1}^K \int_{T_i} v \cdot |\det B_i|^{-1} = \int_{\check{M}} \check{v}$$

and

$$\check{\nabla}(\check{I}v) = 0 \quad .$$

Inserting now  $\check{u} := \check{v} - \check{I}v$  in the Poincaré inequality results in

$$\int_{\check{M}} |\check{v} - \check{I}v|^2 \lesssim \int_{\check{M}} |\check{\nabla} \check{v}|^2 \quad .$$

Obviously  $\underline{c} \leq |\check{T}_i| \leq \bar{c}$  since the number of reference domains  $\check{M}$  is bounded, and each  $\check{M}$  consists only of a bounded number of tetrahedra  $\check{T}_i$ . Hence

$$|\det B_i| = |T_i|/|\check{T}_i| \sim h_{1,T_i} h_{2,T_i} h_{3,T_i} \quad .$$

Since the  $h_{i,T}$  cannot change rapidly one obtains

$$|\det B_i| \sim |\det B_j| \quad \forall T_i, T_j \subset M \quad .$$

Applying the transformation  $F_B : \check{M} \mapsto M$  gives

$$\begin{aligned} \int_M (v - Iv)^2 &= \sum_{i=1}^K \int_{T_i} (v - Iv)^2 = \sum_{i=1}^K \int_{\check{T}_i} (\check{v} - \check{I}\check{v})^2 \cdot |\det B_i| \\ &\lesssim \max_{i=1\dots K} \{|\det B_i|\} \cdot \int_{\check{M}} (\check{v} - \check{I}\check{v})^2 \\ &\lesssim \max_{i=1\dots K} \{|\det B_i|\} \cdot \int_{\check{M}} |\check{\nabla} \check{v}|^2 \\ &\lesssim \sum_{i=1}^K \int_{\check{T}_i} |\check{\nabla} \check{v}|^2 \cdot |\det B_i| \\ &= \sum_{i=1}^K \int_{T_i} |B_i^T \nabla v|^2 = \sum_{i=1}^K \int_{T_i} |B_i^T C_{T_i}^{-T} \cdot C_{T_i}^T \nabla v|^2 \\ &\leq \sum_{i=1}^K \|B_i^T C_{T_i}^{-T}\|_{\mathbb{R}^{3 \times 3}}^2 \cdot \int_{T_i} |C_{T_i}^T \nabla v|^2 \quad . \end{aligned}$$

Lemma 2.1 on page 18 is now utilized to bound the norm of  $B_i^T C_{T_i}^{-T}$ . Let  $T_i - \mathbf{a}$  be the tetrahedron  $T_i$  shifted by  $-\mathbf{a}$ . By definition the mappings  $B_i$  and  $C_{T_i}^{-1}$  act as follows:

$$\begin{array}{ccc} \check{T}_i & \xrightarrow{B_i} & (T_i - \mathbf{a}) \\ & & \xrightarrow{C_{T_i}^{-1}} \hat{T}_i, \\ & & \xrightarrow{C_{T_i}^{-1} B_i} \hat{T}_i \quad . \end{array}$$

The number of tetrahedra  $\check{T}_j \subset \check{M}$  is bounded. Hence the diameters of the inscribed spheres of all tetrahedra  $\check{T}_i$  can be bounded uniformly from below, i.e.  $\varrho(\check{T}_i) \gtrsim 1$ . The longest edge of  $\hat{T}_i$  is bounded from above by  $\sqrt{6}$  (see definition of the mapping  $C_{T_i}$ ). Lemma 2.1 yields readily

$$\|B_i^T C_{T_i}^{-T}\|_{\mathbb{R}^{3 \times 3}} = \|C_{T_i}^{-1} B_i\|_{\mathbb{R}^{3 \times 3}} \leq d(\hat{T}_i) / \varrho(\check{T}_i) \lesssim 1$$

and further

$$\int_M (v - Iv)^2 \lesssim \sum_{i=1}^K \int_{T_i} |C_{T_i}^T \nabla v|^2 = \|C^T(\mathbf{x}) \nabla v\|_M^2 \quad .$$

The orthogonality property of the projection and  $Pv - Iv \in V_h$  then imply

$$\begin{aligned} \|v - Pv\|_M^2 &= \int_M (v - Pv)(v - Pv) = \int_M (v - Pv)(v - Iv) \\ &\leq \|v - Pv\|_M \cdot \|v - Iv\|_M \\ \text{and} \quad \|v - Pv\|_M &\leq \|v - Iv\|_M \lesssim \|C^T(\mathbf{x}) \nabla v\|_M \end{aligned}$$

finishing the second part of the proof. Recall that  $C(\mathbf{x})$  is the global matrix function defined in (2.3).

The last relation is obvious since  $Pv \in \mathbb{P}^0(M)$ . ■

**Remark 3.1** In the case  $\Omega \subset \mathbb{R}^2$  the reference domains can be chosen easily. Assume that the macro element consists of  $K$  triangles. When  $a$  is a inner node then choose  $\check{M}$  to be a regular  $K$ -polygon with the midpoint in the coordinate origin. Figure 3.1 may serve for visualization.

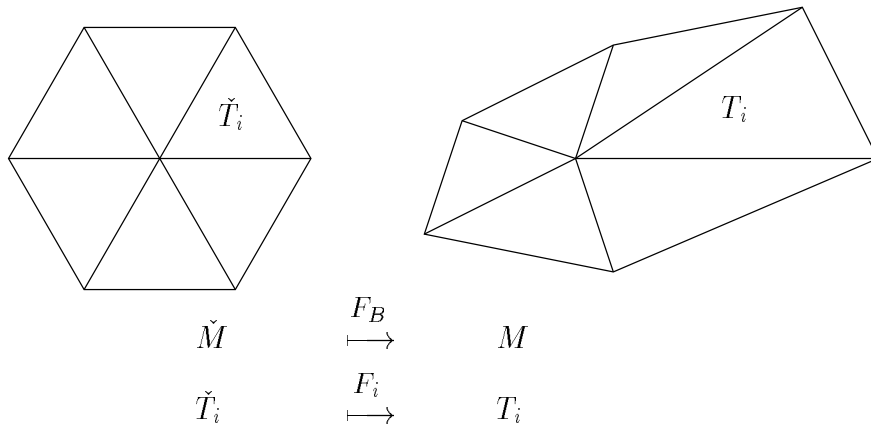


Figure 3.1: Continuous, piecewise affine linear mapping  $F_B$  for  $\Omega \subset \mathbb{R}^2$

If  $a$  is a boundary node then let  $\check{M}$  be the union of those  $K$  (congruent) triangles whose vertices have the polar coordinates  $(0, 0)$ ,  $(1, (i-1)\pi/2K)$  and  $(1, i\pi/2K)$ ,  $i = 1 \dots K$ .

The regular polygon is chosen here only for the convenience of the description, but otherwise completely arbitrary. Any other reference domain could serve the same purpose.

The case  $\Omega \subset \mathbb{R}^3$  is more difficult since generally no regular polyhedra exist. Thus we had to utilize the more technical definition of the reference domains here. □

### The $H^1$ interpolation operator

Now the Clément interpolation operator is constructed. Let  $P_j$  be the aforementioned local  $L_2$  projection over the macro element  $M_j$  of a node  $a_j$ . The interpolation operator  $R$  is defined by

$$Rv := \sum_{a_j \in \mathcal{N}} (P_j v)(a_j) \cdot \varphi_j$$

with  $\varphi_j$  being the (piecewise affine linear) basis function related to node  $a_j$ . Then the following Lemma holds.

**Lemma 3.2** For all  $v \in H^1(\Omega)$  the interpolation operator  $R : H^1(\Omega) \mapsto V_h$  satisfies

$$\begin{aligned} \|v - Rv\| &\lesssim \|v\| \\ \|h_{min}^{-1}(\mathbf{x}) \cdot (v - Rv)\| &\equiv \left( \sum_{T \in \mathcal{T}_h} h_{min,T}^{-2} \|v - Rv\|_T^2 \right)^{1/2} \lesssim m_1(v, \mathcal{T}_h) \cdot \|\nabla v\| \end{aligned}$$

$$\|h_{\min}^{-1}(\mathbf{x}) \cdot C^T(\mathbf{x}) \nabla(v - Rv)\| \equiv \left( \sum_{T \in \mathcal{T}_h} h_{\min, T}^{-2} \|C_T^T \nabla(v - Rv)\|_T^2 \right)^{1/2} \lesssim m_1(v, \mathcal{T}_h) \cdot \|\nabla v\|.$$

The component-wise form of the last inequality is

$$\left\| \frac{h_i(\mathbf{x})}{h_{\min}(\mathbf{x})} \tilde{D}_i(v - Rv) \right\| = \left( \sum_{T \in \mathcal{T}_h} \frac{h_{i, T}^2}{h_{\min, T}^2} \|\tilde{D}_i(v - Rv)\|_T^2 \right)^{1/2} \lesssim m_1(v, \mathcal{T}_h) \cdot \|\nabla v\|$$

for all  $i = 1 \dots d$ .

**Proof:** Let  $T$  be an arbitrary tetrahedron, and  $a_k$  an arbitrary but fixed node thereof. Then  $R$  can be represented over  $T$  as

$$Rv \Big|_T = \sum_{a_j \in \mathcal{N}_T} (P_j v)(a_j) \cdot \varphi_j \Big|_T = P_k v \Big|_T + \sum_{a_j \in \mathcal{N}_T} (P_j v - P_k v)(a_j) \cdot \varphi_j \Big|_T,$$

since 
$$P_k v \Big|_T = \sum_{a_j \in \mathcal{N}_T} (P_k v)(a_j) \cdot \varphi_j \Big|_T .$$

The inverse inequality (2.16) and the triangle inequality imply

$$\begin{aligned} |(P_j v - P_k v)(a_j)| &\leq \|P_j v - P_k v\|_{\infty, T} \\ &\lesssim |T|^{-1/2} \cdot \|P_j v - P_k v\|_T \\ &\leq |T|^{-1/2} \cdot \left( \|v - P_j v\|_T + \|v - P_k v\|_T \right) . \end{aligned}$$

The bound  $\|\varphi_j\|_T \leq |T|^{1/2}$  gives

$$\begin{aligned} |(P_j v - P_k v)(a_j) \cdot \varphi_j|_T &= |(P_j v - P_k v)(a_j)| \cdot \|\varphi_j\|_T \\ &\lesssim \|v - P_j v\|_T + \|v - P_k v\|_T . \end{aligned}$$

Applying this inequality to the representation of  $R$  leads to

$$\begin{aligned} \|v - Rv\|_T &\leq \|v - P_k v\|_T + \left\| \sum_{a_j \in \mathcal{N}_T} (P_j v - P_k v)(a_j) \cdot \varphi_j \right\|_T \\ &\lesssim \|v - P_k v\|_T + \sum_{a_j \in \mathcal{N}_T} \left( \|v - P_j v\|_T + \|v - P_k v\|_T \right) \\ &\lesssim \sum_{a_j \in \mathcal{N}_T} \|v - P_j v\|_T \leq \sum_{a_j \in \mathcal{N}_T} \|v - P_j v\|_{M_j} . \end{aligned}$$

The local approximation inequality (3.6) results in

$$\|v - Rv\|_T \stackrel{(3.6)}{\lesssim} \sum_{a_j \in \mathcal{N}_T} \|v\|_{M_j} \sim \|v\|_{M(T)}$$

with  $M(T) := \bigcup_{a_j \in \mathcal{N}_T} M_j = \bigcup_{T' \cap T \neq \emptyset} T'$ . This holds since every tetrahedron  $T'$  is contained in at most four macro elements  $M_j$ . Then

$$\|v - Rv\|^2 = \sum_{T \in \mathcal{T}_h} \|v - Rv\|_T^2 \lesssim \sum_{T \in \mathcal{T}_h} \|v\|_{M(T)}^2 \sim \|v\|^2$$

because every tetrahedron  $T$  appears only a bounded number of times in the sum. Hence the first inequality is obtained.

For the second inequality we apply (3.7) instead of (3.6) and obtain

$$\|v - Rv\|_T \stackrel{(3.7)}{\lesssim} \sum_{a_j \in \mathcal{N}_T} \|C^T(\mathbf{x}) \nabla v\|_{M_j} \sim \|C^T(\mathbf{x}) \nabla v\|_{M(T)} \quad .$$

Similarly this yields

$$\begin{aligned} \|h_{min}^{-1}(\mathbf{x}) \cdot (v - Rv)\|^2 &= \sum_{T \in \mathcal{T}_h} h_{min,T}^{-2} \|v - Rv\|_T^2 \lesssim \sum_{T \in \mathcal{T}_h} h_{min,T}^{-2} \|C^T(\mathbf{x}) \nabla v\|_{M(T)}^2 \\ &\sim \sum_{T \in \mathcal{T}_h} h_{min,T}^{-2} \cdot \|C_T^T \nabla v\|_T^2 \end{aligned}$$

since  $h_{min,T'}$  does not change rapidly for  $T' \subset M(T)$ .

In order to bound this last sum we utilize the matching function  $m_1(\cdot, \cdot)$  from (3.5) giving

$$\|h_{min}^{-1}(\mathbf{x}) \cdot (v - Rv)\|^2 \lesssim \sum_{T \in \mathcal{T}_h} h_{min,T}^{-2} \cdot \|C_T^T \nabla v\|_T^2 \lesssim m_1(v, \mathcal{T}_h)^2 \cdot \|\nabla v\|^2 \quad .$$

Thus the second result is proven.

The last part is derived similarly, and for this reason only major inequalities are given here. As before

$$C_T^T \nabla Rv \Big|_T = C_T^T \nabla P_k v \Big|_T + \sum_{a_j \in \mathcal{N}_T} (P_j v - P_k v)(a_j) \cdot C_T^T \nabla \varphi_j \Big|_T \quad .$$

Recalling the inverse inequality (2.18)

$$\|C_T^T \nabla \varphi_j\|_T \lesssim \|\varphi_j\|_T \sim |T|^{1/2}$$

leads to

$$\begin{aligned} \|(P_j v - P_k v)(a_j) \cdot C_T^T \nabla \varphi_j\|_T &= |(P_j v - P_k v)(a_j)| \cdot \|C_T^T \nabla \varphi_j\|_T \\ &\lesssim \left( \|v - P_j v\|_T + \|v - P_k v\|_T \right) \end{aligned}$$

analogously as before. Similarly to the second part one obtains

$$\begin{aligned} \|C_T^T \nabla (v - Rv)\|_T &\leq \|C_T^T \nabla (v - P_k v)\|_T + \left\| \sum_{a_j \in \mathcal{N}_T} (P_j v - P_k v)(a_j) \cdot C_T^T \nabla \varphi_j \right\|_T \\ &\stackrel{(3.8)}{\lesssim} \|C^T(\mathbf{x}) \nabla v\|_{M_k} + \|C^T(\mathbf{x}) \nabla v\|_{M(T)} \\ &\lesssim \|C^T(\mathbf{x}) \nabla v\|_{M(T)} \end{aligned}$$

and hence

$$\sum_{T \in \mathcal{T}_h} h_{min,T}^{-2} \|C_T^T \nabla (v - Rv)\|_T^2 \lesssim \sum_{T \in \mathcal{T}_h} h_{min,T}^{-2} \|C^T(\mathbf{x}) \nabla v\|_{M(T)}^2 \sim \sum_{T \in \mathcal{T}_h} h_{min,T}^{-2} \|C_T^T \nabla v\|_T^2.$$

Utilizing the matching function  $m_1(\cdot, \cdot)$  results immediately in

$$\|h_{min}^{-1}(\mathbf{x}) \cdot C^T(\mathbf{x}) \nabla (v - Rv)\| \lesssim m_1(v, \mathcal{T}_h) \cdot \|\nabla v\| \quad .$$

■

### The $H_o^1$ interpolation operator

The interpolation operator  $R$  introduced above has the disadvantage that it does not preserve homogeneous Dirichlet boundary conditions. This is remedied by definition 3.2 below. Moreover, the interpolation operator has to be modified on parts of the Robin boundary  $\Gamma_R$  to treat Robin boundary terms correctly. To this end, introduce a subset of the Robin boundary nodes  $\mathcal{N}_R$  by

$$\mathcal{N}_{R,\alpha} := \left\{ a \in \mathcal{N}_R : \exists E \subset \Gamma_R \text{ with } a \in \mathcal{N}_E \text{ and } \frac{h_E}{h_{\min,E}^2} < \alpha \right\} . \quad (3.9)$$

**Definition 3.2 (Clément interpolation operator)** *The Clément interpolation operator  $R_o : H_o^1(\Omega) \mapsto V_{o,h}$  is defined by*

$$R_o v := \sum_{a_j \in \mathcal{N} \setminus (\mathcal{N}_D \cup \mathcal{N}_{R,\alpha})} (P_j v)(a_j) \cdot \varphi_j . \quad (3.10)$$

This means that the interpolated nodal value 0 is imposed on the Dirichlet boundary and partly on the Robin boundary. Nodes which are surrounded by neither boundary are not affected. The following anisotropic estimates describe the main interpolation results.

**Theorem 3.3** *For all  $v \in H_o^1(\Omega)$  the interpolation operator  $R_o : H_o^1(\Omega) \mapsto V_{o,h}$  satisfies*

$$\|v - R_o v\| \lesssim \|v\| \quad (3.11)$$

$$\|h_{\min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\| \lesssim m_1(v, \mathcal{T}_h) \cdot \|v\| \quad (3.12)$$

$$\|h_{\min}^{-1}(\mathbf{x}) \cdot C^T(\mathbf{x}) \nabla (v - R_o v)\| \lesssim m_1(v, \mathcal{T}_h) \cdot \|v\| \quad (3.13)$$

$$\sum_{E \subset \Omega \setminus \Gamma} \frac{h_E}{h_{\min,E}^2} \cdot \|v - R_o v\|_E^2 \lesssim m_1(v, \mathcal{T}_h)^2 \cdot \|v\|^2 \quad (3.14)$$

$$\sum_{E \subset \Gamma_N} \frac{h_E}{h_{\min,E}^2} \cdot \|v - R_o v\|_E^2 \lesssim m_1(v, \mathcal{T}_h)^2 \cdot \|v\|^2 \quad (3.15)$$

$$\sum_{E \subset \Gamma_R} \max \left\{ \frac{h_E}{h_{\min,E}^2}, \alpha \right\} \cdot \|v - R_o v\|_E^2 \lesssim m_1(v, \mathcal{T}_h)^2 \cdot \|v\|^2 . \quad (3.16)$$

**Proof:** The definition of the interpolation operator  $R_o$  implies

$$R_o v = Rv - \sum_{a_j \in \mathcal{N}_D \cup \mathcal{N}_{R,\alpha}} (P_j v)(a_j) \cdot \varphi_j .$$

Since we want to utilize the previous Lemma it is sufficient to bound terms of the form  $\|(P_j v)(a_j) \cdot \varphi_j\|_T$  and  $\|(P_j v)(a_j) \cdot C_T^T \nabla \varphi_j\|_T$  for boundary nodes  $a_j \in \mathcal{N}_D \cup \mathcal{N}_{R,\alpha}$ .

Thus let  $a_j \in \mathcal{N}_D \cup \mathcal{N}_{R,\alpha}$  be fixed. Let  $T \subset M_j$  be an arbitrary (but fixed) tetrahedron with a boundary face  $E$  containing  $a_j$ . The equivalence relation (2.16) yields

$$|(P_j v)(a_j)| = \|P_j v\|_{\infty,T} \sim |T|^{-1/2} \cdot \|P_j v\|_T \leq |T|^{-1/2} \cdot (\|v\|_T + \|v - P_j v\|_T) .$$

The estimate of the local  $L_2$  projection, and  $\|\varphi_j\|_T \leq |T|^{1/2}$  give

$$\|(P_j v)(a_j) \cdot \varphi_j\|_T \lesssim \|v\|_{M_j} \quad \forall T \subset M_j \quad .$$

Applying this estimate and the arithmetic-quadratic mean inequality results in

$$\begin{aligned} \left\| \sum_{a_j \in \mathcal{N}_D \cup \mathcal{N}_{R,\alpha}} (P_j v)(a_j) \cdot \varphi_j \right\|^2 &= \sum_{T \cap \Gamma \neq \emptyset} \left\| \sum_{\substack{a_j \in \mathcal{N}_T \\ a_j \in \mathcal{N}_D \cup \mathcal{N}_{R,\alpha}}} (P_j v)(a_j) \cdot \varphi_j \right\|_T^2 \\ &\leq \sum_{T \cap \Gamma \neq \emptyset} 4 \cdot \sum_{\substack{a_j \in \mathcal{N}_T \\ a_j \in \mathcal{N}_D \cup \mathcal{N}_{R,\alpha}}} \left\| (P_j v)(a_j) \cdot \varphi_j \right\|_T^2 \\ &\lesssim \sum_{a_j \in \mathcal{N}_D \cup \mathcal{N}_{R,\alpha}} \|v\|_{M_j}^2 \lesssim \|v\|^2 \end{aligned}$$

following the same arguments as in the previous proof. Finally the first inequality is obtained via

$$\|v - R_o v\| \leq \|v - Rv\| + \left\| \sum_{a_j \in \mathcal{N}_D \cup \mathcal{N}_{R,\alpha}} (P_j v)(a_j) \cdot \varphi_j \right\| \lesssim \|v\| \quad .$$

The second inequality is derived similarly. Let  $a_j \in \mathcal{N}_D \cup \mathcal{N}_{R,\alpha}$  be an arbitrary node. Starting point is a face  $E$  which contains the node  $a_j$ . If  $a_j \in \mathcal{N}_{R,\alpha}$  then there exists a face  $E$  such that  $E \subset \Gamma_R$  and  $h_E/h_{\min,E}^2 < \alpha$ . Otherwise  $a_j \in \mathcal{N}_D$  where a face  $E \subset \Gamma_D$  is chosen. Consider the unique tetrahedron  $T \supset E$ . Equivalence relation (2.17), the triangle inequality and the trace inequality (2.14) yield

$$\begin{aligned} |(P_j v)(a_j)| &= \|P_j v\|_{\infty,E} \stackrel{(2.17)}{\sim} |E|^{-1/2} \|P_j v\|_E \\ &\leq |E|^{-1/2} \cdot (\|v\|_E + \|v - P_j v\|_E) \\ &\lesssim |E|^{-1/2} \cdot \|v\|_E + |T|^{-1/2} \cdot (\|v - P_j v\|_T + \|C_T^T \nabla(v - P_j v)\|_T) \quad . \end{aligned}$$

Recalling estimates (3.7) and (3.8) of the local  $L_2$  projection leads to

$$|(P_j v)(a_j)| \lesssim |E|^{-1/2} \cdot \|v\|_E + |T|^{-1/2} \cdot \|C^T(\mathbf{x}) \nabla v\|_{M_j} \quad .$$

If  $E \subset \Gamma_D$  then  $v = 0$  on  $E$ . Together with  $\|\varphi_j\|_{T'} \leq |T'|^{1/2}$  this yields

$$\|(P_j v)(a_j) \cdot \varphi_j\|_{T'} \lesssim \|C^T(\mathbf{x}) \nabla v\|_{M_j} \quad \forall T' \subset M_j \quad .$$

Otherwise  $E \subset \Gamma_R$  has been chosen such that  $|T|/|E| \sim h_E < \alpha \cdot h_{\min,E}^2$  which implies

$$\|(P_j v)(a_j) \cdot \varphi_j\|_{T'} \lesssim h_{\min,E} \cdot \|\alpha^{1/2} \cdot v\|_E + \|C^T(\mathbf{x}) \nabla v\|_{M_j} \quad \forall T' \subset M_j \quad .$$

Applying these estimates and utilizing the matching function  $m_1(\cdot, \cdot)$  results in

$$\begin{aligned} \left\| h_{\min}^{-1}(\mathbf{x}) \sum_{a_j \in \mathcal{N}_D \cup \mathcal{N}_{R,\alpha}} (P_j v)(a_j) \cdot \varphi_j \right\|^2 &\leq \sum_{T \cap \Gamma \neq \emptyset} 4 \cdot \sum_{\substack{a_j \in \mathcal{N}_T \\ a_j \in \mathcal{N}_D \cup \mathcal{N}_{R,\alpha}}} h_{\min,T}^{-2} \|(P_j v)(a_j) \cdot \varphi_j\|_T^2 \\ &\lesssim \sum_{a_j \in \mathcal{N}_{R,\alpha}} \left( \|\alpha^{1/2} v\|_E^2 + h_{\min,E}^{-2} \cdot \|C^T(\mathbf{x}) \nabla v\|_{M_j}^2 \right) + \sum_{a_j \in \mathcal{N}_D \setminus \mathcal{N}_{R,\alpha}} h_{\min,E}^{-2} \cdot \|C^T(\mathbf{x}) \nabla v\|_{M_j}^2 \\ &\lesssim \|\alpha^{1/2} \cdot v\|_{\Gamma_R}^2 + \sum_{T \in \mathcal{T}_h} h_{\min,T}^{-2} \cdot \|C_T^T \nabla v\|_T^2 \\ (3.5) \quad &\lesssim \|\alpha^{1/2} \cdot v\|_{\Gamma_R}^2 + m_1(v, \mathcal{T}_h)^2 \cdot \|\nabla v\|^2 \leq m_1(v, \mathcal{T}_h)^2 \cdot \|v\|^2 \end{aligned}$$

and thus the second inequality is proven.

In order to derive the third inequality of the Theorem we proceed analogously. The only difference is that now  $\varphi_j$  is replaced by  $C_T^T \nabla \varphi_j$ . From inequality (2.18)

$$\|C_T^T \nabla \varphi_j\|_T \lesssim \|\varphi_j\|_T \leq |T|^{1/2}$$

one obtains

$$\begin{aligned} \|(P_j v)(a_j) \cdot C_T^T \nabla \varphi_j\|_T &\lesssim \|C^T(\mathbf{x}) \nabla v\|_{M_j} && \text{for } a_j \in \mathcal{N}_D \\ \|(P_j v)(a_j) \cdot C_T^T \nabla \varphi_j\|_T &\lesssim \|C^T(\mathbf{x}) \nabla v\|_{M_j} + h_{\min, T} \cdot \|\alpha^{1/2} \cdot v\|_E && \text{for } a_j \in \mathcal{N}_{R, \alpha} \end{aligned}$$

for a boundary tetrahedron  $T$ . The remainder of the proof is similar to the lines above and the previous proof and thus it will be omitted here.

The last three inequalities of the Theorem estimate the interpolation error over certain faces. For an interior face  $E \subset \Omega \setminus \Gamma$  or a Neumann boundary face  $E \subset \Gamma_N$  the trace inequality (2.14) is applied to  $\|v - R_o v\|_E$  giving

$$\frac{h_E}{h_{\min, T}^2} \|v - R_o v\|_E^2 \lesssim h_{\min, E}^{-2} (\|v - R_o v\|_T^2 + \|C_T^T \nabla(v - R_o v)\|_T^2)$$

with  $T \supset E$ . In conjunction with the previous interpolation estimates (3.12) and (3.13) the inequalities

$$\begin{aligned} \sum_{E \subset \Omega \setminus \Gamma} \frac{h_E}{h_{\min, E}^2} \cdot \|v - R_o v\|_E^2 &\lesssim \sum_{T \in \mathcal{T}_h} h_{\min, T}^{-2} \cdot (\|v - R_o v\|_T^2 + \|C_T^T \nabla(v - R_o v)\|_T^2) \\ &\lesssim m_1(v, \mathcal{T}_h)^2 \cdot \|v\|^2 \\ \text{and } \sum_{E \subset \Gamma_N} \frac{h_E}{h_{\min, E}^2} \cdot \|v - R_o v\|_E^2 &\lesssim \sum_{T \cap \Gamma_N \neq \emptyset} h_{\min, T}^{-2} \cdot (\|v - R_o v\|_T^2 + \|C_T^T \nabla(v - R_o v)\|_T^2) \\ &\lesssim m_1(v, \mathcal{T}_h)^2 \cdot \|v\|^2 \end{aligned}$$

are derived which prove the fourth and fifth assertion.

Finally, consider  $E \subset \Gamma_R$  and recall that  $R_o v = 0$  when  $h_E/h_{\min, E}^2 < \alpha$ . The application of the trace inequality results in

$$\begin{aligned} \sum_{E \subset \Gamma_R} \max \left\{ \frac{h_E}{h_{\min, E}^2}, \alpha \right\} \cdot \|v - R_o v\|_E^2 &= \\ &= \sum_{\substack{E \subset \Gamma_R \\ h_E/h_{\min, E}^2 \geq \alpha}} \frac{h_E}{h_{\min, E}^2} \cdot \|v - R_o v\|_E^2 + \sum_{\substack{E \subset \Gamma_R \\ h_E/h_{\min, E}^2 < \alpha}} \alpha \cdot \|v - R_o v\|_E^2 \\ &\lesssim \sum_{T \cap \Gamma_R \neq \emptyset} h_{\min, T}^{-2} \cdot (\|v - R_o v\|_T^2 + \|C_T^T \nabla(v - R_o v)\|_T^2) + \sum_{\substack{E \subset \Gamma_R \\ h_E/h_{\min, E}^2 < \alpha}} \|\alpha^{1/2} \cdot v\|_E^2 \\ &\lesssim m_1(v, \mathcal{T}_h)^2 \cdot \|v\|^2 + \|\alpha^{1/2} \cdot v\|_{\Gamma_R}^2 \lesssim m_1(v, \mathcal{T}_h)^2 \cdot \|v\|^2 \end{aligned}$$

which finishes the proof. ■

**Remark 3.2** Note that the matching function  $m_1(\cdot, \cdot)$  does not influence the first interpolation estimate (3.11) but all the other ones.  $\square$

**Remark 3.3** The interpolation error estimate (3.12) for *isotropic* meshes reads without the factor  $m_1(v, \mathcal{T}_h)$  on the right-hand side (more precisely, then  $m_1 \sim 1$ , and  $m_1(v, \mathcal{T}_h)$  merges with the remaining interpolation constant).

In contrast to this *isotropic* interpolation error estimate, the factor  $m_1(v, \mathcal{T}_h)$  is necessary for the interpolation error estimate on *anisotropic* meshes (i.e. the estimate does not hold without  $m_1$ ). To show this, an *analytical* example is presented.

Assume  $\Omega = (0, 1) \times (0, 1)$  and  $\Gamma_D \equiv \partial\Omega$ . Choose the anisotropic function

$$v(x, y) := (1 - e^{-\alpha x} - (1 - e^{-\alpha})x) \cdot 4y(1 - y) \in H_o^1(\Omega) \quad , \quad \alpha \gg 1 \quad ,$$

which exhibits an exponential layer of steepness  $\alpha$  along the line  $x = 0$  (cf. figure 5.1 on page 97). Consider meshes constructed as follows:

- Start with a rectangular mesh with  $m$  uniform intervals in  $x$  direction and  $n$  uniform intervals in  $y$  direction.
- Bisect each rectangle from the lower left to the upper right corner.

Let  $m$  and  $\alpha$  be arbitrary but *fixed* parameters, and obtain the family of triangulations  $\mathcal{T}_h$  by varying  $n$ . Extensive computations yield

$$\begin{aligned} \|h_{min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\|^2 &= \frac{\text{polynomial of order 7 in } n}{\text{polynomial of order 5 in } n} \\ m_1(v, \mathcal{T}_h)^2 &= \frac{\text{polynomial of order 6 in } n}{\text{polynomial of order 4 in } n} \\ \|v\| &= \|\nabla v\| \sim 1 \quad . \end{aligned}$$

A large  $n \gg m$  implies that the mesh  $\mathcal{T}_h$  is anisotropically aligned with the  $x$  direction whereas the function  $v$  is anisotropically aligned with the  $y$  direction. Thus the anisotropic mesh becomes increasingly misaligned with the anisotropic function as  $n$  increases. Not surprisingly, this results in  $\|h_{min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\|$  and  $m_1(v, \mathcal{T}_h)$  to grow like  $\mathcal{O}(n)$ . Hence

$$\|h_{min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\| \sim m_1(v, \mathcal{T}_h) \cdot \|v\| \quad ,$$

i.e. interpolation estimate (3.12) is asymptotically sharp for this example, and  $m_1(v, \mathcal{T}_h)$  can not be omitted. We expect (3.13)–(3.16) to be sharp as well but did not trouble to prove this conjecture.

As an example we have chosen the parameters  $m = 1000$  and  $\alpha = 1000$ . Then the quotient  $\|h_{min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\| / m_1(v, \mathcal{T}_h) / \|\nabla v\| \rightarrow 2.393\dots$  as  $n \rightarrow \infty$ .

It might be worth mentioning that all these results were obtained using a computer algebra system. The ‘formula’ of  $\|h_{min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\|$  occupies about 25 kByte of storage, and thus it is impossible to be absolutely sure of the correct result. Nevertheless numerous tests have been carried out to verify the result, as well as independent checks against a finite element program.

Lastly, the same behaviour can be obtained with an *isotropic* function, say  $v = 4^2 \cdot x(1 - x) \cdot y(1 - y)$ . However, the example from above is better suited to the situation of an *anisotropic* function and anisotropic mesh.  $\square$

**Remark 3.4** The analysis shows that, for example for the interpolation estimate (3.12),

$$\|h_{min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\|^2 \lesssim \|\alpha^{1/2} \cdot v\|_{\Gamma_R}^2 + m_1^2(v, \mathcal{T}_h) \cdot \|\nabla v\|^2 \leq m_1^2(v, \mathcal{T}_h) \cdot \|v\|^2 \quad ,$$

i.e. the  $\Gamma_R$  part of the energy norm is *not* emphasized by the factor  $m_1^2(v, \mathcal{T}_h)$ . Therefore, if  $\|\alpha^{1/2} \cdot v\|_{\Gamma_R}$  dominates  $\|\nabla v\|$  in the energy norm it could be advantageous to re-define the matching function by

$$\tilde{m}_1(v, \mathcal{T}_h) := \|h_{min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\| / \|v\| \quad .$$

This matter, however, requires further investigation.  $\square$

### 3.2.3 Anisotropic residual error estimator

In order to derive lower bounds on the error, the functions  $f \in L_2(\Omega)$ ,  $g_N \in L_2(\Gamma_N)$  and  $g_R \in L_2(\Gamma_R)$  have to be approximated (or replaced) by certain counterparts from finite dimensional spaces. For a rather general setting, let  $P_1$  be an arbitrary mapping from  $L_2(\Omega)$  into the space of piecewise *constant* functions over  $\mathcal{T}_h$ . Thus

$$P_1(v) \Big|_T \in \mathbb{P}^0(T) \quad \forall T \in \mathcal{T}_h$$

(or more precisely, the restriction is constant on the interior of  $T$ ). Similarly, let  $P_2$  be an arbitrary mapping from  $L_2(\Gamma_N)$  into the space of piecewise *constant* functions over the  $\Gamma_N$  faces, i.e.

$$P_2(v) \Big|_E \in \mathbb{P}^0(E) \quad \forall E \subset \Gamma_N \quad .$$

Finally, let  $P_3$  be an arbitrary mapping from  $L_2(\Gamma_R)$  into the space of piecewise *linear* functions over the  $\Gamma_R$  faces, i.e.

$$P_3(v) \Big|_E \in \mathbb{P}^1(E) \quad \forall E \subset \Gamma_R \quad .$$

These mappings and their possible implementations (e.g. by projection operators) are discussed in Remark 3.5 on page 50.

Next, we introduce the so-called element residual and the face residual. These residuals play an important role not only in this Section but also for other error estimators.

**Definition 3.3 (Element and face residual)** *Let  $v_h \in V_{o,h}$  be an arbitrary finite element function. The element residual over a tetrahedron  $T$  is defined by*

$$r_T(v_h) := P_1 f + \Delta v_h \quad . \quad (3.17)$$

*The definition of the face residual depends on whether the corresponding face  $E$  is an interior face or a face from  $\Gamma_N$  or  $\Gamma_R$ . For a finite element function  $v_h \in V_{o,h}$  and  $x \in E$  set*

$$r_E(v_h)(x) := \begin{cases} \lim_{t \rightarrow +0} \left[ \frac{\partial v_h}{\partial n_E}(x + t n_E) + \frac{\partial v_h}{\partial(-n_E)}(x + t(-n_E)) \right] & \text{if } E \subset \Omega \setminus \Gamma \\ P_2 g_N - \frac{\partial v_h}{\partial n}(x) & \text{if } E \subset \Gamma_N \\ \alpha \cdot (P_3 g_R - v_h(x)) - \frac{\partial v_h}{\partial n}(x) & \text{if } E \subset \Gamma_R \end{cases} \quad . \quad (3.18)$$

*Here the vectors  $n_E$  and  $n$  have slightly different meaning – for interior faces  $E$  let  $n_E - E$  be any of the two unitary normal vectors. In contrast, for boundary faces  $E \subset \Gamma_N \cup \Gamma_R$  let  $n - E$  be the outer unitary normal vector. Note that the first face residual is also known as gradient jump or jump residual.*

We remark that the face residual  $r_E(v_h)$  is from

$$r_E(v_h) \in \begin{cases} \mathbb{P}^0(E) & \text{when } E \subset \Omega \setminus \Gamma \text{ or } E \subset \Gamma_N \\ \mathbb{P}^1(E) & \text{when } E \subset \Gamma_R \end{cases} .$$

Obviously  $r_T(v_h) = P_1 f$  holds for piecewise linear basis functions as considered here. The definition above however indicates how to modify this theory to treat higher order basis functions, cf. Remark 3.12 on page 67. Moreover, this element residual is related to the classical form of the differential operator and is as such problem dependent. With the help of the element residuals and the face residuals we now define the residual error estimator.

**Definition 3.4 (Residual error estimator)** *The local residual error estimator  $\eta_{R,T}(u_h)$  for a tetrahedron  $T$  is defined by*

$$\begin{aligned} \eta_{R,T}(u_h) := & h_{\min,T} \cdot \left( \|r_T(u_h)\|_T^2 + \sum_{E \subset \partial T \setminus \Gamma} h_E^{-1} \cdot \|r_E(u_h)\|_E^2 + \sum_{E \subset \partial T \cap \Gamma_N} h_E^{-1} \cdot \|r_E(u_h)\|_E^2 + \right. \\ & \left. + \sum_{E \subset \partial T \cap \Gamma_R} \min \left\{ 1, \frac{h_E}{\alpha h_{\min,T}^2} \right\} \cdot h_E^{-1} \cdot \|r_E(u_h)\|_E^2 \right)^{1/2} . \end{aligned} \quad (3.19)$$

*In order to keep the notation short, we also introduce a local approximation term for a tetrahedron  $T$  by*

$$\begin{aligned} \zeta_T := & h_{\min,T} \cdot \left( \|f - P_1 f\|_{\omega_T}^2 + \sum_{E \subset \partial T \cap \Gamma_N} h_E^{-1} \cdot \|g_N - P_2 g_N\|_E^2 + \right. \\ & \left. + \sum_{E \subset \partial T \cap \Gamma_R} \min \left\{ 1, \frac{h_E}{\alpha h_{\min,T}^2} \right\} \cdot h_E^{-1} \cdot \|\alpha \cdot (g_R - P_3 g_R)\|_E^2 \right)^{1/2} \end{aligned} \quad (3.20)$$

*and the global terms*

$$\eta_R^2(u_h) := \sum_{T \in \mathcal{T}_h} \eta_{R,T}^2(u_h) \quad \text{and} \quad \zeta^2 := \sum_{T \in \mathcal{T}_h} \zeta_T^2 . \quad (3.21)$$

Recall that  $\alpha = \alpha(\mathbf{x})$  is assumed piecewise constant over the Robin boundary faces. The value of  $\alpha$  is, of course, always taken from the face  $E$  under consideration. For example, a term like  $h_E \cdot \alpha^{-1} h_{\min,T}^{-2}$  may illustrate such an (implicit) notation.

We now formulate the main result for the residual error estimator, namely bounds of the error from above and below.

**Theorem 3.4 (Residual error estimation)** *Let  $u \in H_o^1(\Omega)$  be the exact solution and  $u_h \in V_{o,h}$  be the finite element solution.*

*Then the error is bounded locally from below for all  $T \in \mathcal{T}_h$  by*

$$\eta_{R,T}(u_h) \lesssim \| \|u - u_h\| \|_{\omega_T} + \zeta_T . \quad (3.22)$$

*The error is bounded globally from above by*

$$\| \|u - u_h\| \| \lesssim m_1(u - u_h, \mathcal{T}_h) \cdot \left[ \eta_R^2(u_h) + \zeta^2 \right]^{1/2} . \quad (3.23)$$

**Proof:** Firstly, estimate (3.22) will be proven. We start with the norm  $\|r_T(u_h)\|_T$  of the element residual  $r_T = r_T(u_h) := P_1 f + \Delta u_h$ . Since we use linear ansatz functions,  $r_T \in \mathbb{P}^0(T)$  holds. For  $\mathbf{x} \in T$  let

$$w(\mathbf{x}) := r_T(u_h)(\mathbf{x}) \cdot b_T(\mathbf{x}) \quad \in \mathbb{P}^4(T) \cap H_o^1(T) \quad ,$$

where the element bubble function  $b_T$  is from (2.19). Integration by parts yields

$$\begin{aligned} \int_T r_T \cdot w &= \int_T (f + \Delta u_h) \cdot w + \int_T (P_1 f - f) \cdot w \\ &= \int_T \nabla^T(u - u_h) \cdot \nabla w + \int_T (P_1 f - f) \cdot w \quad ; \\ \left| \int_T r_T \cdot w \right| &\leq \|\nabla(u - u_h)\|_T \cdot \|\nabla w\|_T + \|f - P_1 f\|_T \cdot \|w\|_T \quad . \end{aligned}$$

Recalling (2.24), (2.25), and  $0 \leq b_T \leq 1$  gives the following bounds

$$\left. \begin{aligned} \left| \int_T r_T \cdot w \right| &= \|b_T^{1/2} \cdot r_T\|_T^2 \sim \|r_T\|_T^2 \\ \|\nabla w\|_T &= \|\nabla(b_T \cdot r_T)\|_T \lesssim h_{min,T}^{-1} \cdot \|r_T\|_T \\ \|w\|_T &= \|b_T \cdot r_T\|_T \leq \|r_T\|_T \end{aligned} \right\} \quad (3.24)$$

that result in

$$\begin{aligned} \|r_T\|_T^2 &\lesssim \|\nabla(u - u_h)\|_T \cdot h_{min,T}^{-1} \cdot \|r_T\|_T + \|f - P_1 f\|_T \cdot \|r_T\|_T \\ \text{and} \quad h_{min,T}^2 \cdot \|r_T\|_T^2 &\lesssim \|\nabla(u - u_h)\|_T^2 + h_{min,T}^2 \cdot \|f - P_1 f\|_T^2 \quad . \end{aligned}$$

Now we aim at a bound of the norm  $\|r_E(u_h)\|_E$  of the gradient jump across some inner face  $E \subset \Omega \setminus \Gamma$ . Since we use linear ansatz functions  $r_E \in \mathbb{P}^0(E)$  holds. Let  $T_1$  and  $T_2$  be the two tetrahedra that  $E$  belongs to. The right hand side  $f = -\Delta u$  is assumed to be in  $L_2(\Omega)$ . Integration by parts yields for any function  $w \in H_o^1(\omega_E)$

$$\begin{aligned} 0 &= - \int_{\omega_E} \nabla^T w \nabla u + \int_{\omega_E} w \cdot f \\ \text{and} \quad - \int_E w \cdot r_E(u_h) &= \sum_{i=1}^2 \int_{\partial T_i} w \cdot \frac{\partial u_h}{\partial n} = \sum_{i=1}^2 \left( \int_{T_i} \nabla^T w \nabla u_h + \int_{T_i} w \cdot \Delta u_h \right) \\ &= \sum_{i=1}^2 \left( \int_{T_i} \nabla^T w \nabla u_h + \int_{T_i} w \cdot (r_{T_i} - P_1 f) \right) \\ &= \sum_{i=1}^2 \left( \int_{T_i} \nabla^T w \nabla (u_h - u) + \int_{T_i} w \cdot (r_{T_i} + f - P_1 f) \right) \quad . \end{aligned}$$

Let now the function  $w \in H_o^1(\omega_E)$  be defined by

$$w := F_{ext}(r_E(u_h)) \cdot b_E \quad ,$$

with  $F_{ext}$  being the extension operator of (2.23), and  $b_E$  being the face bubble function of (2.20). Because of  $w|_E = r_E \cdot b_E|_E$  we conclude

$$\int_E r_E^2 \cdot b_E \leq \sum_{i=1}^2 \left( \|\nabla(u - u_h)\|_{T_i} \cdot \|\nabla w\|_{T_i} + (\|r_{T_i}\|_{T_i} + \|f - P_1 f\|_{T_i}) \cdot \|w\|_{T_i} \right) \quad .$$

The function  $w$  is piecewise cubic on  $\omega_E$ . The equivalence relations (2.26) – (2.28) imply

$$\left. \begin{aligned} \int_E r_E \cdot w &= \int_E r_E^2 \cdot b_E = \|b_E^{1/2} \cdot r_E\|_E^2 \sim \|r_E\|_E^2 \\ \|\nabla w\|_{T_i} &= \|\nabla (F_{ext}(r_E) \cdot b_E)\|_{T_i} \sim h_E^{1/2} h_{min,T_i}^{-1} \cdot \|r_E\|_E \\ \text{and } \|w\|_{T_i} &= \|F_{ext}(r_E) \cdot b_E\|_{T_i} \sim h_E^{1/2} \cdot \|r_E\|_E \end{aligned} \right\} \quad (3.25)$$

and subsequently lead to

$$\begin{aligned} \|r_E\|_E^2 &\lesssim \sum_{i=1}^2 \left( \|\nabla(u - u_h)\|_{T_i} \cdot h_E^{1/2} h_{min,T_i}^{-1} \|r_E\|_E + \right. \\ &\quad \left. + (\|r_{T_i}\|_{T_i} + \|f - P_1 f\|_{T_i}) \cdot h_E^{1/2} \|r_E\|_E \right) . \end{aligned}$$

The dimensions  $h_E \sim h_{E,T_i}$  and  $h_{min,T_i}$  cannot change rapidly for adjacent tetrahedra. Recalling the bound of  $\|r_T\|_T$  from above we conclude

$$\|r_E\|_E \lesssim h_E^{1/2} h_{min,E}^{-1} \cdot \left( \|\nabla(u - u_h)\|_{\omega_E} + h_{min,E} \|f - P_1 f\|_{\omega_E} \right) .$$

For a fixed tetrahedron  $T = T_1$  we sum up over all (inner) faces  $E \subset \partial T \setminus \Gamma$  and obtain

$$\sum_{E \subset \partial T \setminus \Gamma} \frac{h_{min,T}^2}{h_E} \cdot \|r_E(u_h)\|_E^2 \lesssim \|\nabla(u - u_h)\|_{\omega_T}^2 + h_{min,T}^2 \|f - P_1 f\|_{\omega_T}^2 .$$

Next, the norm  $\|r_E(u_h)\|_E$  for a face  $E \subset \Gamma_N$  of the Neumann boundary is to be bounded. Again  $r_E \in \mathbb{P}^0(E)$ . Let the tetrahedron  $T$  contain  $E$ . Define a function  $w := F_{ext}(r_E(u_h)) \cdot b_E$ , with  $F_{ext}$  from (2.23), and recall that  $w = 0$  on  $\partial T \setminus E$ . Integration by parts and  $g_N = \partial u / \partial n$  then yield

$$\begin{aligned} \int_E r_E(u_h) \cdot w &= \int_E \left( g_N - \frac{\partial u_h}{\partial n} \right) \cdot w + \int_E (P_2 g_N - g_N) \cdot w \\ &= \int_T \nabla^T(u - u_h) \cdot \nabla w - \int_T (P_1 f + \Delta u_h) \cdot w + \int_T (P_1 f - f) \cdot w + \\ &\quad + \int_E (P_2 g_N - g_N) \cdot w \\ &\leq \|\nabla(u - u_h)\|_T \cdot \|\nabla w\|_T + (\|r_T\|_T + \|P_1 f - f\|_T) \cdot \|w\|_T + \\ &\quad + \|g_N - P_2 g_N\|_E \cdot \|w\|_E . \end{aligned}$$

Utilizing the same equivalence relations (3.25) as before, inserting the bound of  $\|r_T\|_T$  from above, and  $\|w\|_E \leq \|r_E\|_E$  gives

$$\begin{aligned} \sum_{E \subset \partial T \cap \Gamma_N} \frac{h_{min,T}^2}{h_E} \cdot \|r_E(u_h)\|_E^2 &\lesssim \|\nabla(u - u_h)\|_T^2 + h_{min,T}^2 \|f - P_1 f\|_T^2 + \\ &\quad + \sum_{E \subset \partial T \cap \Gamma_N} \frac{h_{min,T}^2}{h_E} \cdot \|g_N - P_2 g_N\|_E^2 . \end{aligned}$$

Finally we seek a bound of  $\|r_E(u_h)\|_E$  for a face  $E \subset \Gamma_R$  of the Robin boundary. Recall  $r_E(u_h) = \alpha \cdot (P_3 g_R - u_h) - \partial u_h / \partial n \in \mathbb{P}^1(E)$ , let  $T \supset E$ , and set  $w := F_{ext}(r_E(u_h)) \cdot b_E$ . From  $w = 0$  on  $\partial T \setminus E$ , and by integration by parts it follows that

$$\begin{aligned}
\int_E r_E(u_h) \cdot w &= \int_E \left( \alpha \cdot (g_R - u_h) - \frac{\partial u_h}{\partial n} \right) \cdot w + \int_E \alpha \cdot (P_3 g_R - g_R) \cdot w \\
&= \int_E \left( \alpha(u - u_h) + \frac{\partial(u - u_h)}{\partial n} \right) \cdot w + \int_E \alpha \cdot (P_3 g_R - g_R) \cdot w \\
&= \int_T \nabla^T(u - u_h) \cdot \nabla w - \int_T (f + \Delta u_h) \cdot w + \int_E \alpha(u - u_h) \cdot w \\
&\quad + \int_E \alpha \cdot (P_3 g_R - g_R) \cdot w \\
&\leq \|\nabla(u - u_h)\|_T \cdot \|\nabla w\|_T + (\|r_T\|_T + \|P_1 f - f\|_T) \cdot \|w\|_T + \\
&\quad + \alpha \cdot \left( \|u - u_h\|_E + \|g_R - P_3 g_R\|_E \right) \cdot \|w\|_E \quad .
\end{aligned}$$

As before, we use the equivalence relations (3.25), the bound of  $\|r_T\|_T$ , and  $\|w\|_E \leq \|r_E\|_E$  which eventually implies

$$\begin{aligned}
\min \left\{ 1, \frac{h_E}{\alpha h_{min,T}^2} \right\} \cdot \frac{h_{min,T}^2}{h_E} \cdot \|r_E(u_h)\|_E^2 &\lesssim \\
&\lesssim \|\nabla(u - u_h)\|_T^2 + h_{min,T}^2 \|f - P_1 f\|_T^2 + \min \left\{ 1, \frac{\alpha h_{min,T}^2}{h_E} \right\} \cdot \|\alpha^{1/2}(u - u_h)\|_E^2 + \\
&\quad + \min \left\{ 1, \frac{h_E}{\alpha h_{min,T}^2} \right\} \cdot \frac{h_{min,T}^2}{h_E} \cdot \|\alpha \cdot (g_R - P_3 g_R)\|_E^2 \\
&\leq \|u - u_h\|_T^2 + h_{min,T}^2 \|f - P_1 f\|_T^2 + \\
&\quad + \min \left\{ 1, \frac{h_E}{\alpha h_{min,T}^2} \right\} \cdot \frac{h_{min,T}^2}{h_E} \cdot \|\alpha \cdot (g_R - P_3 g_R)\|_E^2 \quad .
\end{aligned}$$

Summing up over all faces  $E \subset \partial T \cap \Gamma_R$  and collecting the estimates of the norms for all four types of residuals accomplishes the proof of (3.22).

Secondly, in order to derive (3.23) we utilize the orthogonality property of the error

$$a(u - u_h, v_h) = 0 \quad \forall v_h \in V_{o,h} \quad .$$

Integration by parts and the weak formulation (3.2) give for all  $v \in H_o^1(\Omega)$

$$\begin{aligned}
a(u - u_h, v) &= a(u - u_h, v - R_o v) \\
&= \int_\Omega \nabla^T(u - u_h) \cdot \nabla(v - R_o v) + \int_{\Gamma_R} \alpha \cdot (u - u_h) \cdot (v - R_o v) \\
&\stackrel{(3.2)}{=} \sum_{T \in \mathcal{T}_h} \int_T (f + \Delta u_h) \cdot (v - R_o v) + \sum_{E \subset \Omega \setminus \Gamma} \int_E r_E(u_h) \cdot (v - R_o v) + \\
&\quad + \sum_{E \subset \Gamma_N} \int_E \left( g_N - \frac{\partial u_h}{\partial n} \right) \cdot (v - R_o v) +
\end{aligned}$$

$$\begin{aligned}
& + \sum_{E \in \Gamma_R} \int_E \left( \alpha(g_R - u_h) - \frac{\partial u_h}{\partial n} \right) \cdot (v - R_o v) \\
\leq & \left( \sum_{T \in \mathcal{T}_h} h_{min,T}^2 \cdot \|f + \Delta u_h\|_T^2 \right)^{1/2} \cdot \left( \sum_{T \in \mathcal{T}_h} h_{min,T}^{-2} \cdot \|v - R_o v\|_T^2 \right)^{1/2} + \\
& + \left( \sum_{E \subset \Omega \setminus \Gamma} \frac{h_{min,E}^2}{h_E} \|r_E(u_h)\|_E^2 \right)^{1/2} \cdot \left( \sum_{E \subset \Omega \setminus \Gamma} \frac{h_E}{h_{min,E}^2} \cdot \|v - R_o v\|_E^2 \right)^{1/2} + \\
& + \left( \sum_{E \subset \Gamma_N} \frac{h_{min,E}^2}{h_E} \left\| g_N - \frac{\partial u_h}{\partial n} \right\|_E^2 \right)^{1/2} \cdot \left( \sum_{E \subset \Gamma_N} \frac{h_E}{h_{min,E}^2} \|v - R_o v\|_E^2 \right)^{1/2} + \\
& + \left( \sum_{E \in \Gamma_R} \min \left\{ 1, \frac{h_E}{\alpha h_{min,E}^2} \right\} \cdot \frac{h_{min,E}^2}{h_E} \left\| \alpha(g_R - u_h) - \frac{\partial u_h}{\partial n} \right\|_E^2 \right)^{1/2} \\
& \quad \cdot \left( \sum_{E \in \Gamma_R} \max \left\{ \frac{h_E}{h_{min,E}^2}, \alpha \right\} \cdot \|v - R_o v\|_E^2 \right)^{1/2}.
\end{aligned}$$

Every second root term is bounded by  $m_1(v, \mathcal{T}_h) \cdot \|v\|$  by means of the interpolation Theorem 3.3 on page 41. Substituting  $v := u - u_h$  yields an upper bound of the error

$$\begin{aligned}
\|u - u_h\|^2 \lesssim & m_1(u - u_h, \mathcal{T}_h)^2 \cdot \left( \sum_{T \in \mathcal{T}_h} h_{min,T}^2 \cdot \|f + \Delta u_h\|_T^2 + \sum_{E \subset \Omega \setminus \Gamma} \frac{h_{min,E}^2}{h_E} \|r_E(u_h)\|_E^2 \right. \\
& + \sum_{E \subset \Gamma_N} \frac{h_{min,E}^2}{h_E} \left\| g_N - \frac{\partial u_h}{\partial n} \right\|_E^2 + \\
& \left. + \sum_{E \in \Gamma_R} \min \left\{ 1, \frac{h_E}{\alpha h_{min,E}^2} \right\} \cdot \frac{h_{min,E}^2}{h_E} \left\| \alpha(g_R - u_h) - \frac{\partial u_h}{\partial n} \right\|_E^2 \right). \quad (3.26)
\end{aligned}$$

Finally, utilizing the triangle inequality  $\|f + \Delta u_h\|_T \leq \|r_T(u_h)\|_T + \|f - P_1 f\|_T$ , and similarly for  $g_N$  and  $g_R$ , results in the upper bound (3.23) of the error.  $\blacksquare$

The term  $m_1(u - u_h, \mathcal{T}_h)$  is discussed in more detail in the next Section.

**Remark 3.5** The terms  $P_1 f$ ,  $P_2 g_N$  and  $P_3 g_R$  appear both in the definition of the residuals  $r_T(u_h)$  and  $r_E(u_h)$  as well as in inequalities (3.22) and (3.23).

Assume for the moment that these terms are replaced by arbitrary functions from  $L_2(\Omega)$ ,  $L_2(\Gamma_N)$  or  $L_2(\Gamma_R)$ , respectively. Then one would obtain an upper bound of the error similar to (3.23) but a lower bound (3.22) would no longer hold. Choosing the original values  $f, g_N, g_R$  instead of  $P_1 f, P_2 g_N, P_3 g_R$  would, for example, result in the error bound (3.26).

There are two reasons for using these mappings. Mainly, such terms (or similar ones from a finite dimensional space) are required to derive a lower bound of the error.

Furthermore there is a useful interpretation for these projections. It may be difficult to evaluate exactly the integrals over  $f, g_N$  and  $g_R$ , respectively. If  $f$  is suitably smooth (e.g.  $f \in L_2 \cap C^0(T)$ ) then  $P_1 f$  may represent a quadrature rule. For example, the midpoint quadrature rule is equivalent to  $P_1 : L_2(T) \cap C^0(T) \rightarrow \mathbb{P}^0(T)$ ,  $P_1 f(\mathbf{x}) := f(\mathbf{x}_{midpoint})$  on  $T$ . The term  $\|f - P_1 f\|$  then measures the quadrature error. Other choices of  $P_1$  include, for example, projection operators.  $\square$

**Remark 3.6** Siebert [50] proposes a similar error estimator for rectangular or cuboidal finite elements. There the factor of the gradient jump in the definition of the error estimator equals  $h_E$  instead of  $h_{min}^2/h_E$  as in our work (cf. (3.19)). Thus Siebert has to impose an additional condition on  $u_h$  to give a reliable lower bound of the error. This renders our estimator slightly more general.  $\square$

**Remark 3.7** For Robin boundary conditions, the function  $\alpha = \alpha(\mathbf{x})$ ,  $\mathbf{x} \in \Gamma_R$ , has been assumed piecewise constant over the  $\Gamma_R$  faces. If one wants to consider an arbitrary function  $\alpha(\mathbf{x})$  then the  $\Gamma_R$  residual has to be modified (cf. (3.18)) to

$$r_E := P_3\left(\alpha \cdot (g_R - u_h) - \partial u_h / \partial n\right) \quad .$$

If additionally  $P_3$  is a linear operator (into the space of piecewise linear functions over the  $\Gamma_R$  faces) then this face residual can be written as

$$r_E = P_3(\alpha g_R) - P_3(\alpha u_h) - \partial u_h / \partial n \quad .$$

The corresponding entry in the approximation term  $\zeta_{R,T}$  (cf. (3.20)) then becomes

$$\alpha g_R - P_3(\alpha g_R) - (\alpha u_h - P_3(\alpha u_h))$$

and would thus depend on  $u_h$  (the same holds true if  $P_3$  is a nonlinear operator). By assuming  $\alpha$  piecewise constant such a dependence is avoided.  $\square$

### 3.2.4 Discussion of the matching function

The matching function  $m_1$  has been motivated, defined and discussed extensively in Section 3.2.1. It enters the interpolation error estimates (Theorem 3.3). But the knowledge of the value of  $m_1(v, \mathcal{T}_h)$  is not necessary there since these estimates serve only as auxiliary tools.

The situation is different for the finite element error estimates. Firstly, the term  $m_1(u - u_h, \mathcal{T}_h)$  is involved in the upper bound (3.23) of the error. Secondly, the lower and upper error bound readily imply the two-sided estimate

$$c \cdot \eta_R^2 - \zeta^2 \lesssim \|u - u_h\| \lesssim m_1^2(u - u_h, \mathcal{T}_h) \cdot (\eta_R^2 + \zeta^2) \quad .$$

Both results clearly illustrate that  $m_1(u - u_h, \mathcal{T}_h)$  has to be small to enable an efficient error estimation (note  $m_1 \geq 1$ ). Unfortunately the matching function  $m_1(u - u_h, \mathcal{T}_h)$  contains the (usually unknown) exact solution  $u$ . Two questions arise immediately.

1. Can one construct meshes that yield a small  $m_1(u - u_h, \mathcal{T}_h)$ , even if  $u$  is unknown?
2. Can  $m_1(u - u_h, \mathcal{T}_h)$  be estimated or bounded somehow?

Two approaches seem possible at present.

**Heuristic approach:** Choose a ‘suitable’ anisotropic mesh for a given problem (with a certain solution  $u$ ). From a heuristic point of view, it is sensible to aim at a good interpolation of the anisotropic solution  $u$ . Therefore the mesh should be aligned with the (unknown) solution  $u$ , and a possible strategy could be, for example:

- Detect regions of anisotropic behaviour of the solution  $u$  by investigating the finite element solution  $u_h$ .
- Based on that information, determine a (quasi) optimal stretching direction, aspect ratio and size of the finite elements.
- Generate a new mesh.

This strategy resembles the adaptive anisotropic procedure which has been introduced in chapter 1. Indeed, this strategy (which focuses on  $m_1$ ) should be part of any adaptive algorithm. In this way it can be the basis for a small  $m_1(u - u_h, \mathcal{T}_h)$ , and thus for an accurate error estimation.

The meshes of the numerical examples 1 and 4 are constructed on similar principles (but without adaptivity). The corresponding results show that  $m_1$  is small indeed. Hence this heuristic approach is feasible, and it can yield the anticipated effects. Note that so far  $m_1(u - u_h, \mathcal{T}_h)$  is *not* evaluated or estimated.

**Numerical approach:** A different (or additional) possibility is to *estimate*

$$m_1(u - u_h, \mathcal{T}_h) = \left( \sum_{T \in \mathcal{T}_h} h_{\min, T}^{-2} \cdot \|C_T^T \nabla(u - u_h)\|_T^2 \right)^{1/2} / \|\nabla(u - u_h)\| \quad ,$$

cf. (3.5). For example, one can replace the unknown gradient  $\nabla u$  of the solution by an approximation. For this we utilize here a recovered gradient  $\nabla^R u_h$  giving

$$m_1^R(u - u_h, \mathcal{T}_h) := \left( \sum_{T \in \mathcal{T}_h} h_{\min, T}^{-2} \cdot \|C_T^T (\nabla^R u_h - \nabla u_h)\|_T^2 \right)^{1/2} / \|\nabla^R u_h - \nabla u_h\| \quad . \quad (3.27)$$

For simplicity we have chosen the recovered gradient  $\nabla^R u_h$  to be the linear Lagrange interpolate at the nodes of the mesh. The value at a node  $a$  is given by

$$\nabla^R u_h(a) := \sum_{T: a \in \mathcal{N}_T} \frac{|T|}{|\omega_a|} \cdot \nabla u_h|_T \quad \text{with} \quad \omega_a := \bigcup_{T: a \in \mathcal{N}_T} T \quad , \quad (3.28)$$

cf. [58] for example. We remark that other techniques are possible, e.g. utilizing superconvergence points, or by the superconvergence patch recovery technique (cf. [67]). A closer investigation, in particular in the anisotropic context here, is certainly desirable but beyond the scope of our work. Hence the recovered gradient mentioned above shall suffice, although we are aware that the accuracy may deteriorate on adaptively refined meshes.

We note that, strictly speaking,  $m_1^R(u - u_h, \mathcal{T}_h)$  depends on  $u_h$  but not on  $u$ . Nevertheless the notation  $m_1^R(u - u_h, \mathcal{T}_h)$  is used to indicate that it approximates  $m_1(u - u_h, \mathcal{T}_h)$ .

### 3.3 A local Dirichlet problem error estimator

#### 3.3.1 Introduction and definition

Local problem error estimators have been known for a long time [14, 16, 57, 58]. In this Section we demonstrate on the example of a local Dirichlet problem that these error estimators can be applied to anisotropic meshes as well. As far as we know this is the first rigorous analytical investigation for this type of anisotropic error estimator.

Let us first recall the situation on *isotropic meshes*. The basic idea is to solve a local problem with higher accuracy. The difference between its solution and the original finite element solution serves as error estimator. The existing estimators differ mainly about the *local problem space*  $V_T$ , i.e. the space in which the local problem is to be solved. Certain inequalities and equivalence results as well as bubble functions play a vital role in the analysis. With their help one can show the equivalence of the local problem error estimator and the residual error estimator.

On *anisotropic meshes*, we basically employ the same methodology. The careful selection of the local problem space  $V_T$  requires particular attention. Our exemplary choice here enables us to derive Lemma 3.5. This Lemma states an important equivalence relation over the local problem space  $V_T$ . While the proof of the Lemma is trivial on isotropic meshes, it is fairly technical on anisotropic meshes. Here we derive a new technique which utilizes compactness arguments. Additionally the role of an appropriate local problem space  $V_T$  becomes apparent in the Lemma and its proof. Next, Theorem 3.6 states the equivalence of the local problem error estimator  $\eta_{D,T}$  and the residual error estimator  $\eta_{R,T}$ . The main results, namely lower and upper bounds of the error, are given in Theorem 3.7 on page 64. In Section 3.3.4 it is shown that the local Dirichlet problem is well-conditioned. Finally, we remark on Robin boundary conditions and on higher order ansatz functions in Section 3.3.5.

Practical considerations suggest that local problem error estimators should comply with the following, rather general guidelines (cf. [1, 58]).

1. The estimator should catch the local behaviour of the error.
2. The local finite element space should be richer than the original one ( $V_{o,h}$ ) to extract more information on the solution. The degrees of freedom (DOF) of the local problem should correspond to the residual terms of the finite element solution  $u_h$  (see below).
3. The estimator should be cheap to compute.

To derive the estimator, recall that  $a(\cdot, \cdot)$  is the bilinear form, and  $u$  and  $u_h$  denote the exact and the finite element solution, respectively. Let  $T$  be an arbitrary but fixed tetrahedron.

In accordance with points 1 and 3 from above, the subdomain of the local problem is chosen to be  $\omega_T$  (which is formed by  $T$  and all (at most four) adjacent tetrahedra that have a common face with  $T$ ). Let

$$H_o^1(\omega_T) := \{v \in H^1(\omega_T) : v = 0 \text{ on } (\partial\omega_T \setminus \Gamma) \cup \Gamma_D\} \quad ,$$

i.e. every function of this space can be seen as a function from  $H_o^1(\Omega)$  with support in  $\omega_T$ . The error  $e = u - u_h$  then satisfies

$$a(u - u_h, v) = \int_{\omega_T} f \cdot v - \int_{\omega_T} \nabla^T u_h \nabla v + \int_{\Gamma_N \cap \partial\omega_T} g_N \cdot v + \int_{\Gamma_R \cap \partial\omega_T} \alpha \cdot (g_R - u_h) \cdot v \quad \forall v \in H_o^1(\omega_T) \quad .$$

A straightforward approximation of the space  $H_o^1(\omega_T)$  by some local, finite dimensional space  $V_T \subset H_o^1(\omega_T)$  leads to the problem:

$$\begin{aligned} \text{Find } e_T \in V_T : \quad & a(e_T, v_T) = a(u - u_h, v_T) \quad \forall v_T \in V_T \\ \text{or } \int_{\omega_T} \nabla^T e_T \nabla v_T + \int_{\Gamma_R \cap \partial \omega_T} \alpha \cdot e_T \cdot v_T &= \int_{\omega_T} f \cdot v_T - \int_{\omega_T} \nabla^T u_h \nabla v_T + \\ &+ \int_{\Gamma_N \cap \partial \omega_T} g_N \cdot v_T + \int_{\Gamma_R \cap \partial \omega_T} \alpha \cdot (g_R - u_h) \cdot v_T \quad . \end{aligned}$$

Then  $\|e_T\|_{\omega_T}$  could serve as error estimator. Furthermore, in the light of Remark 3.5 on page 50 we want to use  $P_1 f$  instead of  $f$ , and similar projections of  $g_N$  and  $g_R$ .

Next, the local finite element space  $V_T$  is to be defined. In the context of the residual error estimator and the point 2 from above, we list the residuals which appear in  $\eta_{R,T}$  and their degrees of freedom on the one hand, and on the other the corresponding functions from  $V_T$ :

	$\eta_{R,T}$	DOF	$\eta_{D,T}$
$T$	$r_T \in \mathbb{P}^0(T)$	1	$b_T \in H_o^1(T)$
$E \subset \Omega \setminus \Gamma$	$r_E \in \mathbb{P}^0(E)$	1	$b_E \in H_o^1(\omega_E)$
$E \subset \Gamma_N$	$r_E \in \mathbb{P}^0(E)$	1	$b_E \in H^1(T)$
$E \subset \Gamma_R$	$r_E \in \mathbb{P}^1(E)$	3	$V_E \subset H^1(T)$

For  $E \subset \Gamma_R$  the subspace  $V_E$  is defined below. The first three entries of the table have their analogies in the isotropic analysis (e.g. [58]) but the fourth one seems to be new and thus requires some explanation.

On a Robin boundary face, the residual  $r_E \in \mathbb{P}^1(E)$  is linear (in contrast to the other  $\mathbb{P}^0$  residuals, cf. (3.17) and (3.18)). The three degrees of freedom of  $r_E$  should have a counterpart in  $V_T$ , preferably also with three degrees of freedom. Alternatively, from an analytic point of view we want  $b_E \cdot F_{ext}(r_E) \in V_T$ . Any  $r_E \in \mathbb{P}^1(E)$  can be written as a linear combination of the (linear) functions  $\lambda_{T,i}(\mathbf{x})|_E$ ,  $i = 1, 2, 3$ . Here  $\lambda_{T,i}(\mathbf{x}) = \lambda_{T,i}$  are the barycentric coordinates of  $T$  such that  $\lambda_{T,4}$  is related to the unique node which is not in  $E$ . Recalling the definition of the linear extension operator  $F_{ext}$ , one thus sets

$$b_{E,i} := b_E \cdot F_{ext}(\lambda_{T,i}(\mathbf{x})) = b_E \cdot \left( \lambda_{T,i} + \frac{\lambda_{T,4}}{3} \right) \quad i = 1, 2, 3, \quad (3.29)$$

to achieve that aim. All three functions  $b_{E,i}$  form the subspace  $V_E$  via

$$V_E := \text{span}\{b_{E,i} : i = 1, 2, 3\} \quad .$$

The local finite element space  $V_T$ , the local problem and the estimator are now defined as follows.

**Definition 3.5 (Local problem error estimator)** *Let the local finite element space  $V_T \subset H^1_0(\omega_T)$  be given by*

$$V_T := \text{span}\{b_T, b_E, b_{E',i} : E \subset \partial T \setminus (\Gamma_D \cup \Gamma_R), E' \subset \partial T \cap \Gamma_R, i = 1, 2, 3\} \quad . \quad (3.30)$$

*Find a solution  $e_T \in V_T$  of the local variational problem*

$$\begin{aligned} a(e_T, v_T) &\equiv \int_{\omega_T} \nabla^T e_T \nabla v_T + \int_{\Gamma_R \cap \partial \omega_T} \alpha \cdot e_T \cdot v_T \\ &\stackrel{!}{=} \int_{\omega_T} P_1 f \cdot v_T - \int_{\omega_T} \nabla^T u_h \nabla v_T + \int_{\Gamma_N \cap \partial \omega_T} P_2 g_N \cdot v_T + \\ &\quad + \int_{\Gamma_R \cap \partial \omega_T} \alpha \cdot (P_3 g_R - u_h) \cdot v_T \quad \forall v_T \in V_T \quad . \end{aligned} \quad (3.31)$$

*The local and global Dirichlet problem error estimator are then*

$$\eta_{D,T} := \|e_T\|_{\omega_T} \quad \text{and} \quad \eta_D^2(u_h) := \sum_{T \in \mathcal{T}_h} \eta_{D,T}^2(u_h) \quad . \quad (3.32)$$

Note that the particular choice of the local finite element space  $V_T$  (namely  $v_T = 0$  on  $\partial \omega_T \setminus \partial T$ ) reduces certain boundary integrals and norms, e.g.

$$\int_{\partial \omega_T \cap \Gamma_N} w \cdot v_T = \int_{\partial T \cap \Gamma_N} w \cdot v_T \quad \forall v_T \in V_T, w \in L_2(\partial \omega_T) \quad .$$

Two equivalent formulations of the local problem are as follows. Find  $e_T \in V_T$  such that

$$\begin{aligned} a(e_T, v_T) &= a(u - u_h, v_T) + \int_{\omega_T} (P_1 f - f) \cdot v_T + \int_{\Gamma_N \cap \partial \omega_T} (P_2 g_N - g_N) \cdot v_T + \\ &\quad + \int_{\Gamma_R \cap \partial \omega_T} \alpha \cdot (P_3 g_R - g_R) \cdot v_T \quad \forall v_T \in V_T \quad (3.33) \end{aligned}$$

$$\text{or} \quad a(e_T, v_T) = \sum_{T' \in \omega_T} \int_{T'} r_{T'} \cdot v_T + \sum_{E \subset \partial T \setminus \Gamma_D} \int_E r_E \cdot v_T \quad \forall v_T \in V_T \quad . \quad (3.34)$$

The weak formulation in the definition above can be seen as the discrete analogue of the local Dirichlet problem

$$\begin{aligned} -\Delta \varphi &= P_1 f && \text{in } \omega_T \\ \varphi &= u_h && \text{on } \partial \omega_T \setminus (\Gamma_N \cup \Gamma_R) \\ \partial \varphi / \partial n &= P_2 g_N && \text{on } \partial \omega_T \cap \Gamma_N \\ \partial \varphi / \partial n &= \alpha \cdot (P_3 g_R - \varphi) && \text{on } \partial \omega_T \cap \Gamma_R \end{aligned}$$

which is solved on the manifold  $u_h + V_T$ .

**Remark 3.8** The local problem error estimator presented here differs from the isotropic counterpart of Verfürth [58] by the local ansatz space  $V_T$ . More precisely, we employ less basis functions. Furthermore, other local problem error estimators are certainly possible (for some isotropic varieties see e.g. [58, Section 1.3]). Then modified definitions and/or a different analysis become necessary.  $\square$

### 3.3.2 Preliminary result

**Lemma 3.5** *The following relations hold for all  $v_T \in V_T$ .*

$$\|v_T\|_{\omega_T} \sim h_{\min,T} \cdot \|\nabla v_T\|_{\omega_T} \quad (3.35)$$

$$\|v_T\|_E \lesssim h_E^{-1/2} \cdot h_{\min,T} \cdot \|\nabla v_T\|_{\omega_T} \quad \forall E \subset \partial T \quad . \quad (3.36)$$

If  $T$  has a boundary face then the constants depend on the interior angles of the corners and edges of the domain  $\Omega$ .

**Proof:** Some temporary notation is introduced first. Let  $T_0 := T$ , and denote the remaining tetrahedra of  $\omega_T$  by  $T_1 \dots T_k$ , where  $k = 4$  for an interior tetrahedron  $T$  and  $k < 4$  if  $T$  has a boundary face. Enumerate the faces  $E_i$  of  $T$  such that interior faces of  $\omega_T$  come first (i.e.  $E_i := T \cap T_i$ ,  $i = 1 \dots k$ ), followed by Dirichlet boundary faces, and finally  $\Gamma_N$  or  $\Gamma_R$  faces.

Let us start with the proof of (3.35), and consider the tetrahedron  $T_i$ ,  $0 \leq i \leq k$ . The transformation matrices  $C_{T_i}$  and  $H_{T_i}$  will play a vital role (the transformation via  $A_{T_i}$  is unsuitable here). The corresponding equations and transformations, with the abbreviations introduced below, yield

$$\|v_T\|_{\omega_T}^2 = \sum_{i=0}^k \|v_T\|_{T_i}^2 = \sum_{i=0}^k 6|T_i| \cdot \|\bar{v}_T\|_{T_i}^2 \sim |T| \cdot \sum_{i=0}^k r_i \sim |T| \cdot r$$

$$\text{with } r_i := \|\bar{v}_T\|_{T_i}^2 \geq 0 \quad \text{and} \quad r := \sum_{i=0}^k r_i \geq 0 \quad ,$$

$$\text{and } \|\nabla v_T(\mathbf{x})\|_{\mathbb{R}^3} \stackrel{(2.11)}{=} \|H_{T_i}^{-1} C_{T_i}^T \nabla v_T(\mathbf{x})\|_{\mathbb{R}^3} = \|H_{T_i}^{-1} \cdot \hat{\nabla} \hat{v}_T(\hat{\mathbf{x}})\|_{\mathbb{R}^3}$$

$$\begin{aligned} \|\nabla v_T\|_{\omega_T}^2 &= \sum_{i=0}^k \|\nabla v_T\|_{T_i}^2 = \sum_{i=0}^k 6|T_i| \cdot \|H_{T_i}^{-1} \cdot \hat{\nabla} \hat{v}_T\|_{T_i}^2 \\ &= 6 \cdot \sum_{i=0}^k |T_i| \cdot h_{\min,T_i}^{-2} \cdot \|\text{diag}(\gamma_{1,i}, \gamma_{2,i}, 1) \cdot \hat{\nabla} \hat{v}_T\|_{T_i}^2 \end{aligned}$$

$$\sim h_{\min,T}^{-2} \cdot |T| \cdot \sum_{i=0}^k s_i \sim h_{\min,T}^{-2} \cdot |T| \cdot s$$

$$\text{with } \gamma_{1,i} := \frac{h_{\min,T_i}}{h_{1,T_i}} \quad \text{and} \quad \gamma_{2,i} := \frac{h_{\min,T_i}}{h_{2,T_i}}$$

$$s_i := \|\text{diag}(\gamma_{1,i}, \gamma_{2,i}, 1) \cdot \hat{\nabla} \hat{v}_T\|_{T_i}^2 \geq 0 \quad \text{and} \quad s := \sum_{i=0}^k s_i \geq 0 .$$

Before presenting the proof, we outline it here for a better understanding. To this end we consider  $r$  and  $s$  over a set of certain variables which define them. Then we prove that the maximum of  $r$  and  $s$  is bounded from above, and the minimum is bounded from below and away from 0, respectively.

To accomplish this, several cases have to be considered. For all of them, the general setting is similar, and thus only the first case will be discussed extensively.

**Case 1:  $T$  has no face from  $\Gamma_N \cup \Gamma_R$**

The representation of  $v_T$  is here

$$v_T = \beta_0 \cdot b_T + \sum_{i=1}^k \beta_i \cdot b_{E_i} \quad \beta_i \in \mathbb{R}.$$

The case  $v_T \equiv 0$  is trivial so assume  $v_T \not\equiv 0$ . The functions  $b_T$  and  $b_{E_i}$  are linearly independent. Hence the coefficients  $\beta_i$  can not vanish simultaneously, and without loss of generality we demand  $\sum_{i=0}^k \beta_i^2 = 1$ .

The term  $s_i$  depends on  $\hat{T}_i$ ,  $\gamma_{j,i}$  and  $\hat{v}_T$ , i.e.

$$s_i = s_i(\hat{x}_{2,T_i}, \hat{x}_{3,T_i}, \hat{y}_{3,T_i}, \gamma_{1,i}, \gamma_{2,i}, \beta_0 \dots \beta_k) \quad .$$

The restrictions on  $\hat{T}_i$  and  $T_i$  imply

- $0 < \hat{x}_{2,T_i} \leq 1/2, \quad 0 < \hat{x}_{3,T_i} < 1, \quad -1 < \hat{y}_{3,T_i} < 1,$
- $0 < \gamma_{1,i}, \gamma_{2,i} \leq 1 \quad .$

Now we omit the geometrical meaning that stand behind  $s_i$ . Instead we view  $s_i$  as a purely analytic term that depends on the compact set  $X_i \times G_i \times B$ , with

$$\begin{aligned} X_i &:= \left\{ (\hat{x}_{2,T_i}, \hat{x}_{3,T_i}, \hat{y}_{3,T_i}) : 0 \leq \hat{x}_{2,T_i} \leq 1/2, 0 \leq \hat{x}_{3,T_i} \leq 1, -1 \leq \hat{y}_{3,T_i} \leq 1 \right\} \quad , \\ G_i &:= \left\{ (\gamma_{1,i}, \gamma_{2,i}) : 0 \leq \gamma_{1,i}, \gamma_{2,i} \leq 1 \right\} \quad , \\ B &:= \left\{ (\beta_0, \dots, \beta_k) : \sum_{i=0}^k \beta_i^2 = 1 \right\} \quad . \end{aligned}$$

Note that the set  $X_i \times G_i \times B$  is no longer related to the physical (geometrical) reality. This means that the set  $X_i$  still describes tetrahedra  $\hat{T}_i$ , but not all of them are reference tetrahedra of actual, existing tetrahedra  $T_i$ .

The total sum  $s = s_0 + \dots + s_k$  depends on all  $s_i$  and thus on all  $X_i$ ,  $G_i$  and on  $B$ . Hence  $s$  is analysed over the compact set

$$K := \bigtimes_{i=0}^k X_i \quad \times \quad \bigtimes_{i=0}^k G_i \quad \times \quad B \quad .$$

The terms  $r_i$  and  $r$  depend only on  $\bar{v}_T|_{\hat{T}_i}$  and thus on  $B$ . It is easily verified that  $r$  and  $s$  vary continuously over the compact set  $K$ . Therefore both terms attain their maximum and minimum, respectively. The maximum clearly satisfies

$$\max_K r \sim \max_K s \sim 1 \quad .$$

Also, one has  $r > 0$  if (and only if)  $v_T \not\equiv 0$  which implies  $0 < \min_K r \sim 1$  giving

$$r \sim 1 \quad .$$

In order to show  $s \sim 1 \sim r$  one requires  $0 < \min_K s \sim 1$  which is proven now.

Assume the contrary, i.e. choose that values of  $K$  that yield  $s = s_i = 0$ . Then one has for the outer tetrahedra (i.e.  $i = 1 \dots k$ )

$$0 = s_i = \|\text{diag}(\gamma_{1,i}, \gamma_{2,i}, 1) \cdot \hat{\nabla} \hat{v}_T\|_{\hat{T}_i}^2 \geq \|\partial \hat{v}_T / \partial \hat{z}\|_{\hat{T}_i}^2 = \beta_i^2 \cdot \|\partial \hat{b}_{E_i} / \partial \hat{z}\|_{\hat{T}_i}^2$$

since  $v_T|_{T_i} = \beta_i \cdot b_{E_i}|_{T_i}$ . The last norm is always positive implying  $\beta_i = 0$ ,  $i = 1 \dots k$ . Then  $v_T$  is reduced to  $v_T = \beta_0 \cdot b_T$  giving

$$0 = s_0 = \|\text{diag}(\gamma_{1,0}, \gamma_{2,0}, 1) \cdot \hat{\nabla} \hat{v}_T\|_{\hat{T}_0}^2 \geq \|\partial \hat{v}_T / \partial \hat{z}\|_{\hat{T}_0}^2 = \beta_0^2 \cdot \|\partial \hat{b}_T / \partial \hat{z}\|_{\hat{T}_0}^2 \quad .$$

The last norm is positive again which yields  $\beta_0 = 0$ . This contradicts the assumption  $\sum_{i=0}^k \beta_i^2 = 1$ . Hence  $s > 0$  for all  $v_T \not\equiv 0$  resulting in  $0 < \min_K s \sim 1$ . Thus one has

$$\begin{aligned} s &\sim 1 \sim r \\ \text{or} \quad \|v_T\|_{\omega_T} &\sim h_{\min,T} \cdot \|\nabla v_T\|_{\omega_T} \quad . \end{aligned}$$

**Case 2:  $T$  has exactly one face from  $\Gamma_N \cup \Gamma_R$**

The notation from above implies that  $E_4$  is the one face from  $\partial T \cap (\Gamma_N \cup \Gamma_R)$ . We proceed similar to case 1 and only repeat major steps. The representation of  $v_T$  is either

$$\begin{aligned} v_T &= \beta_0 \cdot b_T + \sum_{i=1}^k \beta_i \cdot b_{E_i} + \beta_4 \cdot b_{E_4} && \text{when } E_4 \subset \Gamma_N \\ \text{or} \quad v_T &= \beta_0 \cdot b_T + \sum_{i=1}^k \beta_i \cdot b_{E_i} + \sum_{j=1}^3 \beta_{4,j} \cdot b_{E_{4,j}} && \text{when } E_4 \subset \Gamma_R \quad . \end{aligned}$$

The set  $B$  is defined analogously such that the sum of all squared coefficients equals 1. Similarly one obtains

$$r \sim 1 \quad \text{and} \quad \max_K s \sim 1 \quad .$$

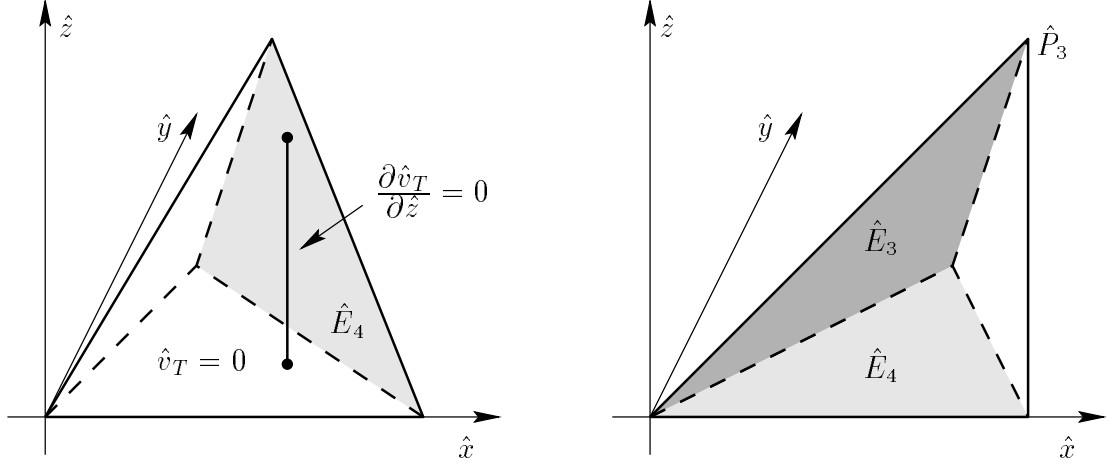
In order to prove  $\min_K s > 0$  assume again the contrary, and choose that values of  $K$  that yield  $s = s_i = 0$ . Start with the outer tetrahedra  $T_i$ ,  $i = 1 \dots k$  and obtain exactly as before  $\beta_i = 0$ ,  $i = 1 \dots k$ . Hence  $v_T$  is reduced to

$$v_T = \beta_0 \cdot b_T + \beta_4 \cdot b_{E_4} \quad \text{or} \quad v_T = \beta_0 \cdot b_T + \sum_{j=1}^3 \beta_{4,j} \cdot b_{E_{4,j}} \quad .$$

Consider now  $\hat{T}$ . Recall  $0 = s_0 \geq \|\partial \hat{v}_T / \partial \hat{z}\|_{\hat{T}_0}^2$  implying  $\partial \hat{v}_T / \partial \hat{z} \equiv 0$ . On the face  $\hat{E}_4$  the representation of  $\hat{v}_T$  is

$$\hat{v}_T = \beta_4 \cdot \hat{b}_{E_4} \quad \text{or} \quad \hat{v}_T = \sum_{j=1}^3 \beta_{4,j} \cdot \hat{b}_{E_{4,j}} \quad .$$

Assume that  $\hat{E}_4$  is not parallel to the  $\hat{z}$  axis. Then every line which is parallel to this  $\hat{z}$  axis, and which goes through  $\hat{E}_4$  also intersects another face  $\hat{E}_i$  on which  $\hat{v}_T = 0$  (cf. left part of figure 3.2). Because of  $\partial \hat{v}_T / \partial \hat{z} \equiv 0$  one has  $\hat{v}_T = 0$  on  $\hat{E}_4$  and  $\beta_4 = 0$  or  $\beta_{4,j} = 0$ ,  $j = 1, 2, 3$ . The case where  $\hat{E}_4$  is parallel to the  $\hat{z}$  axis is dealt with similarly since then  $\hat{v}_T = 0$  on  $\partial \hat{E}_4$

Figure 3.2: Reference tetrahedron  $\hat{T}$  in case 2 (left) and case 3.2 (right)

Hence  $v_T$  is further reduced to  $v_T = \beta_0 \cdot b_T$  from which the contradiction to  $s = s_0 = 0$  follows exactly as in case 1. Therefore  $0 < \min_K s \sim 1$  and  $s \sim 1 \sim r$ .

**Case 3:  $T$  has exactly two faces from  $\Gamma_N \cup \Gamma_R$**

The same notation as before is used which implies that  $E_3$  and  $E_4$  are the two faces from  $\partial T \cap (\Gamma_N \cup \Gamma_R)$ . We proceed similar to case 1 but have to consider sub-cases. The representation of  $v_T$  is here

$$v_T = \beta_0 \cdot b_T + \sum_{i=1}^k \beta_i \cdot b_{E_i} + \begin{cases} \beta_3 \cdot b_{E_3} \\ \sum_{j=1}^3 \beta_{3,j} \cdot b_{E_{3,j}} \end{cases} + \begin{cases} \beta_4 \cdot b_{E_4} \\ \sum_{j=1}^3 \beta_{4,j} \cdot b_{E_{4,j}} \end{cases} \quad \begin{array}{l} \text{if } E_3 \text{ or } E_4 \subset \Gamma_N \\ \text{if } E_3 \text{ or } E_4 \subset \Gamma_R. \end{array}$$

The set  $B$  is defined analogously such that the sum of all squared coefficients equals 1. Similarly one obtains  $r \sim 1$  and  $\max_K s \sim 1$ . In order to prove  $\min_K s > 0$  assume again the contrary, and choose that values of  $K$  that yield  $s = s_i = 0$ . Start with the outer tetrahedra  $T_i$ ,  $i = 1 \dots k$ , and obtain exactly as before  $\beta_i = 0$ ,  $i = 1 \dots k$ . Hence  $v_T$  is reduced to

$$v_T = \beta_0 \cdot b_T + \begin{cases} \beta_3 \cdot b_{E_3} \\ \sum_{j=1}^3 \beta_{3,j} \cdot b_{E_{3,j}} \end{cases} + \begin{cases} \beta_4 \cdot b_{E_4} \\ \sum_{j=1}^3 \beta_{4,j} \cdot b_{E_{4,j}} \end{cases} \quad \begin{array}{l} \text{if } E_3 \subset \Gamma_N \text{ or } E_4 \subset \Gamma_N \\ \text{if } E_3 \subset \Gamma_R \text{ or } E_4 \subset \Gamma_R \end{array}.$$

Consider now  $\hat{T}$ .

**Case 3.1:** The projections of  $\hat{E}_3$  and  $\hat{E}_4$  onto the  $\hat{x}\hat{y}$  plane *differ*.

Then there are lines which are parallel to the  $\hat{z}$  axis and which intersect only one of the faces  $\hat{E}_3$  or  $\hat{E}_4$ , say  $\hat{E}_4$  without loss of generality. Denote the intersection of such lines with  $\hat{E}_4$  by  $\hat{F}$ .

All the lines through  $\hat{F}$  (and which are parallel to the  $\hat{z}$  axis) intersect  $\partial \hat{T}$  in another face ( $\hat{E}_1$  or  $\hat{E}_2$ ) where  $\hat{v}_T = 0$  holds. Analogously to previous cases derive  $\partial \hat{v}_T / \partial \hat{z} \equiv 0$  giving  $\hat{v}_T = 0$  on  $\hat{F}$  and thus on the whole of  $\hat{E}_4$ . This implies  $\beta_4 = 0$  or  $\beta_{4,j} = 0$ ,  $j = 1, 2, 3$ . The remaining representation of  $\hat{v}_T$  coincides with the case 2 which has already been considered.

**Case 3.2:** The projections of  $\hat{E}_3$  and  $\hat{E}_4$  onto the  $\hat{x}\hat{y}$  plane *coincide*.

This corresponds to  $\hat{P}_3$  being one of the points  $(0, 0, 1)^T$ ,  $(1, 0, 1)^T$  or  $(\hat{x}_2, 1, 1)^T$ . Furthermore,  $\hat{P}_0\hat{P}_1\hat{P}_2$  is either  $\hat{E}_3$  or  $\hat{E}_4$ . The right part of figure 3.2 may illustrate such a situation. Here, however, we cannot proceed exactly as in cases 1 and 2 since now there exists a  $\hat{v}_T \neq 0$  with  $\partial\hat{v}_T/\partial\hat{z} \equiv 0$ . Hence a refined analysis becomes necessary.

The angle between  $E_3$  and  $E_4$  is the angle between  $\Gamma_N$  faces or  $\Gamma_R$  faces of  $\Omega$ , i.e. it depends on  $\Omega$  but not on  $\mathcal{T}_h$  or  $T$ . Basic geometry implies  $h_{2,T} \sim h_{3,T}$  and therefore  $\gamma_{2,0} \sim 1$ . A closer investigation of  $s_0$  yields

$$0 = s_0 = \|\text{diag}(\gamma_{1,0}, \gamma_{2,0}, 1) \cdot \hat{\nabla}\hat{v}_T\|_{\hat{T}_0}^2 \gtrsim \|\partial\hat{v}_T/\partial\hat{y}\|_{\hat{T}_0}^2 + \|\partial\hat{v}_T/\partial\hat{z}\|_{\hat{T}_0}^2$$

giving  $\partial\hat{v}_T/\partial\hat{y} = \partial\hat{v}_T/\partial\hat{z} \equiv 0$  in  $\hat{T}_0$ .

Similar to above, consider now  $\hat{E}_1$  and  $\hat{E}_2$  and all edges of  $T$ , and recall that  $v_T = 0$  on them. This leads easily to  $v_T \equiv 0$  on the whole of  $T$  which contradicts the assumption. Hence  $0 < \min_K s \sim 1$  and  $s \sim 1 \sim r$ .

**Case 4:**  $T$  has exactly three faces from  $\Gamma_N \cup \Gamma_R$

Start exactly like in cases 1, 2 and 3 and obtain  $\beta_i = 0$ ,  $i = 1 \dots k$ , and  $\partial\hat{v}_T/\partial\hat{z} \equiv 0$  on  $\hat{T}$ . Hence  $v_T$  is reduced to

$$v_T = \beta_0 \cdot b_T + \left\{ \begin{array}{l} \beta_2 \cdot b_{E_2} \\ \sum_{j=1}^3 \beta_{2,j} \cdot b_{E_{2,j}} \end{array} \right\} + \left\{ \begin{array}{l} \beta_3 \cdot b_{E_3} \\ \sum_{j=1}^3 \beta_{3,j} \cdot b_{E_{3,j}} \end{array} \right\} + \left\{ \begin{array}{l} \beta_4 \cdot b_{E_4} \\ \sum_{j=1}^3 \beta_{4,j} \cdot b_{E_{4,j}} \end{array} \right\}$$

depending on whether  $E_2, E_3, E_4$  are from  $\Gamma_N$  or  $\Gamma_R$ . Consider now  $\hat{T}$ . One has  $v_T = 0$  on  $E_1$ .

If  $\hat{E}_1$  is *not parallel* to the  $\hat{z}$  axis then we proceed analogously to case 3.1 and prove that  $\hat{v}_T = 0$  on another face of  $\hat{T}$ . Hence  $v_T$  is further reduced, and case 3 applies.

Otherwise  $\hat{E}_1$  is *parallel* to the  $\hat{z}$  axis and perpendicular to the  $\hat{x}\hat{y}$  plane. The angles between any two of the faces  $E_2, E_3, E_4$  are angles between  $\Gamma_N$  faces or  $\Gamma_R$  faces of  $\Omega$ , i.e. they depend on  $\Omega$  but not on  $\mathcal{T}_h$  or  $T$ . Basic geometry yields then  $h_{2,T} \sim h_{3,T}$  which, similarly to case 3.2, leads to  $\partial\hat{v}_T/\partial\hat{y} = \partial\hat{v}_T/\partial\hat{z} \equiv 0$  in  $\hat{T}$  and  $v_T \equiv 0$  in  $T$ . This contradicts the assumption and proves  $0 < \min_K s \sim 1$  and  $s \sim 1 \sim r$ .

After the proof of the first relation of Lemma 3.5 we now turn to inequality (3.36). Let  $E$  be an arbitrary face of  $T$ . It is easily verified that

$$\|\bar{v}_T\|_E \lesssim \|\bar{v}_T\|_{\hat{T}} \quad \forall v_T \in V_T$$

holds on the standard tetrahedron  $T$  because  $\|\bar{v}_T\|_E$  is a seminorm over the finite dimensional space  $V_{\hat{T}}$  of linearly independent functions, whereas the right-hand side is a norm over the same space. Hence

$$\|v_T\|_E^2 = 2|E| \cdot \|\bar{v}_T\|_E \lesssim |E| \cdot \|\bar{v}_T\|_{\hat{T}} \sim \frac{|E|}{|T|} \cdot \|v_T\|_T^2 \sim h_E^{-1} \cdot h_{\min,T}^2 \cdot \|\nabla v_T\|_{\omega_T}^2$$

which proves Lemma 3.5. ■

**Remark 3.9** Note that in (3.35) the norm over the whole domain  $\omega_T$  is essential. In particular the inequality

$$\|v_T\|_T \lesssim h_{\min,T} \cdot \|\nabla v_T\|_T \quad \forall v_T \in V_T$$

does *not* hold.

Consider for example a tetrahedron  $T$  with vertices  $P_0 = \mathbf{0}$ ,  $P_1 = \mathbf{e}_1$ ,  $P_2 = \mathbf{e}_2$ , and  $P_3 = h \cdot \mathbf{e}_3$  with  $h \rightarrow 0$ . Choose  $v_T := b_{E_1} + b_{E_2}$ , with  $b_{E_1} = 27xy(1 - x - y - z/h)$  and  $b_{E_2} = 27xyz/h$  being two face bubble functions. Then

$$\begin{aligned} \|v_T\|_T &= \sqrt{27/560} \cdot h^{1/2} \\ \text{and} \quad h_{\min,T} \cdot \|\nabla v_T\|_T &= \sqrt{81/35} \cdot \frac{h^{3/2}}{\sqrt{1+2h^2}} \quad . \end{aligned}$$

Thus the abovementioned inequality does not hold with a multiplicative constant independent of  $h$ . Note also that the corresponding isotropic estimates are much easier to derive.  $\square$

**Remark 3.10** Assume that  $\partial\Omega$  consists solely of the Dirichlet boundary  $\Gamma_D$ . Assume further that the domain  $\omega_T$  is contained in a rectangular prism with minimal side length  $l \sim h_{\min,T}$ . Then the  $\lesssim$  part of (3.35) coincides with the Friedrichs inequality. But we have not shown such a geometrical condition and thus had to proceed in the way described above.  $\square$

### 3.3.3 Equivalence and bounds of the local problem error estimator

Set

$$\eta_{R,\omega_T}^2 := \sum_{T' \subset \omega_T} \eta_{R,T'}^2 \quad \text{and} \quad \eta_{D,\omega_T}^2 := \sum_{T' \subset \omega_T} \eta_{D,T'}^2 \quad . \quad (3.37)$$

**Theorem 3.6 (Equivalence with the residual error estimator)** *The local problem error estimator is equivalent to the residual error estimator  $\eta_{R,T}$  in the following sense:*

$$\eta_{D,T} \lesssim \eta_{R,\omega_T} \quad (3.38)$$

$$\eta_{R,T} \lesssim \eta_{D,\omega_T} \quad . \quad (3.39)$$

If  $T$  has a boundary face then the constant in (3.38) depends on the interior angles of the corners and edges of the domain  $\Omega$ .

**Proof:** Let  $T$  be an arbitrary but fixed tetrahedron throughout the proof.

For the first inequality recall the definition (3.32) of  $\eta_{D,T}$ . The residuals  $r_T(u_h)$  and  $r_E(u_h)$  are given in definition 3.3 on page 45. Furthermore, take into account how  $\omega_T$  and  $V_T$  are modified if  $T$  has a boundary face, and observe in particular that  $e_T = 0$  on  $\partial\omega_T \setminus \partial T$ . By integration by parts one obtains

$$\begin{aligned} \eta_{D,T}^2 &= \|e_T\|_{\omega_T}^2 = a(e_T, e_T) \\ &\stackrel{(3.34)}{=} \sum_{T' \in \omega_T} \int_{T'} r_{T'} \cdot e_T + \sum_{E \subset \partial T \setminus \Gamma_D} \int_E r_E \cdot e_T \\ &\leq \left( \sum_{T' \subset \omega_T} \|r_{T'}(u_h)\|_{T'}^2 \right)^{1/2} \cdot \|e_T\|_{\omega_T} + \sum_{E \subset \partial T \setminus \Gamma_D} \|r_E(u_h)\|_E \cdot \|e_T\|_E \quad . \end{aligned}$$

Now  $\|e_T\|_{\omega_T}$  and  $\|e_T\|_E$ ,  $E \subset \partial T \setminus (\Gamma_D \cup \Gamma_R)$ , are bounded using Lemma 3.5 on page 56. For  $E \subset \partial T \cap \Gamma_R$ , the norm  $\|e_T\|_E$  requires a refined analysis. This yields

$$\begin{aligned}
 & \left. \begin{aligned}
 & E \subset \partial T \setminus (\Gamma_D \cup \Gamma_R): \quad \begin{aligned}
 & \|e_T\|_{\omega_T} \stackrel{(3.35)}{\lesssim} h_{min,T} \cdot \|\nabla e_T\|_{\omega_T} \leq h_{min,T} \cdot \|e_T\|_{\omega_T}, \\
 & \|e_T\|_E \stackrel{(3.36)}{\lesssim} \frac{h_{min,T}}{h_E^{1/2}} \cdot \|\nabla e_T\|_{\omega_T} \leq \frac{h_{min,T}}{h_E^{1/2}} \cdot \|e_T\|_{\omega_T},
 \end{aligned} \\
 & E \subset \partial T \cap \Gamma_R: \quad \begin{aligned}
 & \|e_T\|_E \leq \alpha^{-1/2} \cdot \|e_T\|_{\omega_T}, \\
 & \|e_T\|_E \stackrel{(3.36)}{\lesssim} \frac{h_{min,T}}{h_E^{1/2}} \cdot \|\nabla e_T\|_{\omega_T} \leq \frac{h_{min,T}}{h_E^{1/2}} \cdot \|e_T\|_{\omega_T},
 \end{aligned} \\
 & \implies \|e_T\|_E \lesssim \min \left\{ \alpha^{-1/2}, \frac{h_{min,T}}{h_E^{1/2}} \right\} \cdot \|e_T\|_{\omega_T}.
 \end{aligned} \right\} \quad (3.40)
 \end{aligned}$$

(Note that the constants appearing here depend on  $\Omega$  if  $T$  has a boundary face). Inserting these inequalities results in

$$\begin{aligned}
 \eta_{D,T}^2 & \lesssim h_{min,T} \cdot \left( \sum_{T' \subset \omega_T} \|r_{T'}(u_h)\|_{T'} + \sum_{E \subset \partial T \setminus (\Gamma_D \cup \Gamma_R)} h_E^{-1/2} \|r_E(u_h)\|_E + \right. \\
 & \quad \left. + \sum_{E \subset \partial T \cap \Gamma_R} \min \left\{ 1, \frac{h_E^{1/2}}{\alpha^{1/2} h_{min,T}} \right\} \cdot h_E^{-1/2} \cdot \|r_E(u_h)\|_E \right) \cdot \|e_T\|_{\omega_T}
 \end{aligned}$$

which, together with  $\|e_T\|_{\omega_T} = \eta_{D,T}$ , proves the desired inequality (3.38)

$$\eta_{D,T} \lesssim \sum_{T' \subset \omega_T} \eta_{R,T'} \quad .$$

For the proof of the second inequality we require bounds of  $\eta_{R,T}$ , and thus of  $\|r_T\|_T$  and  $\|r_E\|_E$ . For reasons that will become clear below, we first bound the term  $\|r_{T'}\|_{T'}$ , with  $T' \subset \omega_T$  being an arbitrary tetrahedron. Recall definition (2.19) of the bubble function  $b_{T'}$  and set  $v_{T'} := b_{T'} \cdot r_{T'}$ . Then  $b_{T'}$  and  $v_{T'}$  belong to the finite element space  $V_{T'}$  (and not to  $V_T$  if  $T' \neq T$ ). Hence the local problem related to  $T'$  has to be invoked.

For the derivation the local problem (3.31) will be utilized. Note that some terms in that formulation vanish since  $v_{T'} = 0$  on  $\partial T'$ . The equivalence (2.24) and integration by parts imply

$$\begin{aligned}
 \|r_{T'}\|_{T'}^2 & \sim \|b_{T'}^{1/2} \cdot r_{T'}\|_{T'}^2 = \int_{T'} r_{T'} \cdot v_{T'} \\
 & = \sum_{T'' \subset \omega_{T'}} \int_{T''} (P_1 f + \Delta u_h) \cdot v_{T'} \quad \text{since } v_{T'} \in H_o^1(T') \\
 & = \int_{\omega_{T'}} P_1 f \cdot v_{T'} - \int_{\omega_{T'}} \nabla^T u_h \cdot \nabla v_{T'} \\
 & \stackrel{(3.31)}{=} a(e_{T'}, v_{T'}) \leq \|e_{T'}\|_{T'} \cdot \|v_{T'}\|_{T'} \quad ,
 \end{aligned}$$

where  $e_{T'} \in V_{T'}$  denotes the solution of the local problem over  $\omega_{T'}$ . The equivalence (2.25) and  $v_{T'} = 0$  on  $\partial T'$  result in

$$\|v_{T'}\|_{T'} = \|\nabla v_{T'}\|_{T'} = \|\nabla(b_{T'} \cdot r_{T'})\|_{T'} \stackrel{(2.25)}{\sim} h_{min,T'}^{-1} \cdot \|r_{T'}\|_{T'} \quad .$$

Combining both inequalities yields

$$\|r_{T'}\|_{T'} \lesssim h_{min,T}^{-1} \cdot \|e_{T'}\|_{T'} \leq h_{min,T}^{-1} \cdot \eta_{D,T'} \quad \forall T' \in \omega_T$$

since  $h_{min,T'}$  does not change rapidly across adjacent tetrahedra  $T'$ .

The norm of  $r_E \in \mathbb{P}^0(E)$  for an interior face  $E \subset \partial T \setminus \Gamma$  is bounded similarly. Recall the definitions (2.20) and (2.23) of the bubble function  $b_E$  and the extension operator  $F_{ext}$ , respectively, and set  $v_E := b_E \cdot F_{ext}(r_E) \in V_T \cap H_o^1(\omega_E)$ . Equivalence (2.26),  $v_E = 0$  on  $\partial T \cap \Gamma$ , and integration by parts imply

$$\begin{aligned} \|r_E\|_E^2 &\sim \|b_E^{1/2} \cdot r_E\|_E^2 = \int_E r_E \cdot v_E = - \sum_{T' \subset \omega_E} \int_{\partial T'} \frac{\partial u_h}{\partial n} \cdot v_E \\ &= - \sum_{T' \subset \omega_E} \int_{T'} \Delta u_h \cdot v_E - \int_{\omega_E} \nabla^T u_h \cdot \nabla v_E \\ &\stackrel{(3.31)}{=} - \sum_{T' \subset \omega_E} \int_{T'} \Delta u_h \cdot v_E - \int_{\omega_E} P_1 f \cdot v_E + a(e_T, v_E) \\ &\leq \sum_{T' \subset \omega_E} \|r_{T'}\|_{T'} \cdot \|v_E\|_{T'} + \|e_T\|_{\omega_E} \cdot \|v_E\|_{\omega_E} \quad . \end{aligned}$$

The equivalences (2.27) and (2.28) imply

$$\begin{aligned} \|v_E\|_{T'} &= \|b_E \cdot F_{ext}(r_E)\|_{T'} \sim h_E^{1/2} \cdot \|r_E\|_E \\ \|v_E\|_{\omega_E} &= \|\nabla v_E\|_{\omega_E} = \|\nabla(b_E \cdot F_{ext}(r_E))\|_{\omega_E} \sim h_E^{1/2} \cdot h_{min,T}^{-1} \cdot \|r_E\|_E \quad . \end{aligned}$$

Now the previous bound of  $\|r_{T'}\|_{T'}$  is required for all tetrahedra  $T' \subset \omega_E$ . Combining all estimates yields

$$\|r_E\|_E \lesssim h_E^{1/2} \cdot h_{min,T}^{-1} \cdot \sum_{T' \subset \omega_E} \eta_{D,T'} \quad \forall E \subset \partial T \setminus \Gamma \quad .$$

The norm of  $r_E \in \mathbb{P}^0(E)$  for a Neumann boundary face  $E \subset \partial T \cap \Gamma_N$  is bounded analogously. Set  $v_E := b_E \cdot F_{ext}(r_E) \in V_T$ , recall  $v_E = 0$  on  $\partial T \setminus E$ , and derive

$$\begin{aligned} \|r_E\|_E^2 &\sim \|b_E^{1/2} \cdot r_E\|_E^2 = \int_E r_E \cdot v_E = \int_E \left( P_2 g_N - \frac{\partial u_h}{\partial n} \right) \cdot v_E \\ &= \int_E P_2 g_N \cdot v_E - \int_T \nabla^T u_h \cdot \nabla v_E - \int_T \Delta u_h \cdot v_E \\ &\stackrel{(3.31)}{=} a(e_T, v_E) - \int_T r_T \cdot v_E \leq \|e_T\|_T \cdot \|v_E\|_T + \|r_T\|_T \cdot \|v_E\|_T \quad . \end{aligned}$$

The norms  $\|v_E\|_T$ ,  $\|v_E\|_T$  and  $\|r_T\|_T$  are bounded as before giving

$$\|r_E\|_E \lesssim h_E^{1/2} \cdot h_{min,T}^{-1} \cdot \|e_T\|_T \leq h_E^{1/2} \cdot h_{min,T}^{-1} \cdot \eta_{D,T} \quad \forall E \subset \partial T \cap \Gamma_N \quad .$$

Finally the norm of linear residual  $r_E \in \mathbb{P}^1(E)$  for a Robin boundary face  $E \subset \partial T \cap \Gamma_R$  is to be bounded. Proceed in analogy and set  $v_E := b_E \cdot F_{ext}(r_E) \in V_E \subset V_T$ . Integration

by parts and the local problem (3.31) yield

$$\begin{aligned}
\|r_E\|_E^2 &\sim \|b_E^{1/2} \cdot r_E\|_E^2 = \int_E r_E \cdot v_E = \int_E \left( \alpha \cdot (P_3 g_R - u_h) - \frac{\partial u_h}{\partial n} \right) \cdot v_E \\
&= \int_E \alpha \cdot (P_3 g_R - u_h) \cdot v_E - \int_T \nabla^T u_h \cdot \nabla v_E - \int_T \Delta u_h \cdot v_E \\
&\stackrel{(3.31)}{=} a(e_T, v_E) - \int_T r_T \cdot v_E \leq \|e_T\|_T \cdot \|v_E\|_T + \|r_T\|_T \cdot \|v_E\|_T \quad .
\end{aligned}$$

The terms  $\|v_E\|_T$  and  $\|r_T\|_T$  are bounded as before whereas for  $\|v_E\|_T^2$  the equivalence (2.28) is invoked implying

$$\begin{aligned}
\|v_E\|_T^2 &= \|\nabla(F_{ext}(r_E) \cdot b_E)\|_T^2 + \alpha \cdot \|r_E \cdot b_E\|_E^2 \\
&\lesssim h_E h_{min,T}^{-2} \cdot \|r_E\|_E^2 + \max \left\{ \alpha, \frac{h_E}{h_{min,T}^2} \right\} \cdot \|r_E\|_E^2 \\
&\lesssim h_E h_{min,T}^{-2} \cdot \max \left\{ 1, \frac{\alpha h_{min,T}^2}{h_E} \right\} \cdot \|r_E\|_E^2
\end{aligned}$$

and resulting in

$$\min \left\{ 1, \frac{h_E}{\alpha h_{min,T}^2} \right\} \cdot \frac{h_{min,T}^2}{h_E} \cdot \|r_E\|_E^2 \lesssim \|e_T\|_T^2 \leq \eta_{D,T}^2 \quad \forall E \subset \partial T \cap \Gamma_R \quad .$$

Collecting all the results for  $\|r_T\|_T$  and  $\|r_E\|_E$  and inserting them into the definition of  $\eta_{R,T}$  gives the desired estimate (3.39).  $\blacksquare$

With the help of the previous Theorem 3.6 we easily derive the main result of this Section, namely upper and lower error bounds by means of the local problem error estimator.

**Theorem 3.7 (Local problem error estimation)** *The lower bound of the error is*

$$\eta_{D,T}(u_h) \leq \|u - u_h\|_{\omega_T} + c \cdot \zeta_T \quad . \quad (3.41)$$

*The error is bounded globally from above by*

$$\|u - u_h\| \lesssim m_1(u - u_h, \mathcal{T}_h) \cdot \left[ \eta_D^2(u_h) + \zeta^2 \right]^{1/2} \quad . \quad (3.42)$$

*If  $T$  has a boundary face then the constant in (3.41) depends on the interior angles of the corners and edges of the domain  $\Omega$ .*

**Proof:** In order to prove the lower bound of the error, apply the formulation (3.33) of the local problem, recall how  $\omega_T$  and  $V_T$  are modified if  $T$  has a boundary face, and observe in particular that  $e_T = 0$  on  $\partial\omega_T \setminus \partial T$ . Together with integration by parts one obtains

$$\begin{aligned}
\eta_{D,T}^2 &= \|e_T\|_{\omega_T}^2 = a(e_T, e_T) \\
&\stackrel{(3.33)}{=} a(u - u_h, e_T) + \int_{\omega_T} (P_1 f - f) \cdot e_T + \int_{\Gamma_N \cap \partial\omega_T} (P_2 g_N - g_N) \cdot e_T + \\
&\quad + \int_{\Gamma_R \cap \partial\omega_T} \alpha \cdot (P_3 g_R - g_R) \cdot e_T \\
&\leq \|u - u_h\|_{\omega_T} \cdot \|e_T\|_{\omega_T} + \|f - P_1 f\|_{\omega_T} \cdot \|e_T\|_{\omega_T} + \\
&\quad + \|g_N - P_2 g_N\|_{\Gamma_N \cap \partial T} \cdot \|e_T\|_{\Gamma_N \cap \partial T} + \|\alpha \cdot (g_R - P_3 g_R)\|_{\Gamma_R \cap \partial T} \cdot \|e_T\|_{\Gamma_R \cap \partial T}.
\end{aligned}$$

By means of the previous bounds from (3.40) one readily obtains the desired estimate (3.41). Note that here the only constant appears at the approximation term.

Finally inequality (3.42) follows immediately from the error bound (3.23) of the residual error estimator, and from the equivalence (3.39) of  $\eta_{R,T}$  and  $\eta_{D,T}$ .  $\blacksquare$

### 3.3.4 Condition number of the finite element matrix of the local problem

The error estimator  $\eta_{D,T}$  requires the solution of a local variational problem. We will employ a finite element method with a proper choice of the ansatz and test functions. It can be shown that then the variational problem is well-behaved, i.e. the condition number of the corresponding finite element matrix is bounded independently of the aspect ratio of the elements under consideration.

Basically, the ansatz and test functions will be the bubble functions that span the space  $V_T$ , with a modification for the Robin face bubble functions. For a precise definition enumerate here the faces of  $T$  such that  $E_1 \dots E_k$  denote all interior faces and Neumann faces,  $E_{k+1} \dots E_m$  denote the Robin faces, and finally  $E_{m+1} \dots E_4$  denote the Dirichlet faces,  $1 \leq k \leq m \leq 4$ . Define the row vector  $\Phi$  of ansatz and test functions by

$$\Phi := \{b_T, b_{E_1} \dots b_{E_k}, \tilde{b}_{E_{k+1},1} \dots \tilde{b}_{E_{k+1},3}, \dots, \tilde{b}_{E_m,1} \dots \tilde{b}_{E_m,3}\}$$

$$\text{with } \tilde{b}_{E_i,j} := \min \left\{ 1, \frac{h_{E_i}^{1/2}}{\alpha^{1/2} h_{\min,T}} \right\} \cdot b_{E_i,j},$$

i.e. the ansatz functions related to a Robin face are simply the scaled bubble functions  $b_{E_i,j}$  from (3.29). Clearly the functions from  $\Phi$  span the space  $V_T$ . An arbitrary function  $v_T \in V_T$  can then be written as

$$v_T = \beta_0 \cdot b_T + \sum_{i=1}^k \beta_i \cdot b_{E_i} + \sum_{i=k+1}^m \sum_{j=1}^3 \beta_{i,j} \cdot \tilde{b}_{E_i,j} \quad \beta_i, \beta_{i,j} \in \mathbb{R} \quad .$$

For the remainder of this Section, define the vector

$$\mathbf{v} := (\beta_0, \beta_1 \dots \beta_k, \beta_{k+1,1} \dots \beta_{k+1,3}, \dots, \beta_{m,1} \dots \beta_{m,3})^T \in \mathbb{R}^{1+k+3(m-k)}.$$

By means of the finite element isomorphism

$$v_T = \Phi \cdot \mathbf{v} \in V_T \quad \longleftrightarrow \quad \mathbf{v} \in \mathbb{R}^{1+3m-2k}$$

one obtains

$$a(v_T, w_T) = (K_T \mathbf{v}, \mathbf{w}) \quad \forall w_T = \Phi \cdot \mathbf{w} \in V_T \quad .$$

Here  $K_T \in \mathbb{R}^{(1+3m-2k) \times (1+3m-2k)}$  is the usual finite element stiffness matrix which is symmetric and positive definite.

**Theorem 3.8 (Condition number)** *The condition number  $\kappa(K_T)$  of the local problem stiffness matrix  $K_T$  is bounded independently of  $T$ :*

$$\kappa(K_T) \sim 1 \quad \forall T \in \mathcal{T}_h \quad .$$

**Proof:** The condition number is given by

$$\kappa(K_T) = \frac{\max_{\mathbf{v} \neq \mathbf{0}} (K_T \mathbf{v}, \mathbf{v}) / (\mathbf{v}, \mathbf{v})}{\min_{\mathbf{w} \neq \mathbf{0}} (K_T \mathbf{w}, \mathbf{w}) / (\mathbf{w}, \mathbf{w})} .$$

The scalar product  $(K_T \mathbf{v}, \mathbf{v})$  equals

$$(K_T \mathbf{v}, \mathbf{v}) = a(v_T, v_T) = \|v_T\|_{\omega_T}^2 = \|\nabla v_T\|_{\omega_T}^2 + \sum_{i=k+1}^m \alpha \cdot \|v_T\|_{E_i}^2 .$$

Both norms are now investigated separately.

Starting with  $\|\nabla v_T\|_{\omega_T}^2$ , equivalence (3.35) of Lemma 3.5 states

$$\|\nabla v_T\|_{\omega_T}^2 \sim h_{min,T}^{-2} \cdot \|v_T\|_{\omega_T}^2 .$$

Apply now the same techniques as in the proof of Lemma 3.5 to show easily

$$\begin{aligned} \|v_T\|_{\omega_T}^2 &= \sum_{T' \subset \omega_T} \|v_T\|_{T'}^2 \sim |T| \cdot \sum_{T' \subset \omega_T} \|\bar{v}_T\|_{T'}^2 \\ &\sim |T| \cdot \sum_{T' \subset \omega_T} \left\| \beta_0 \bar{b}_T + \sum_{i=1}^k \beta_i \bar{b}_{E_i} + \sum_{i=k+1}^m \sum_{j=1}^3 \beta_{i,j} \cdot \min \left\{ 1, \frac{h_{E_i}^{1/2}}{\alpha^{1/2} h_{min,T}} \right\} \bar{b}_{E_{i,j}} \right\|_{T'}^2 \\ &\sim |T| \cdot \left( \sum_{i=0}^k \beta_i^2 + \sum_{i=k+1}^m \sum_{j=1}^3 \min \left\{ 1, \frac{h_{E_i}}{\alpha h_{min,T}^2} \right\} \cdot \beta_{i,j}^2 \right) . \end{aligned}$$

For a Robin boundary face  $E_i \subset \Gamma_R \cap \partial T$  utilize standard scaling arguments (for the standard face) and recall the representation of  $v_T$  and the definition of  $\bar{b}_{E_{i,j}}$  to prove

$$\begin{aligned} \alpha \cdot \|v_T\|_{E_i}^2 &= \alpha \cdot \left\| \sum_{j=1}^3 \beta_{i,j} \cdot \min \left\{ 1, \frac{h_{E_i}^{1/2}}{\alpha^{1/2} h_{min,T}} \right\} \cdot b_{E_{i,j}} \right\|_{E_i}^2 \\ &= \min \left\{ \alpha, \frac{h_{E_i}}{h_{min,T}^2} \right\} \cdot \left\| \sum_{j=1}^3 \beta_{i,j} \cdot b_{E_{i,j}} \right\|_{E_i}^2 \\ &\sim \min \left\{ \alpha, \frac{h_{E_i}}{h_{min,T}^2} \right\} \cdot |E_i| \cdot \sum_{j=1}^3 \beta_{i,j}^2 \\ &\sim h_{min,T}^{-2} \cdot |T| \cdot \min \left\{ \frac{\alpha h_{min,T}^2}{h_{E_i}}, 1 \right\} \cdot \sum_{j=1}^3 \beta_{i,j}^2 . \end{aligned}$$

Combining all relations yields

$$\begin{aligned} (K_T \mathbf{v}, \mathbf{v}) &\sim \frac{|T|}{h_{min,T}^2} \cdot \left( \sum_{i=0}^k \beta_i^2 + \right. \\ &\quad \left. + \sum_{i=k+1}^m \sum_{j=1}^3 \left[ \min \left\{ 1, \frac{h_{E_i}}{\alpha h_{min,T}^2} \right\} + \min \left\{ \frac{\alpha h_{min,T}^2}{h_{E_i}}, 1 \right\} \right] \cdot \beta_{i,j}^2 \right) \\ &\sim h_{min,T}^{-2} \cdot |T| \cdot \left( \sum_{i=0}^k \beta_i^2 + \sum_{i=k+1}^m \sum_{j=1}^3 \beta_{i,j}^2 \right) = h_{min,T}^{-2} \cdot |T| \cdot \|\mathbf{v}\|_{\mathbb{R}^{1+3m-2k}}^2 \end{aligned}$$

which gives the assertion. ■

### 3.3.5 Comments and Remarks

#### Remark 3.11 (Robin boundary conditions)

Up to now Robin boundary conditions have, to our knowledge, not been incorporated into *a posteriori* error estimators. Hence our choice and analysis of the Robin boundary residual seems to be completely new, even in an isotropic context.

In the literature the treatment of Robin boundary conditions for *a posteriori* error estimators is often neglected. Partly this might be due to the fact that they do not occur in elasticity problems but mainly in heat conduction problems. Also, there seems to be a general opinion that Robin boundary conditions can be treated analogously to Neumann boundary conditions. Our analysis, however, shows that this is only partly true.

Consider a family of problems where the exchange parameter  $\alpha$  of the Robin boundary condition may vary. In an *isotropic* context, the norm of a Robin face residual is multiplied by an additional weight

$$\min \left\{ 1, \frac{h_E}{\alpha h_{min,T}^2} \right\} \sim \min\{1, \alpha^{-1} \cdot h_T^{-1}\} \quad (\text{isotropic case}) \quad ,$$

cf. definition (3.19) of the error estimator. Hence the correct choice of the error estimator does not only depend on the differential equation but also on the actual discretization  $\mathcal{T}_h$ . When  $\alpha \cdot h_T > 1$  then the Robin residual has to be scaled in the way described in the previous Sections.

Finally, let us discuss the different interpretations of the Robin boundary condition for the two cases where  $\alpha \not\approx 1$ . When  $\alpha$  is small, i.e.  $\alpha \ll 1$ , then the Robin boundary condition represents almost a homogeneous Neumann boundary condition, i.e.  $\partial u_h / \partial n \approx \alpha \cdot (g_R - u_h) \approx 0$ . Hence one would expect that the representation of the Robin residual and the Neumann residual is similar in the residual error estimator.

If, however,  $\alpha$  becomes large, i.e.  $\alpha \gg 1$ , then the Robin boundary condition approaches an (inhomogeneous) Dirichlet boundary condition, i.e.  $g_R - u_h \approx \alpha^{-1} \cdot \partial u_h / \partial n \approx 0$ . Therefore the Robin residual should almost vanish in the residual error estimator.

Indeed, when comparing both points of view with the residual error estimator presented, one experiences that the scaling factor  $\min\{1, h_E \cdot \alpha^{-1} h_{min,T}^{-2}\}$  of the Robin residual  $\|r_E\|_E$  behaves exactly as described above.  $\square$

#### Remark 3.12 (Higher order ansatz functions in $V_{o,h}$ )

Often one is interested not only in the use of piecewise linear ansatz functions in the finite element method but also in quadratic (or even higher) ansatz functions. Here we will briefly discuss how the theory of the residual error estimator and the local problem error estimator has to be modified to accommodate to a higher order ansatz space  $V_{o,h}$ . Assume for the moment that elements of order  $m$  are used,  $m \geq 1$ .

Let us start with the modifications which would become necessary for the *residual error estimator*.

1. The equivalences (or inverse inequalities in their original form) of Lemma 2.7 on page 27 have to hold for

$$\varphi_T \in \mathbb{P}^{m-2}(T) \quad \text{and} \quad \varphi_E \in \mathbb{P}^m(E) \quad .$$

If no Robin boundary exists then  $\varphi_E \in \mathbb{P}^{m-1}(E)$  suffices.

2. The extension operator  $F_{ext}$  of (2.23) has to act from the space  $\mathbb{P}^m(E) \ni r_E$ . If no Robin boundary exists then the mapping from the space  $\mathbb{P}^{m-1}(E) \ni r_E$  is sufficient.
3. Higher order approximation operators instead of  $P_1, P_2, P_3$  may be useful in order to make the approximation term  $\zeta_T$  small in comparison to the error estimators  $\eta_{R,T}$  or  $\eta_{D,T}$ , respectively.

For the *local problem error estimator* we now list the modifications which are necessary to prove the equivalence to the residual error estimator.

1. All points mentioned for the residual error estimator (i.e. extension operator and inverse inequalities) are here required too.
2. The local finite element space  $V_T$  of (3.30) has to contain the functions
  - $r_T \cdot b_T$
  - $F_{ext}(r_E) \cdot b_E, \quad \forall E \subset \partial T \setminus \Gamma_D$  .

Thus  $V_T$  has to be enhanced.

3. The essential Lemma 3.5 has to be proven for the enhanced space  $V_T$ .

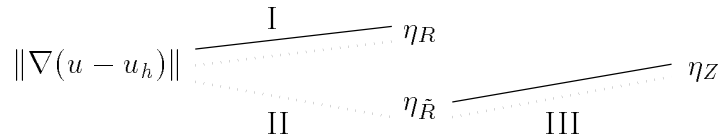
We may note here that some of the abovementioned modifications and effects can already be studied when comparing the (constant) residual  $r_E \in \mathbb{P}^0(E)$  on a Neumann boundary and the (linear) residual  $r_E \in \mathbb{P}^1(E)$  on a Robin boundary. One can fairly easily identify that parts of the proofs where the Robin residual requires a modified analysis.  $\square$

## 3.4 Zienkiewicz-Zhu like error estimators

### 3.4.1 Introduction

Error estimators based on gradient recovery have become quite popular in recent years. First ideas go back to Zienkiewicz and Zhu [66] and utilize an averaged gradient. Later the estimator has been improved by the ‘superconvergent patch recovery’ [67]. We extend the first error estimator to anisotropic meshes. Although we did not investigate the second estimator, we think that the superconvergence analysis on anisotropic meshes may become a fruitful field for further reliable and robust error estimators.

The analysis of the recovered gradient Zienkiewicz-Zhu (ZZ) error estimator (temporarily denoted by  $\eta_Z$ ) heavily relies on a *modified* residual error estimator  $\eta_{\tilde{R}}$ . This  $\eta_{\tilde{R}}$  is obtained from the usual residual error estimator  $\eta_R$  by omitting the element residual  $r_T$ , and keeping only the face residual  $r_E$  (cf. [44, 46]). The following figure visualizes the present state of the analysis. It shows which equivalences are proven for *anisotropic* meshes (depicted by —) and for *isotropic* meshes (depicted by  $\cdots$ ). Details are described afterwards.



**I:** The relation between the error and  $\eta_R$  is quite well understood for *isotropic* meshes. For *anisotropic* tetrahedral/triangular meshes, it is given in Theorem 3.4 on page 46 of our work.

**II:** Only recently Carstensen and Verfürth [23] have filled the gap between  $\|\nabla(u - u_h)\|$  and the modified residual error estimator  $\eta_{\tilde{R}}$  for triangular isotropic meshes. Their proof requires an additional but weak mesh assumption. A corresponding anisotropic proof is not known to us.

**III:** Recovered gradient ZZ error estimators are, as far as we know, always related to a *modified* residual error estimator  $\eta_{\tilde{R}}$ . For proofs on *isotropic* meshes, see for example [44] and [46]. For *anisotropic* meshes, our results are presented below. To our knowledge no other proofs exist. We may stress here that our result does not imply an equivalence between the error  $\|\nabla(u - u_h)\|$  and  $\eta_Z$  as long as step II is missing for anisotropic meshes.

The remainder of this Section is organized as follows. Although anisotropic *cuboidal* meshes are no main topic of this paper we have derived results, and will present them in the next Section for two reasons. Firstly, no anisotropic Zienkiewicz-Zhu like error estimator has been derived so far, and secondly, the structure of this estimator might give some clue for constructing estimators for tetrahedral meshes.

The last Section is devoted to anisotropic tetrahedral/triangular meshes. We motivate and introduce three different Zienkiewicz-Zhu like error estimators. For one of them an equivalence proof is established on tetrahedral meshes of tensor product type.

### 3.4.2 Cuboidal or rectangular mesh

An analysis of this estimator for an isotropic mesh which consists of rectangles (2D) and bilinear basis functions is done by Rank and Zienkiewicz [44]. The extension to rectangular prisms (3D) and trilinear basis functions is obvious.

The modification of the estimator for anisotropic rectangular or cuboidal meshes is almost straightforward. In order to get some impression of the kind of these modifications our result is stated here.

Consider a mesh formed by rectangular prisms  $T$  whose edges are aligned with the coordinate axes. Denote the edge lengths by  $h_{1,T}, h_{2,T}, h_{3,T}$ , and define the matrix  $H_T := \text{diag}(h_{1,T}, h_{2,T}, h_{3,T})$ . Let  $h_{\min} := \min\{h_{1,T}, h_{2,T}, h_{3,T}\}$ .

Define the *recovered gradient*  $\nabla^R u_h$  as the trilinear Lagrange interpolate of the nodal values

$$\nabla^R u_h(a_i) := \frac{1}{8} \sum_{T:a_i \in \mathcal{N}_T} \nabla u_h \Big|_T \quad a_i \in \mathcal{N}_I \quad .$$

(For a boundary node set the recovered derivative which is normal to the boundary equal to the true derivative). Now define the *anisotropic Zienkiewicz-Zhu* error estimator by

$$\eta_{Z,T}(u_h) := h_{\min,T} \cdot \|H_T^{-1} (\nabla^R u_h - \nabla u_h)\|_T$$

and the *modified residual error estimator* by

$$\eta_{\tilde{R},T}(u_h) := h_{\min,T} \cdot \left( \sum_{E \subset \partial T \setminus \Gamma_D} h_E^{-1} \cdot \|r_E(u_h)\|_E^2 \right)^{1/2} ,$$

i.e. only the jump residuals are utilized here, and the element residual is omitted. Simple algebra (which is analogous to [44]) shows the anisotropic Zienkiewicz-Zhu like estimator to be equivalent to the modified residual error estimator, i.e.

$$\eta_{Z,T} \sim \eta_{\tilde{R},T}$$

which is also true when suitable different weights (non-negative and bounded away from 0) of the recovered gradient are used.

### 3.4.3 Tetrahedral or triangular meshes

#### Tetrahedral mesh of tensor product type

For isotropic triangular grids (in two dimensions) a proof completely different from the one for cuboidal meshes is given by Rodriguez [46]. We have extended his ideas to the isotropic three-dimensional case and arbitrary non-negative weights of the recovered gradient, and corrected a minor mistake.

For anisotropic meshes we derive an error estimator, and present an equivalence Theorem for *tensor product type meshes*. By this we understand meshes where six tetrahedra can be found that form a rectangular prism. Unfortunately we failed to obtain an estimator for general tetrahedral, anisotropic meshes. Hence we omit the technical proof and only present the estimator and the results.

For the remainder of this paragraph assume a tetrahedral mesh of tensor product type. Let the recovered gradient be the linear Lagrange interpolate at the nodes of the mesh. The value at a node  $a$  is given by

$$\nabla^R u_h(a) := \sum_{T:a \in \mathcal{N}_T} \beta_{a,T} \cdot \nabla u_h \Big|_T \quad \sum_{T:a \in \mathcal{N}_T} \beta_{a,T} = 1 \quad ,$$

with non-negative, arbitrary weights  $\beta_{a,T} \geq 0$ .

**Definition 3.6** Define the modified anisotropic residual error estimator and the first anisotropic ZZ error estimator by

$$\begin{aligned}\eta_{\tilde{R},T}^2(u_h) &:= \sum_{E \subset \partial T \setminus \Gamma} \frac{h_{min,T}^2}{h_E} \cdot \|r_E(u_h)\|_E^2 \\ \eta_{Z_1,T}(u_h) &:= h_{min,T} \cdot \|C_T^{-T} (\nabla^R u_h - \nabla u_h)\|_T \\ \eta_{Z_1}(u_h) &:= \left( \sum_{T \in \mathcal{T}_h} \eta_{Z_1,T}^2(u_h) \right)^{1/2} = \|h_{min}(\mathbf{x}) \cdot C^{-T}(\mathbf{x}) (\nabla^R u_h - \nabla u_h)\| \quad . \quad (3.43)\end{aligned}$$

For the proof of an equivalence we require node-related quantities which can be derived easily. Denote by  $\tilde{h}_{1,T}, \tilde{h}_{2,T}, \tilde{h}_{3,T}$  the side lengths of the circumscribing rectangular prisms of a tetrahedron  $T$ . Set  $\tilde{H}_T := \text{diag}\{\tilde{h}_{1,T}, \tilde{h}_{2,T}, \tilde{h}_{3,T}\}$ . Then

$$\begin{aligned}\eta_{Z_1,T}^2(u_h) &\sim h_{min,T}^2 \cdot \left\| \tilde{H}_T^{-T} (\nabla^R u_h - \nabla u_h) \right\|_T^2 \\ &\sim h_{min,T}^2 \cdot |T| \cdot \sum_{a \in \mathcal{N}_T} \left\| \tilde{H}_T^{-T} (\nabla^R u_h - \nabla u_h)(a) \right\|_{\mathbb{R}^3}^2 \quad ,\end{aligned}$$

implying the definition

$$\eta_{Z_1,a}^2 := \sum_{T:a \in \mathcal{N}_T} \left\| \tilde{H}_T^{-T} (\nabla^R u_h - \nabla u_h|_T)(a) \right\|_{\mathbb{R}^3}^2$$

for the anisotropic ZZ error estimator. From

$$\eta_{\tilde{R},T}^2(u_h) = \sum_{E \subset \partial T \setminus \Gamma} \frac{h_{min,T}^2}{h_E} \cdot \|r_E(u_h)\|_E^2 = 3|T| \sum_{E \subset \partial T \setminus \Gamma} \frac{h_{min,T}^2}{h_E^2} \cdot r_E^2(u_h)$$

one easily identifies the node related quantity of the modified residual error estimator as

$$\eta_{\tilde{R},a}^2(u_h) := \sum_{\substack{E:a \in \mathcal{N}_E \\ E \subset \Gamma}} h_E^{-2} \cdot r_E^2(u_h) \quad .$$

**Theorem 3.9** Assume a tetrahedral, tensor product type mesh. Then the following relations hold.

$$\begin{aligned}\eta_{Z_1,a} &\sim \eta_{\tilde{R},a} \quad , \\ \sum_{T \in \mathcal{T}_h} \eta_{Z_1,T}^2 &\sim \sum_{T \in \mathcal{T}_h} \eta_{\tilde{R},T}^2 \quad , \\ \eta_{Z_1,T} &\lesssim \sum_{T' \cap T \neq \emptyset} \eta_{\tilde{R},T'} \quad \text{and} \quad \eta_{\tilde{R},T} &\lesssim \sum_{T' \cap T \neq \emptyset} \eta_{Z_1,T'} \quad .\end{aligned}$$

As mentioned above, the technical proof of this Theorem is omitted.

**Remark 3.13** There exist meshes  $\mathcal{T}_h$  (which are not of tensor product type) and functions  $u_h$  such that  $\eta_{Z_1,a} \not\sim \eta_{\tilde{R},a}$ . Note that this does not allow a prediction whether  $\|\nabla(u - u_h)\|$  is equivalent with  $\eta_{Z_1}$ , or not.  $\square$

### General tetrahedral mesh

We introduce two further ZZ error estimators. For both we know that there exist meshes  $\mathcal{T}_h$  and functions  $u_h$  such that the ZZ error estimator is *not* equivalent to the modified residual error estimator, not even for tensor product meshes. Although this may seem to be a disadvantage, the ZZ estimators may still yield a satisfying relation to the error, cf. above.

**Definition 3.7** Define the second and third ZZ error estimator by

$$\eta_{Z_2}(u_h) := \|\nabla^R u_h - \nabla u_h\| \quad (3.44)$$

$$\eta_{Z_3}(u_h) := \left\| h_{min}^{-1}(\mathbf{x}) \cdot C^T(\mathbf{x}) (\nabla^R u_h - \nabla u_h) \right\| \quad . \quad (3.45)$$

The second ZZ error estimator is completely analogous to the isotropic version. In some sense it contains the anisotropy of the solution via the recovered gradient  $\nabla^R u_h$ .

The third ZZ error estimator, however, is motivated by the matching function  $m_1$ . Suppose that the recovered gradient  $\nabla^R u_h$  approximates  $\nabla u$  well. This may require a well adapted mesh which, in turn, may yield a small matching function  $m_1(u - u_h, \mathcal{T}_h)$ . Then one has

$$\begin{aligned} \eta_{Z_3}(u_h) &= \left\| h_{min}^{-1}(\mathbf{x}) \cdot C^T(\mathbf{x}) (\nabla^R u_h - \nabla u_h) \right\| \\ &\approx \left\| h_{min}^{-1}(\mathbf{x}) \cdot C^T(\mathbf{x}) \nabla(u - u_h) \right\| \\ &\stackrel{(3.5)}{=} m_1(u - u_h, \mathcal{T}_h) \cdot \|\nabla(u - u_h)\| \\ \text{or} \quad \|\nabla(u - u_h)\| &\approx \frac{1}{m_1(u - u_h, \mathcal{T}_h)} \cdot \eta_{Z_3}(u_h) \quad . \end{aligned}$$

Again further analysis is necessary, as well as numerical experiments which may indicate whether these estimators are useful.

### 3.4.4 Conclusions

First an Zienkiewicz-Zhu (ZZ) like error estimator for anisotropic *cuboidal* meshes has been presented. The structure of that error estimator has given some clues on how to define the first ZZ like error estimator  $\eta_{Z_1}$  for *tetrahedral* anisotropic meshes. Indeed, the equivalence of  $\eta_{Z_1}$  and a modified residual error estimator  $\eta_{\tilde{R}}$  has been proven for *tensor product type meshes*. Two further ZZ like error estimators ( $\eta_{Z_2}$  and  $\eta_{Z_3}$ ) for anisotropic tetrahedral meshes have been motivated by heuristic considerations.

The theoretical foundation of our ZZ like error estimators is partially insufficient. So far, no equivalence of  $\eta_{Z_i}$  and the error  $\|\nabla(u - u_h)\|$  is proven. This requires the equivalence of  $\|\nabla(u - u_h)\|$  and  $\eta_{\tilde{R}}$  which is not available yet for anisotropic meshes (cf. step II in Section 3.4.1). Secondly, it might even be advantageous to omit the detour via the modified residual error estimator  $\eta_{\tilde{R}}$ . Instead, one can try to show an equivalence of  $\eta_{Z_i}$  and  $\|\nabla(u - u_h)\|$  directly. Then new techniques are required (e.g. a superconvergence analysis).

Keeping these restrictions in mind, some numerical experiments will be carried out in chapter 5. The results of example 4 there are promising and justify further research.

Lastly, the approximation of the exact gradient  $\nabla u$  by a recovered gradient  $\nabla^R u_h$  has been addressed in Section 3.2.4 (cf. ‘Numerical approach’ on page 52).

### 3.5 $L_2$ error estimator

An  $L_2$  error estimator for non-uniform *isotropic* meshes has been derived by Eriksson and Johnson [29]. Slightly different presentations are given by Ainsworth/Oden [1] and Becker/Rannacher [17]. Verfürth [58] additionally proves a lower error bound. Here we aim at an  $L_2$  error estimator that is suitable for *anisotropic* meshes.

In order to bound the  $L_2$  norm of the error globally from *above* we extend the ideas of [29, 1] to our case. To accomplish this, one basically requires certain anisotropic interpolation error estimates.

We also prove a local *lower bound* of the error. Its derivation is similar to the one of the residual error estimator in the energy norm. Namely, we utilize special anisotropic  $L_2$  bubble function (which are different from the usual ones of (2.19) and (2.20), or of [58, page 61]). The main difficulties lie in their construction and in the proof of the corresponding inverse inequalities.

The bound of the error from above and below renders our  $L_2$  error estimator reliable and efficient. The analysis of the estimator is organized as follows.

In Section 3.5.1 special  $L_2$  bubble functions and their inverse inequalities are given. Section 3.5.2 is devoted to the relation between the anisotropic mesh and the anisotropic solution. This relation describes the  $L_2$  counterpart of the matching function  $m_1(\cdot, \mathcal{T}_h)$  for the  $H^1$  error estimates. Anisotropic interpolation error estimates are derived in Section 3.5.3, and the  $L_2$  error estimator is presented in Section 3.5.4. There we also present the main results, namely upper and lower error bounds in the  $L_2$  norm (Theorem 3.14). Firstly, however, some useful notation is introduced.

Set  $D^2v := \left( \frac{\partial^2 v}{\partial x_i \partial x_j} \right)_{i,j=1}^d$ . Let  $M := (m_{i,j})_{i,j=1}^d$  be a matrix of  $L_2$  functions  $m_{i,j} \in L_2(\omega)$ . Define

$$\|M\|_\omega^2 := \sum_{i,j=1}^d \|m_{i,j}\|_\omega^2 \quad \text{and} \quad |M|^2 = |M(\mathbf{x})|^2 := \sum_{i,j=1}^d |m_{i,j}(\mathbf{x})|^2 \quad .$$

Finally, note that for the  $L_2$  error estimation we use special bubble functions that differ from the general bubble functions defined previously in Section 2.3.3. For simplicity the same notation  $b_T$  and  $b_E$  is used here.

#### 3.5.1 Special $L_2$ bubble functions and their inverse inequalities

For the proof of the lower bound of the error we utilize bubble functions of a higher smoothness, i.e. we now demand  $b_T \in H_o^2(T)$  and  $b_E \in H_o^2(\omega_E)$ .

Let  $T \in \mathcal{T}_h$  be an arbitrary tetrahedron, and denote by  $\lambda_{T,1}, \dots, \lambda_{T,4}$  its barycentric coordinates. The *element bubble function*  $b_T \in \mathbb{P}^8(T) \cap H_o^2(T)$  is defined by

$$b_T := 4^8 \cdot \lambda_{T,1}^2 \cdot \lambda_{T,2}^2 \cdot \lambda_{T,3}^2 \cdot \lambda_{T,4}^2 \quad \text{on } T \quad . \quad (3.46)$$

We also need a bubble function  $b_E$  defined on  $\omega_E = T_1 \cup T_2$ . The technical definition is due to the smoothness requirement  $b_E \in H_o^2(\omega_E)$ . Consider an arbitrary inner face (triangle)  $E$  of  $\mathcal{T}_h$  and the domain  $\omega_E = T_1 \cup T_2$ . The bubble function is defined separately on each tetrahedron; so let  $T$  be any of the two tetrahedra. Assume that its barycentric

coordinates are numbered such that the ones associated with the three nodal points of  $E$  are  $\lambda_1 \dots \lambda_3$ . Firstly we define three cut-off functions  $b_i \in H^2(T)$ ,  $i = 1 \dots 3$ , by

$$b_i = b_i(\lambda_1, \lambda_2, \lambda_3, \lambda_4) := \begin{cases} -128\lambda_i^3 + 48\lambda_i^2 & \text{if } \lambda_i \leq 1/4 \\ 1 & \text{if } \lambda_i > 1/4 \end{cases} .$$

A function  $b_{0,E} \in H^2(E)$  is defined by

$$b_{0,E} := \begin{cases} 12^6 \cdot (\lambda_1 - \frac{1}{4})^2 \cdot (\lambda_2 - \frac{1}{4})^2 \cdot (\lambda_3 - \frac{1}{4})^2 & \text{if } \lambda_1, \lambda_2, \lambda_3 \geq 1/4 \\ 0 & \text{otherwise} \end{cases} .$$

Let  $n_E$  be a unitary normal vector of  $E$ . A function  $b_0 \in H^2(T)$  is then given by

$$b_0(\mathbf{x}_E + t \cdot n_E) := b_{0,E}(\mathbf{x}_E) \quad \mathbf{x}_E \in E, t \in \mathbb{R} \text{ such that } \mathbf{x}_E + t \cdot n_E \in T .$$

A bubble function on the tetrahedron  $T$  is defined by

$$b_{E,T} := b_0 \cdot b_1 \cdot b_2 \cdot b_3 .$$

Finally consider the two tetrahedra  $T_1 \cup T_2 = \omega_E$ , and let the *face bubble function*  $b_E \in H^2_o(\omega_E)$  be

$$b_E(\mathbf{x}) := \begin{cases} b_{E,T_1}(\mathbf{x}) & \text{if } \mathbf{x} \in T_1 \\ b_{E,T_2}(\mathbf{x}) & \text{if } \mathbf{x} \in T_2 \end{cases} . \quad (3.47)$$

In order to visualize the construction of the face bubble function in the two-dimensional case, figure 3.3 shows the cut-off function  $b_1$  as well as  $b_0$  and  $b_E$ . They are depicted on the standard triangle.

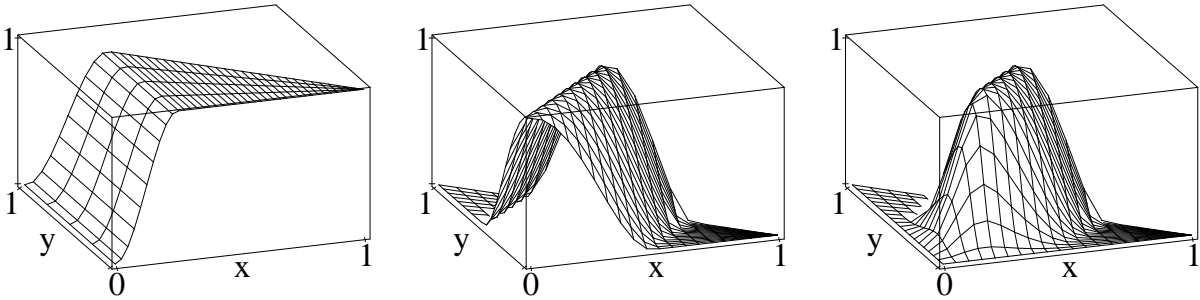


Figure 3.3: Functions  $b_1$ ,  $b_0$  and  $b_E$  (on a single triangle)

To end off, it is easily seen that

$$0 \leq b_T, b_E \leq 1 \quad \text{and} \quad \max_{\mathbf{x} \in T} b_T(\mathbf{x}) = \max_{\mathbf{x} \in \omega_E} b_E(\mathbf{x}) = 1 .$$

**Remark 3.14** To use a normal vector  $n_E$  guarantees the continuity of the derivative of  $b_0$  and  $b_E$  across  $E$  which implies  $b_E \in C^1(\omega_E) \cap H^2_o(\omega_E)$ . The functions  $b_1 \dots b_3$  act as cut-off functions to meet the zero boundary conditions of  $H^2_o$  on  $\partial T \setminus E$ .  $\square$

**Lemma 3.10 (Inverse inequalities)** *Let  $F_{ext}$  be the extension operator of (2.23). The following inverse inequalities (or equivalences) hold for all  $\varphi_T \in \mathbb{P}^0(T)$  and  $\varphi_E \in \mathbb{P}^0(E)$ .*

$$\|b_T^{1/2} \cdot \varphi_T\|_T \sim \|\varphi_T\|_T \quad (3.48)$$

$$\|b_T \cdot \varphi_T\|_T \leq \|\varphi_T\|_T \quad (3.49)$$

$$\|\Delta(b_T \cdot \varphi_T)\|_T \lesssim h_{min,T}^{-2} \cdot \|\varphi_T\|_T \quad (3.50)$$

$$\|b_E^{1/2} \cdot \varphi_E\|_E \sim \|\varphi_E\|_E \quad (3.51)$$

$$\|F_{ext}(\varphi_E) \cdot b_E\|_T \lesssim h_E^{1/2} \cdot \|\varphi_E\|_E \quad \text{for } E \subset T \quad (3.52)$$

$$\|\Delta(F_{ext}(\varphi_E) \cdot b_E)\|_T \lesssim h_E^{1/2} \cdot h_{min,T}^{-2} \cdot \|\varphi_E\|_E \quad \text{for } E \subset T \quad (3.53)$$

**Proof:** The inequalities (3.48) and (3.51) are derived analogously to inequalities (2.24) and (2.26) of Lemma 2.7. Inequality (3.49) results immediately from  $0 \leq b_T \leq 1$ .

In order to prove (3.50) we utilize the transformation technique which yields for general  $w \in H^2(T)$

$$\begin{aligned} \|\Delta w\|_T^2 &\leq 3 \cdot \|D^2 w\|^2 = 3 \int_T |A_T^{-T} \cdot A_T^T \cdot D^2 w \cdot A_T \cdot A_T^{-1}|^2 \\ &\lesssim \|A_T^{-1}\|_{\mathbb{R}^{3 \times 3}}^4 \cdot \int_T |A_T^T \cdot D^2 w \cdot A_T|^2 \\ &= \|A_T^{-1}\|_{\mathbb{R}^{3 \times 3}}^4 \cdot |\det A_T| \cdot \int_T |\bar{D}^2 \bar{w}|^2. \end{aligned}$$

For  $\mathbf{x} \in T$  set now

$$w(\mathbf{x}) := \varphi_T(\mathbf{x}) \cdot b_T(\mathbf{x}) \quad \in \mathbb{P}^8(T) \cap H_o^2(T) \quad .$$

The bound  $\|A_T^{-1}\|_{\mathbb{R}^{3 \times 3}} \lesssim h_{min,T}^{-1}$  of (2.8) and the equivalence of norms over the finite dimensional space  $\mathbb{P}^0(\bar{T}) \ni \bar{\varphi}_T$  imply

$$\begin{aligned} \|\bar{D}^2 \bar{w}\|_{\bar{T}} &= \|\bar{D}^2(\overline{\varphi_T \cdot b_T})\|_{\bar{T}} \lesssim \|\bar{\varphi}_T\|_{\bar{T}} \\ \text{and} \quad \|\Delta w\|_T &\lesssim h_{min,T}^{-2} \cdot |\det A_T|^{1/2} \cdot \|\bar{\varphi}_T\|_{\bar{T}} = h_{min,T}^{-2} \cdot \|\varphi_T\|_T \quad . \end{aligned}$$

Thus (3.50) is obtained.

Inequality (3.52) utilizes the facts that  $0 \leq b_E \leq 1$  and that  $\varphi_E \in \mathbb{P}^0(E)$  is a constant function. This yields

$$\|F_{ext}(\varphi_E) \cdot b_E\|_T \leq |T|^{1/2} \cdot |\varphi_E(\mathbf{x})| \lesssim h_E^{1/2} \cdot \|\varphi_E\|_E$$

and the desired estimate is obtained.

Inequality (3.53) requires a closer investigation. Consider a face  $E$  and any of the two tetrahedra forming  $\omega_E$ . Recall  $b_E|_T = b_0 \cdot b_1 \cdot b_2 \cdot b_3$  from the definition. As an intermediate result we will investigate the first and second derivatives of  $b_0$  and  $b_1 \dots b_3$ . With its help  $\Delta b_E$  will be bounded.

Start with any of the cut-off functions, say  $b_1$ . Recall  $\lambda_4 = 1 - \lambda_1 - \lambda_2 - \lambda_3$ . The definition of  $b_1$  (in terms of the barycentric coordinates  $\lambda_1, \lambda_2, \lambda_3$ ) is

$$b_1(\lambda_1, \lambda_2, \lambda_3) = \begin{cases} -128\lambda_1^3 + 48\lambda_1^2 & \text{if } \lambda_1 \leq 1/4 \\ 1 & \text{if } \lambda_1 > 1/4 \end{cases}$$

yielding

$$|b_1(\mathbf{x})| \leq 1 \quad \forall \mathbf{x} \in T \quad .$$

The first derivative of  $b_1$  with respect to the coordinates  $(x_1, x_2, x_3)$  can be expressed as

$$\frac{\partial b_1}{\partial x_j}(\mathbf{x}) = \sum_{i=1}^3 \frac{\partial b_1}{\partial \lambda_i}(\mathbf{x}) \cdot \frac{\partial \lambda_i}{\partial x_j}(\mathbf{x}) \quad .$$

The definition of the barycentric coordinates (which can be viewed as the linear finite element ansatz functions on  $T$ ) immediately yield

$$\left| \frac{\partial \lambda_i}{\partial x_j}(\mathbf{x}) \right| \lesssim h_{min,T}^{-1} \quad \text{or} \quad \|\nabla \lambda_i(\mathbf{x})\|_{\mathbb{R}^3} \lesssim h_{min,T}^{-1} \quad \forall \mathbf{x} \in T \quad .$$

Furthermore

$$\frac{\partial b_1}{\partial \lambda_1} = \begin{cases} -384\lambda_1^2 + 96\lambda_1 & \text{if } \lambda_1 \leq 1/4 \\ 0 & \text{if } \lambda_1 > 1/4 \end{cases} \quad \text{and} \quad \frac{\partial b_1}{\partial \lambda_2} = \frac{\partial b_1}{\partial \lambda_3} = 0$$

resulting in

$$\left| \frac{\partial b_1}{\partial \lambda_i}(\mathbf{x}) \right| \lesssim 1 \quad i = 1 \dots 3, \forall \mathbf{x} \in T \quad .$$

Combining the last two inequalities implies

$$\left| \frac{\partial b_1}{\partial x_j}(\mathbf{x}) \right| \lesssim h_{min,T}^{-1} \quad j = 1 \dots 3, \forall \mathbf{x} \in T \quad .$$

The second derivatives of  $b_1$  can be expressed as

$$\frac{\partial^2 b_1}{\partial x_\alpha \partial x_\beta}(\mathbf{x}) = \sum_{i,j=1}^3 \frac{\partial \lambda_i}{\partial x_\alpha}(\mathbf{x}) \cdot \frac{\partial^2 b_1}{\partial \lambda_i \partial \lambda_j}(\mathbf{x}) \cdot \frac{\partial \lambda_j}{\partial x_\beta}(\mathbf{x}) \quad 1 \leq \alpha, \beta \leq 3 \quad ,$$

from which one similarly concludes

$$\left| \frac{\partial^2 b_1}{\partial x_\alpha \partial x_\beta}(\mathbf{x}) \right| \lesssim h_{min,T}^{-2} \quad 1 \leq \alpha, \beta \leq 3, \forall \mathbf{x} \in T \quad .$$

Hence we obtain for the cut-off functions

$$\left| \frac{\partial^k b_i}{\partial x_j^k}(\mathbf{x}) \right| \lesssim h_{min,T}^{-k} \quad \begin{array}{l} k = 0, 1, 2 \\ \forall i = 1 \dots 3, \forall \mathbf{x} \in T \\ j = 1 \dots 3 \end{array} \quad .$$

In a slightly different fashion  $b_0$  is investigated. Clearly  $|b_0(\mathbf{x})| \leq 1$  holds for all  $\mathbf{x} \in T$ . In order to bound the derivatives of  $b_0$  we first introduce a particular coordinate system. Let  $\mathbf{l}_1$  and  $\mathbf{l}_2$  be two orthogonal unitary vectors in that plane that contains  $E$ , i.e.  $|\mathbf{l}_i| = 1$ ,  $\mathbf{l}_1 \perp \mathbf{l}_2$ ,  $E \subset \text{span}(\mathbf{l}_1, \mathbf{l}_2)$ . Set  $\mathbf{l}_3 := \mathbf{n}_E$ . Then  $(\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3)$  forms a Cartesian coordinate system.

The function  $b_0$  has been defined via  $b_{0,E}$ . The directional derivative of  $b_{0,E}$  with respect to  $\mathbf{l}_1$  or  $\mathbf{l}_2$  is expressed as

$$\frac{\partial b_{0,E}}{\partial \mathbf{l}_j}(\mathbf{x}) = \sum_{i=1}^3 \frac{\partial b_{0,E}}{\partial \lambda_i}(\mathbf{x}) \cdot \frac{\partial \lambda_i}{\partial \mathbf{l}_j}(\mathbf{x}) \quad \forall j = 1, 2, \forall \mathbf{x} \in E \quad .$$

Similar to above one concludes for  $j = 1, 2$

$$\left| \frac{\partial b_{0,E}}{\partial \lambda_i}(\mathbf{x}) \right| \lesssim 1 \quad \text{and} \quad \left| \frac{\partial \lambda_i}{\partial \mathbf{l}_j}(\mathbf{x}) \right| = |\mathbf{l}_j \cdot \nabla \lambda_i(\mathbf{x})| \leq \|\nabla \lambda_i(\mathbf{x})\|_{\mathbb{R}^3} \lesssim h_{min,T}^{-1} \quad \forall \mathbf{x} \in E$$

$$\text{giving} \quad \left| \frac{\partial b_{0,E}}{\partial \mathbf{l}_j}(\mathbf{x}) \right| \lesssim h_{min,T}^{-1} \quad \forall x \in E \quad \text{and} \quad \left| \frac{\partial b_0}{\partial \mathbf{l}_j}(\mathbf{x}) \right| \lesssim h_{min,T}^{-1} \quad \forall x \in T.$$

Additionally the definition of  $b_0$  implies

$$\frac{\partial b_0}{\partial \mathbf{l}_3}(\mathbf{x}) = 0 \quad \forall x \in T \quad .$$

The coordinate system  $(\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3)$  is transformed into the system  $(x_1, x_2, x_3)$  by a simple rotation. This immediately results in

$$\left| \frac{\partial b_0}{\partial x_j}(\mathbf{x}) \right| \lesssim h_{min,T}^{-1} \quad \forall j = 1 \dots 3, \forall x \in T \quad .$$

For the second derivatives proceed analogously to  $b_1$ , and obtain an equivalent bound. Combining the results for  $b_0$  and  $b_1 \dots b_3$ , we now dispose of

$$\left| \frac{\partial^k b_i}{\partial x_j^k}(\mathbf{x}) \right| \lesssim h_{min,T}^{-k} \quad \begin{array}{l} k = 0, 1, 2 \\ \forall i = 0 \dots 3, \forall \mathbf{x} \in T \\ j = 1 \dots 3 \end{array} \quad .$$

The product rule of differentiation now yields

$$|\Delta b_E(\mathbf{x})| = |\Delta(b_0 \cdot b_1 \cdot b_2 \cdot b_3)(\mathbf{x})| \lesssim h_{min,T}^{-2} \quad \forall \mathbf{x} \in T \quad .$$

Recalling  $\varphi_E \in \mathbb{P}^0(E)$  being constant results in

$$\|\Delta(F_{ext}(\varphi_E) \cdot b_E)\|_T \lesssim h_E^{1/2} \cdot h_{min,T}^{-2} \cdot \|\varphi_E\|_E \quad \text{for } E \subset T \quad .$$

Thus the desired inequality is proven. ■

**Remark 3.15** The choice of the element bubble function  $b_T$  is straightforward. In contrast, the definition of the face bubble function  $b_E$  is substantially more technical because of the smoothness requirement across  $E$ . One point seems particularly worth mentioning. Namely, inverse inequality (3.53) is only a one-sided inequality, i.e.

$$\|b_E\|_E \lesssim h_{min,T}^2 \cdot h_E^{-1/2} \cdot \|\Delta b_E\|_T \quad \text{for } E \subset T$$

does not hold for general meshes. Such an inequality is required (or at least advantageous) e.g. for an  $L_2$  error estimator based on a local problem. Then a modification of  $b_E$  becomes necessary. This application, however, is not discussed in this paper. □

### 3.5.2 The matching function for $L_2$ error estimation

The residual error estimation in the  $H^1$  seminorm heavily relies on anisotropic interpolation error estimates, cf. Section 3.2. The quality of the interpolation estimates has been measured by a matching function  $m_1(v, \mathcal{T}_h)$  which, roughly speaking, evaluates how good

an anisotropic mesh  $\mathcal{T}_h$  is aligned with an anisotropic function  $v$  (cf. Sections 3.2.1 and 3.2.4).

To prove an anisotropic error estimate in the  $L_2$  norm, we likewise apply certain anisotropic interpolation error estimates. Not surprisingly, their quality again depends on the alignment of mesh and function which is now measured by a matching function  $m_2(v, \mathcal{T}_h)$ .

**Definition 3.8 (Matching function  $m_2$ )** *Let  $v \in H^2(\Omega)$  be an arbitrary non-linear function, and  $\mathcal{F}$  be a family of triangulations of  $\Omega$ .*

*Define the matching function  $m_2(\cdot, \cdot) : H^2(\Omega) \times \mathcal{F} \mapsto \mathbb{R}$  by*

$$m_2(v, \mathcal{T}_h) := \left( \sum_{T \in \mathcal{T}_h} h_{\min, T}^{-4} \cdot \|C_T^T \cdot D^2 v \cdot C_T\|_T^2 \right)^{1/2} / \|D^2 v\|_\Omega \quad . \quad (3.54)$$

The definitions of  $C_T$  and  $D^2 v$  readily imply

$$1 \leq m_2(v, \mathcal{T}_h) \quad .$$

### 3.5.3 Anisotropic interpolation estimates

The imbedding Theorem for Sobolev spaces implies  $H^2(\Omega) \hookrightarrow C^0(\bar{\Omega})$ . Hence for a function  $v \in H^2(\Omega)$  the Lagrange interpolate  $\text{Int}(v) \in C^0(\bar{\Omega})$  is well-defined (i.e. the linear nodal interpolate).

First we state well-known interpolation estimates on the unitary tetrahedron  $\bar{T}$ .

**Lemma 3.11** *The following estimates hold for all  $\bar{v} \in H^2(\bar{T})$ .*

$$\begin{aligned} \|\bar{v} - \text{Int } \bar{v}\|_{\bar{T}} &\lesssim \|\bar{D}^2 \bar{v}\|_{\bar{T}} \\ \|\bar{\nabla}(\bar{v} - \text{Int } \bar{v})\|_{\bar{T}} &\lesssim \|\bar{D}^2 \bar{v}\|_{\bar{T}} \quad . \end{aligned}$$

Note that all derivatives are with respect to the reference coordinate system.

Using scaling arguments we now obtain interpolation estimates for an  $L_2$  adapted function on the actual tetrahedron  $T$ .

**Theorem 3.12** *The following interpolation estimates hold for all  $v \in H^2(\Omega)$ .*

$$\begin{aligned} \sum_{T \in \mathcal{T}_h} h_{\min, T}^{-4} \cdot \|v - \text{Int } v\|_T^2 &\lesssim m_2(v, \mathcal{T}_h)^2 \cdot \|D^2 v\|_\Omega^2 \\ \sum_{T \in \mathcal{T}_h} h_{\min, T}^{-4} \cdot \|C_T^T \nabla(v - \text{Int } v)\|_T^2 &\lesssim m_2(v, \mathcal{T}_h)^2 \cdot \|D^2 v\|_\Omega^2 \quad . \end{aligned}$$

**Proof:** The transformation technique yields

$$\|v - \text{Int } v\|_T^2 = |\det A_T| \cdot \|\bar{v} - \text{Int } \bar{v}\|_{\bar{T}}^2 \lesssim |\det A_T| \cdot \|\bar{D}^2 \bar{v}\|_{\bar{T}}^2 \quad .$$

The second derivative is transformed via

$$\bar{D}^2 \bar{v} = A_T^T \cdot D^2 v \cdot A_T$$

resulting in

$$\begin{aligned}
|\det A_T| \cdot \|\bar{D}^2 \bar{v}\|_T^2 &= |\det A_T| \cdot \int_T |\bar{D}^2 \bar{v}|^2 \\
&= \int_T |A_T^T \cdot D^2 v \cdot A_T|^2 = \int_T |A_T^T C_T^{-T} \cdot C_T^T \cdot D^2 v \cdot C_T \cdot C_T^{-1} A_T|^2 \\
&\lesssim \|C_T^{-1} A_T\|_{\mathbb{R}^{3 \times 3}}^4 \cdot \int_T |C_T^T \cdot D^2 v \cdot C_T|^2 \stackrel{(2.4)}{\lesssim} \|C_T^T \cdot D^2 v \cdot C_T\|_T^2 .
\end{aligned}$$

Applying the matching function  $m_2(\cdot, \cdot)$  completes the first part of the proof:

$$\sum_{T \in \mathcal{T}_h} h_{min,T}^{-4} \cdot \|v - \text{Int } v\|_T^2 \lesssim \sum_{T \in \mathcal{T}_h} h_{min,T}^{-4} \cdot \|C_T^T \cdot D^2 v \cdot C_T\|_T^2 \stackrel{(3.54)}{\lesssim} m_2(v, \mathcal{T}_h)^2 \cdot \|D^2 v\|_\Omega^2 .$$

The second part of the proof utilizes (2.5) giving

$$\begin{aligned}
\|C_T^T \cdot \nabla(v - \text{Int } v)\|_T^2 &= \|C_T^T A_T^{-T} \cdot A_T^T \cdot \nabla(v - \text{Int } v)\|_T^2 \\
&\leq \|C_T^T A_T^{-T}\|_{\mathbb{R}^{3 \times 3}}^2 \cdot \|A_T^T \cdot \nabla(v - \text{Int } v)\|_T^2 \\
&\stackrel{(2.5)}{\lesssim} |\det A_T| \cdot \|\bar{\nabla}(\bar{v} - \text{Int } \bar{v})\|_T^2 \lesssim |\det A_T| \cdot \|\bar{D}^2 \bar{v}\|_T^2 \\
&\lesssim \|C_T^T \cdot D^2 v \cdot C_T\|_T^2
\end{aligned}$$

as above. Inserting the matching function  $m_2(\cdot, \cdot)$  we conclude

$$\sum_{T \in \mathcal{T}_h} h_{min,T}^{-4} \cdot \|C_T^T \nabla(v - \text{Int } v)\|_T^2 \lesssim \sum_{T \in \mathcal{T}_h} h_{min,T}^{-4} \cdot \|C_T^T \cdot D^2 v \cdot C_T\|_T^2 \stackrel{(3.54)}{\lesssim} m_2(v, \mathcal{T}_h)^2 \cdot \|D^2 v\|_\Omega^2$$

analogously to the first part of the proof. ■

**Lemma 3.13** *Let  $v \in H^2(\Omega) \cap H_o^1(\Omega)$  be an arbitrary function. The following estimates hold for all  $f \in L_2(\Omega)$  and  $w_h \in V_{o,h}$ .*

$$\begin{aligned}
|(f, v - \text{Int } v)| &\lesssim m_2(v, \mathcal{T}_h) \cdot \left( \sum_{T \in \mathcal{T}_h} h_{min,T}^4 \cdot \|f\|_T^2 \right)^{1/2} \cdot \|D^2 v\|_\Omega \\
|(\nabla w_h, \nabla(v - \text{Int } v))| &\lesssim m_2(v, \mathcal{T}_h) \cdot \left( \sum_{E \subset \Omega \setminus \Gamma} \frac{h_{min,T}^4}{h_E} \cdot \|r_E(w_h)\|_E^2 \right)^{1/2} \cdot \|D^2 v\|_\Omega .
\end{aligned}$$

**Proof:** The first result is readily obtained by Cauchy's inequality and Lemma 3.12.

$$\begin{aligned}
|(f, v - \text{Int } v)| &= \left| \sum_{T \in \mathcal{T}_h} \int_T f \cdot (v - \text{Int } v) \right| \\
&\leq \sum_{T \in \mathcal{T}_h} h_{min,T}^2 \|f\|_T \cdot h_{min,T}^{-2} \|v - \text{Int } v\|_T \\
&\leq \left( \sum_{T \in \mathcal{T}_h} h_{min,T}^4 \|f\|_T^2 \right)^{1/2} \cdot \left( \sum_{T \in \mathcal{T}_h} h_{min,T}^{-4} \|v - \text{Int } v\|_T^2 \right)^{1/2} \\
&\lesssim \left( \sum_{T \in \mathcal{T}_h} h_{min,T}^4 \|f\|_T^2 \right)^{1/2} \cdot m_2(v, \mathcal{T}_h) \cdot \|D^2 v\|_\Omega .
\end{aligned}$$

To prove the second estimate we integrate by parts and apply Cauchy's inequality to conclude for any  $g \in H_o^1(\Omega)$

$$\begin{aligned} (\nabla w_h, \nabla g) &= \sum_{T \in \mathcal{T}_h} \int_T \nabla^T w_h \cdot \nabla g = \sum_{T \in \mathcal{T}_h} \int_{\partial T} \frac{\partial w_h}{\partial n} \cdot g = \sum_{E \subset \Omega \setminus \Gamma} \int_E r_E(w_h) \cdot g \\ &\leq \sum_{E \subset \Omega \setminus \Gamma} \frac{h_{min,T}^2}{h_E^{1/2}} \|r_E(w_h)\|_E \cdot \frac{h_E^{1/2}}{h_{min,T}^2} \|g\|_E \\ &\leq \left( \sum_{E \subset \Omega \setminus \Gamma} \frac{h_{min,T}^4}{h_E} \cdot \|r_E(w_h)\|_E^2 \right)^{1/2} \cdot \left( \sum_{E \subset \Omega \setminus \Gamma} \frac{h_E}{h_{min,T}^4} \|g\|_E^2 \right)^{1/2}. \end{aligned}$$

Utilizing the trace inequality (2.14)

$$\|g\|_E^2 \lesssim h_E^{-1} (\|g\|_T^2 + \|C_T^T \nabla g\|_T^2)$$

and rewriting the sum over all faces  $E$  as a sum over all tetrahedra  $T$  implies

$$\sum_{E \subset \Omega \setminus \Gamma} \frac{h_E}{h_{min,T}^4} \|g\|_E^2 \lesssim \sum_{T \in \mathcal{T}_h} h_{min,T}^{-4} (\|g\|_T^2 + \|C_T^T \nabla g\|_T^2) \quad .$$

Substituting  $g := v - \text{Int } v \in H_o^1(\Omega)$  and applying Lemma 3.12 results immediately in

$$\sum_{E \subset \Omega \setminus \Gamma} \frac{h_E}{h_{min,T}^4} \|g\|_E^2 \lesssim m_2(v, \mathcal{T}_h)^2 \cdot \|D^2 v\|_\Omega^2 \quad .$$

Thus the desired estimate is proven. ■

### 3.5.4 Anisotropic $L_2$ error estimator

**Definition 3.9** ( $L_2$  error estimator) *For an arbitrary tetrahedron  $T$  let the local  $L_2$  error estimator  $\eta_{R,L_2,T}(u_h)$  be defined by*

$$\eta_{R,L_2,T}(u_h) := h_{min,T}^2 \cdot \left( \|r_T(u_h)\|_T^2 + \sum_{E \subset \partial T \setminus \Gamma_D} h_E^{-1} \cdot \|r_E(u_h)\|_E^2 \right)^{1/2} \quad . \quad (3.55)$$

*In order to keep the notation short, we also introduce a local approximation term for a tetrahedron  $T$  by*

$$\zeta_{L_2,T} := h_{min,T}^2 \cdot \|f - P_1 f\|_{\omega_T} \quad (3.56)$$

*and the global terms*

$$\eta_{R,L_2}^2(u_h) := \sum_{T \in \mathcal{T}_h} \eta_{R,L_2,T}^2(u_h) \quad \text{and} \quad \zeta_{L_2}^2 := \sum_{T \in \mathcal{T}_h} \zeta_{L_2,T}^2 \quad . \quad (3.57)$$

In order to obtain an upper bound of the  $L_2$  error we utilize the Aubin-Nitsche trick [11, 40]. The following Theorem presents the main result of the  $L_2$  error estimation.

**Theorem 3.14** ( $L_2$  error estimation) *Let  $u \in H_o^1(\Omega)$  be the exact solution and  $u_h \in V_{o,h}$  be the finite element solution.*

*Then the error (in the  $L_2$  norm) is bounded locally from below for all  $T \in \mathcal{T}_h$  by*

$$\eta_{R,L_2,T}(u_h) \lesssim \|u - u_h\|_{\omega_T} + \zeta_{L_2,T} \quad . \quad (3.58)$$

*Assume further that  $\Omega$  is a convex polyhedral domain. Let  $v_D \in H^2(\Omega) \cap H_o^1(\Omega)$  be the solution of the dual problem*

$$-\Delta v_D = u - u_h \quad \text{in } \Omega, \quad v_D = 0 \quad \text{on } \partial\Omega \quad .$$

*Then the error (in the  $L_2$  norm) is bounded globally from above by*

$$\|u - u_h\| \lesssim m_2(v_D, \mathcal{T}_h) \cdot \left[ \eta_{R,L_2}^2(u_h) + \zeta_{L_2}^2 \right]^{1/2} \quad . \quad (3.59)$$

**Proof:** Firstly, estimate (3.58) will be proven.

We start with the norm  $\|r_T(u_h)\|_T$  of the element residual  $r_T = r_T(u_h) := P_1 f + \Delta u_h$ . Since we use linear ansatz functions  $r_T \in \mathbb{P}^0(T)$  holds. For  $\mathbf{x} \in T$  let

$$w(\mathbf{x}) := r_T(u_h)(\mathbf{x}) \cdot b_T(\mathbf{x}) \quad \in \mathbb{P}^8(T) \cap H_o^2(T) \quad .$$

Integration by parts and  $w \in H_o^2(T)$  then results in

$$\begin{aligned} \int_T r_T \cdot w &= \int_T (f + \Delta u_h) \cdot w + \int_T (P_1 f - f) \cdot w \\ &= \int_T (u_h - u) \cdot \Delta w + \int_T (P_1 f - f) \cdot w \\ \left| \int_T r_T \cdot w \right| &\leq \|u - u_h\|_T \cdot \|\Delta w\|_T + \|f - P_1 f\|_T \cdot \|w\|_T \quad . \end{aligned}$$

Recalling the inverse estimates (3.48) – (3.50) we conclude

$$\begin{aligned} \|r_T\|_T^2 &\lesssim \|u - u_h\|_T \cdot h_{min,T}^{-2} \cdot \|r_T\|_T + \|f - P_1 f\|_T \cdot \|r_T\|_T \\ \text{and} \quad h_{min,T}^4 \cdot \|r_T\|_T^2 &\lesssim \|u - u_h\|_T^2 + h_{min,T}^4 \cdot \|f - P_1 f\|_T^2 \quad . \end{aligned}$$

Now we aim at a bound of the norm  $\|r_E(u_h)\|_E$  of the gradient jump across some inner face (triangle)  $E$ . Since we use linear ansatz functions  $r_E \in \mathbb{P}^0(E)$  holds. Let  $T_1$  and  $T_2$  be the two tetrahedra that  $E$  belongs to. Define a function  $w \in H_o^2(\omega_E)$  by

$$w := F_{ext}(r_E(u_h)) \cdot b_E \quad ,$$

with  $F_{ext}$  being the extension operator of (2.23). Utilize that the right hand side  $f = -\Delta u$  is in  $L_2(\Omega)$ . Integration by parts yields

$$\begin{aligned} - \int_E w \cdot r_E(u_h) &= \int_{\omega_E} \nabla^T w \nabla (u_h - u) + \sum_{i=1}^2 \int_{T_i} w \cdot (r_{T_i} + f - P_1 f) \\ &= \int_{\omega_E} \Delta w \cdot (u - u_h) + \sum_{i=1}^2 \int_{T_i} w \cdot (r_{T_i} + f - P_1 f) \quad . \end{aligned}$$

Because of  $w|_E = r_E \cdot b_E|_E$  we conclude

$$\left| \int_E r_E^2 \cdot b_E \right| \leq \sum_{i=1}^2 \left( \|u - u_h\|_{T_i} \cdot \|\Delta w\|_{T_i} + (\|r_{T_i}\|_{T_i} + \|f - P_1 f\|_{T_i}) \cdot \|w\|_{T_i} \right) \quad .$$

Utilizing the inverse inequalities (3.51) – (3.53) results in

$$\begin{aligned} \|r_E\|_E^2 &\lesssim \sum_{i=1}^2 \left( \|u - u_h\|_{T_i} \cdot h_E^{1/2} h_{min,T_i}^{-2} \|r_E\|_E + \right. \\ &\quad \left. + (\|r_{T_i}\|_{T_i} + \|f - P_1 f\|_{T_i}) \cdot h_E^{1/2} \|r_E\|_E \right) \quad . \end{aligned}$$

The dimensions  $h_E = h_{E,T_i}$  and  $h_{min,T_i}$  cannot change rapidly for adjacent tetrahedra. Recalling the bound of  $\|r_T\|_T$  from above we conclude

$$\|r_E\|_E \lesssim h_E^{1/2} h_{min,T_1}^{-2} \cdot \left( \|u - u_h\|_{\omega_E} + h_{min,T_1}^2 \|f - P_1 f\|_{\omega_E} \right) \quad .$$

For a fixed tetrahedron  $T = T_1$  we sum up over all (inner) faces  $E \subset \partial T \setminus \Gamma_D$  and obtain

$$\sum_{E \subset \partial T \setminus \Gamma_D} \frac{h_{min,T}^4}{h_E} \cdot \|r_E(u_h)\|_E^2 \lesssim \|u - u_h\|_{\omega_T}^2 + h_{min,T}^4 \|f - P_1 f\|_{\omega_T}^2 \quad .$$

This accomplishes the proof of (3.58).

Secondly, in order to derive (3.59) we integrate by parts, utilize the dual solution  $v_D \in H^2(\Omega) \cap H_o^1(\Omega)$ , and apply Lemma 3.13 yielding

$$\begin{aligned} \|u - u_h\|^2 &= (u - u_h, -\Delta v_D) = (\nabla(u - u_h), \nabla v_D) = (\nabla(u - u_h), \nabla(v_D - \text{Int } v_D)) \\ &= (f, v_D - \text{Int } v_D) - (\nabla u_h, \nabla(v_D - \text{Int } v_D)) \\ &\lesssim \left( \sum_{T \in \mathcal{T}_h} h_{min,T}^4 \cdot \|f\|_T^2 + \sum_{E \subset \Omega \setminus \Gamma} \frac{h_{min,T}^4}{h_E} \cdot \|r_E(u_h)\|_E^2 \right)^{1/2} \cdot m_2(v_D, \mathcal{T}_h) \cdot \|D^2 v_D\| \\ &\lesssim \left( \sum_{T \in \mathcal{T}_h} h_{min,T}^4 \cdot \|f\|_T^2 + \sum_{E \subset \Omega \setminus \Gamma} \frac{h_{min,T}^4}{h_E} \cdot \|r_E(u_h)\|_E^2 \right)^{1/2} \cdot m_2(v_D, \mathcal{T}_h) \cdot \|u - u_h\| \end{aligned}$$

since  $\|D^2 v_D\| \leq c_\Omega \cdot \|\Delta v_D\| = c_\Omega \cdot \|u - u_h\|$  holds. Utilizing the triangle inequality  $\|f\|_T \leq \|P_1 f\|_T + \|f - P_1 f\|_T$  results in the upper bound (3.59) of the error.  $\blacksquare$

**Remark 3.16** The problem in applying this error estimation lies clearly in evaluating  $m_2(v_D, \mathcal{T}_h)$  for the solution  $v_D$  of the dual problem.

Additionally one may argue that the dual solution procedure is inappropriate for an anisotropic solution where probably even singularities occur.  $\square$

# Chapter 4

## A singularly perturbed reaction-diffusion equation

### 4.1 Analytical Background

Important real life problems where anisotropic solutions can occur include diffusion-convection-reaction problems, for example singularly perturbed problems. There so-called interior layers or boundary layers (of different kind) with strong anisotropic behaviour can evolve. In order to decide if error estimators can be applied in conjunction with anisotropic meshes, we have chosen the following model problem.

Let us consider a singularly perturbed reaction-diffusion equation whose *classical formulation* reads

$$\text{Find } u \in C^2(\Omega) \cap C(\bar{\Omega}) : \quad \left. \begin{array}{l} -\varepsilon \Delta u + u = f \quad \text{in } \Omega, \\ u = 0 \quad \text{on } \Gamma_D = \partial\Omega \end{array} \right\} \quad (4.1)$$

The positive parameter  $\varepsilon$  is supposed to be very small,  $\varepsilon \ll 1$ , and has much influence on the solution. Under suitable smoothness assumptions on the data (i.e.  $f$  and  $\partial\Omega$ ) the differential equation (4.1) yields a unique *classical solution*.

Similar to the Poisson equation of chapter 3, the classical formulation is too restrictive to describe real-world problems properly. Hence the so-called *variational* or *weak formulation* is more appropriate:

$$\text{Find } u \in H_o^1(\Omega) : \quad \left. \begin{array}{l} a(u, v) = \langle f, v \rangle \quad \forall v \in H_o^1(\Omega) \\ \text{with } a(u, v) := \int_{\Omega} \varepsilon \cdot \nabla^T u \nabla v + u v \\ \langle f, v \rangle := \int_{\Omega} f \cdot v \quad . \end{array} \right\} \quad (4.2)$$

The corresponding *weak solution*  $u$  is sought in the better suited space  $H_o^1(\Omega)$ . Note that the energy norm is defined by the bilinear form and depends on  $\varepsilon$ :

$$\|v\|^2 := a(v, v) = \varepsilon \|\nabla v\|^2 + \|v\|^2 \quad .$$

Throughout this chapter we demand

$$f \in L_2(\Omega) \quad .$$

Then the assumption of the Lax-Milgram Lemma (cf. [22, 26]) are satisfied, namely

- $\mathfrak{f} \in [H_o^1(\Omega)]^* = H^{-1}(\Omega)$
- $a(\cdot, \cdot)$  is elliptic, i.e.  $\exists \mu_1 > 0 : a(v, v) \geq \mu_1 \cdot \|v\|_{H_o^1(\Omega)}^2 \quad \forall v \in H_o^1(\Omega)$
- $a(\cdot, \cdot)$  is bounded, i.e.  $|a(v, w)| \leq \mu_2 \cdot \|v\|_{H_o^1(\Omega)} \cdot \|w\|_{H_o^1(\Omega)} \quad \forall v, w \in H_o^1(\Omega)$  ,

with  $\mu_1 = \varepsilon$  and  $\mu_2 = 1$ . That Lemma answers the question of the existence and uniqueness of a weak solution to the positive.

The *finite element formulation* is analogous to the Poisson equation of Section 3.1, i.e.

$$\text{Find } u_h \in V_{o,h} : \quad a(u_h, v_h) = \langle \mathfrak{f}, v_h \rangle \quad \forall v_h \in V_{o,h} \quad . \quad (4.3)$$

The model problem (4.2) is of interest since one can usually expect boundary layers when a non-vanishing right-hand side  $f$  meets homogeneous Dirichlet boundary conditions. Inside  $\Omega$  and sufficiently far away from the boundary the solution is usually smooth provided  $f$  is smooth enough too. Thus the boundary layers mark the domain of interest, and their resolution requires increased numerical effort. Note however that (4.1) is only a model problem insofar as

- the differential operator is still symmetric and elliptic.
- it can be solved using a standard FEM (with suitable meshes), i.e. no modifications like the Galerkin least squares method or the streamline diffusion method are necessary.

For a more detailed introduction to the analysis and numerical treatment of singularly perturbed differential equations (convection-diffusion and flow problems) see Roos, Stynes and Tobiska [48], and the literature cited therein. Miller, O’Riordan and Shishkin [39] investigate singularly perturbed problems with emphasis on numerical methods and *a priori* estimates.

We are interested in error estimators in particular. *Isotropic* estimators for diffusion-convection-reaction problems can be roughly divided into two major classes. *A priori* error estimators (in conjunction with adapted numerical methods) are known for some time.

For *a posteriori* error estimators, however, the knowledge has been unsatisfactory for a long time. Most estimators yield upper and lower bounds on the error that are not asymptotically equivalent. By this we mean that the upper and lower bound differ by a factor that increases, for example, as the discretization parameter  $h \rightarrow 0$ , or as  $\varepsilon \rightarrow 0$  in the case of a singularly perturbed problem. The first *a posteriori* error estimate with asymptotically equivalent upper and lower bound on the error is, to our knowledge, due to Angermann [3]. He measures the error in the somewhat strange norm

$$\|v\|_{V_0} := \sup_{v \in V_0} \frac{a(v, v)}{\|v\|_{H^1}}$$

which is weaker than the energy norm, i.e.  $\sqrt{\varepsilon} \| \|v\| \| \leq \|v\|_{V_0} \lesssim \| \|v\| \|$ . Angermann himself stated that estimates in this norm are mainly of theoretical interest.

Only recently Verfürth [59] derived the first *a posteriori* error estimator in the energy norm for the model problem (4.1) where upper and lower bounds are asymptotically equivalent.

In the remainder of this chapter an *a posteriori* error estimator for model problem (4.1) is derived that can be applied to *anisotropic* meshes. The upper and lower error bounds involve the same terms and are asymptotically equivalent. Our estimator is partially influenced by Verfürth's isotropic version. The results coincide when our estimator is applied to isotropic meshes.

## 4.2 Residual error estimator

Now the residual error estimator for anisotropic meshes is derived. The theory partly employs the methodology of Verfürth [59], and of the residual error estimator for the Poisson problem (cf. Section 3.2).

In order to bound the error from below, we utilize special bubble functions. Motivated by Verfürth's proposition, these functions are defined in Section 4.2.1 such that they can be applied to anisotropic elements. Section 4.2.2 is devoted to anisotropic interpolation error estimates which are required to bound the error from above. The anisotropic interpolation Theorem 3.3 which has been obtained in the previous chapter helps to shorten the analysis substantially. Finally, in section 4.2.3 the residual error estimator is defined. The main results, namely the bounds of the error from above and below, are presented in Theorem 4.4.

### 4.2.1 Special face bubble functions

In this Section special face bubble functions are defined, and the corresponding inverse inequalities will be derived. The definition and the proof are given first for the standard tetrahedron  $\bar{T}$  and then for the actual tetrahedron  $T$ .

Consider the standard tetrahedron  $\bar{T}$  and the face  $\bar{E}_1$  thereof. For a real number  $\delta \in (0, 1]$  define a linear mapping  $F_\delta : \mathbb{R}^d \mapsto \mathbb{R}^d$  by

$$\begin{aligned} F_\delta(x_1, \dots, x_d) &:= (\delta \cdot x_1, x_2, \dots, x_d)^T \\ \text{or } F_\delta(\mathbf{x}) &= B_\delta \cdot \mathbf{x} \quad \text{with } B_\delta = \text{diag}\{\delta, 1, \dots, 1\} \in \mathbb{R}^{d \times d} \quad . \end{aligned}$$

Obviously this yields

$$|\det B_\delta| = \delta \quad \text{and} \quad \|B_\delta^{-1}\|_{\mathbb{R}^{d \times d}} = \delta^{-1} \quad .$$

Set  $\bar{T}_\delta := F_\delta(\bar{T})$ , i.e.  $\bar{T}_\delta$  is the tetrahedron with the face  $\bar{E}_1$  and a vertex at  $\delta \cdot \mathbf{e}_1$ .

Let  $\bar{b}_{E_1}$  be the usual face bubble function of  $\bar{E}_1$  on  $\bar{T}$  (cf. (2.20)). Define the special face bubble function  $\bar{b}_\delta$  by

$$\bar{b}_\delta = \bar{b}_{E_1, \delta} := \bar{b}_{E_1} \circ F_\delta^{-1}$$

i.e.  $\bar{b}_\delta$  is the usual face bubble function of the face  $\bar{E}_1$  on the tetrahedron  $\bar{T}_\delta$ . For clarity we recall  $\bar{b}_\delta = 0$  on  $\bar{T} \setminus \bar{T}_\delta$ .

Then the following inverse inequalities hold.

#### Lemma 4.1 (Inverse inequalities on the standard tetrahedron)

Assume  $\bar{\varphi} \in \mathbb{P}^0(\bar{E}_1)$ , and let  $F_{ext}$  be the extension operator of (2.23). The following inverse inequalities hold.

$$\begin{aligned} \|\bar{b}_\delta \cdot F_{ext}(\bar{\varphi})\|_{\bar{T}} &\lesssim \delta^{1/2} \cdot \|\bar{\varphi}\|_{\bar{E}_1} \\ \|\nabla(\bar{b}_\delta \cdot F_{ext}(\bar{\varphi}))\|_{\bar{T}} &\lesssim \delta^{-1/2} \cdot \|\bar{\varphi}\|_{\bar{E}_1} \end{aligned}$$

**Proof:** We employ standard scaling techniques via  $F_\delta$  and utilize the inverse inequalities (2.27) and (2.28) on  $\bar{T}$ . Hence the desired estimates

$$\begin{aligned}
\|\bar{b}_\delta \cdot F_{ext}(\bar{\varphi})\|_{\bar{T}} &= \|\bar{b}_\delta \cdot F_{ext}(\bar{\varphi})\|_{\bar{T}_\delta} = |\det B_\delta|^{1/2} \cdot \|\bar{b}_{E_1} \cdot F_{ext}(\bar{\varphi})\|_{\bar{T}} \\
&\stackrel{(2.27)}{\lesssim} |\det B_\delta|^{1/2} \cdot h_{\bar{E}_1, \bar{T}}^{1/2} \cdot \|\bar{\varphi}\|_{\bar{E}_1} = \delta^{1/2} \cdot \|\bar{\varphi}\|_{\bar{E}_1} \\
\text{and } \|\nabla(\bar{b}_\delta \cdot F_{ext}(\bar{\varphi}))\|_{\bar{T}} &= \|\nabla(\bar{b}_\delta \cdot F_{ext}(\bar{\varphi}))\|_{\bar{T}_\delta} \\
&= |\det B_\delta|^{1/2} \cdot \|B_\delta^{-T} \cdot \nabla(\bar{b}_{E_1} \cdot F_{ext}(\bar{\varphi}))\|_{\bar{T}} \\
&\lesssim \delta^{1/2} \cdot \|B_\delta^{-1}\|_{\mathbb{R}^{d \times d}} \cdot \|\nabla(\bar{b}_{E_1} \cdot F_{ext}(\bar{\varphi}))\|_{\bar{T}} \\
&\stackrel{(2.28)}{\lesssim} \delta^{-1/2} \cdot h_{\bar{E}_1, \bar{T}}^{1/2} \cdot h_{\min, \bar{T}}^{-1} \cdot \|\bar{\varphi}\|_{\bar{E}_1} \sim \delta^{-1/2} \cdot \|\bar{\varphi}\|_{\bar{E}_1}
\end{aligned}$$

are obtained.  $\blacksquare$

**Remark 4.1** All inverse inequalities of this previous Lemma are valid for any face  $\bar{E}$  of  $\bar{T}$  (i.e. not only for  $\bar{E}_1$ ) if the face bubble function  $\bar{b}_\delta$  is defined correspondingly.  $\square$

Consider now an actual tetrahedron  $T$ . The special face bubble function  $b_\delta = b_{E, \delta} \in H^1(T)$  of a face  $E$  of  $T$  is defined by

$$b_\delta = b_{E, \delta} := \bar{b}_{E, \delta} \circ F_A^{-1} \quad . \quad (4.4)$$

**Lemma 4.2 (Inverse inequalities on the actual tetrahedron)**

Let  $E$  be an arbitrary face of  $T$ . Assume  $\varphi_E \in \mathbb{P}^0(E)$ . The following inverse inequalities hold.

$$\|b_\delta \cdot F_{ext}(\varphi_E)\|_T \lesssim \delta^{1/2} \cdot h_E^{1/2} \cdot \|\varphi_E\|_E \quad (4.5)$$

$$\|\nabla(b_\delta \cdot F_{ext}(\varphi_E))\|_T \lesssim \delta^{-1/2} \cdot h_E^{1/2} \cdot h_{\min, T}^{-1} \cdot \|\varphi_E\|_E \quad (4.6)$$

**Proof:** Standard scaling arguments and the previous Lemma readily imply

$$\|b_\delta \cdot F_{ext}(\varphi_E)\|_T^2 = 6|T| \cdot \|\bar{b}_\delta \cdot F_{ext}(\bar{\varphi}_E)\|_{\bar{T}}^2 \lesssim 6|T| \cdot \delta \cdot \|\bar{\varphi}_E\|_{\bar{E}}^2 = \delta \cdot h_E \cdot \|\varphi_E\|_E^2 \quad .$$

The other inequality is derived completely analogously and thus left to the reader.  $\blacksquare$

## 4.2.2 Anisotropic interpolation estimates

The interpolation estimates sought contain the energy norm  $\|\cdot\|$  on the right-hand side. For this reason the term  $\varepsilon$  (which is related to the differential operator and not to the interpolation operator) enters the left-hand side. More precisely, define the auxiliary term

$$\alpha_T := \min\{1, \varepsilon^{-1/2} \cdot h_{\min, T}\} \quad . \quad (4.7)$$

The following Lemma is valid.

**Lemma 4.3** Let  $R_o$  be the Clément interpolation operator defined in (3.10). The following interpolation estimates hold for any  $v \in H_o^1(\Omega)$ .

$$\sum_{T \in \mathcal{T}_h} \alpha_T^{-2} \cdot \|v - R_o v\|_T^2 \lesssim m_1(v, \mathcal{T}_h)^2 \cdot \|v\|^2 \quad (4.8)$$

$$\varepsilon^{1/2} \sum_{T \in \mathcal{T}_h} \sum_{E \subset \partial T \setminus \Gamma_D} \alpha_T^{-1} \cdot \frac{h_{E, T}}{h_{\min, T}} \|v - R_o v\|_E^2 \lesssim m_1(v, \mathcal{T}_h)^2 \cdot \|v\|^2 \quad . \quad (4.9)$$

**Proof:** The definition of  $\alpha_T$  implies

$$\alpha_T^{-1} = \max \{1, \varepsilon^{1/2} \cdot h_{min,T}^{-1}\} \quad .$$

For a better understanding we repeat here the anisotropic interpolation estimates of Theorem 3.3 on page 41. Let  $v \in H_o^1(\Omega)$ . Then

$$\begin{aligned} \|v - R_o v\| &\lesssim \|v\| \\ \|h_{min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\| &\lesssim m_1(v, \mathcal{T}_h) \cdot \|\nabla v\| \\ \|h_{min}^{-1}(\mathbf{x}) \cdot C^T(\mathbf{x}) \nabla(v - R_o v)\| &\lesssim m_1(v, \mathcal{T}_h) \cdot \|\nabla v\| \quad . \end{aligned}$$

With that help one obtains

$$\begin{aligned} \sum_{T \in \mathcal{T}_h} \alpha_T^{-2} \cdot \|v - R_o v\|_T^2 &= \sum_{\substack{T \in \mathcal{T}_h \\ 1 \geq \varepsilon \cdot h_{min,T}^{-2}}} \|v - R_o v\|_T^2 + \sum_{\substack{T \in \mathcal{T}_h \\ 1 < \varepsilon \cdot h_{min,T}^{-2}}} \varepsilon h_{min,T}^{-2} \cdot \|v - R_o v\|_T^2 \\ &\leq \|v - R_o v\|^2 + \varepsilon \cdot \|h_{min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\|^2 \\ &\lesssim \|v\|^2 + \varepsilon \cdot m_1(v, \mathcal{T}_h)^2 \cdot \|\nabla v\|^2 \leq m_1(v, \mathcal{T}_h)^2 \cdot \|v\|^2 \end{aligned}$$

which proves the first inequality.

For the second estimate the trace inequality (2.15) is invoked giving

$$h_{E,T} \cdot \|v - R_o v\|_E^2 \lesssim \|v - R_o v\|_T \cdot (\|v - R_o v\|_T + \|C_T^T \nabla(v - R_o v)\|_T) .$$

Utilizing the first result (4.8), the Cauchy–Schwarz inequality, and Theorem 3.3 on page 41 results in

$$\begin{aligned} \varepsilon^{1/2} \sum_{T \in \mathcal{T}_h} \sum_{EC \in \partial T \setminus \Gamma_D} \alpha_T^{-1} \cdot \frac{h_{E,T}}{h_{min,T}} \|v - R_o v\|_E^2 &\lesssim \\ &\lesssim \varepsilon^{1/2} \sum_{T \in \mathcal{T}_h} \left[ \alpha_T^{-1} \cdot \|v - R_o v\|_T \cdot h_{min,T}^{-1} \cdot (\|v - R_o v\|_T + \|C_T^T \nabla(v - R_o v)\|_T) \right] \\ &\lesssim \varepsilon^{1/2} \cdot \left( \sum_{T \in \mathcal{T}_h} \alpha_T^{-2} \cdot \|v - R_o v\|_T^2 \right)^{1/2} \\ &\quad \cdot (\|h_{min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\|^2 + \|h_{min}^{-1}(\mathbf{x}) \cdot C^T(\mathbf{x}) \nabla(v - R_o v)\|^2)^{1/2} \\ &\lesssim \varepsilon^{1/2} \cdot m_1(v, \mathcal{T}_h) \cdot \|v\| \cdot m_1(v, \mathcal{T}_h) \cdot \|\nabla v\| \leq m_1(v, \mathcal{T}_h)^2 \cdot \|v\|^2 \quad . \end{aligned}$$

Hence the second estimate is proven. ■

### 4.2.3 Anisotropic residual error estimator

Let the element residual over a tetrahedron  $T$  be defined by

$$r_T(v_h) := P_1 f - (-\varepsilon \cdot \Delta v_h + v_h) \quad . \quad (4.10)$$

Obviously this residual of  $v_h$  is related to the strong form of the differential operator. Therefore the definition of  $r_T$  is problem dependent and in particular different to the definition for the Poisson equation.

**Definition 4.1 (Residual error estimator)** Define the local residual error estimator  $\eta_{\varepsilon,R,T}(u_h)$  for a tetrahedron  $T$  by

$$\eta_{\varepsilon,R,T}(u_h) := \left( \alpha_T^2 \cdot \|r_T(u_h)\|_T^2 + \varepsilon^{3/2} \cdot \alpha_T \cdot \sum_{E \subset \partial T \setminus \Gamma_D} \frac{h_{\min,T}}{h_E} \cdot \|r_E(u_h)\|_E^2 \right)^{1/2}. \quad (4.11)$$

We now present the main result for the singularly perturbed equation (4.2).

**Theorem 4.4 (Residual error estimation)** Let  $u \in H_o^1(\Omega)$  be the exact solution and  $u_h \in V_{o,h}$  be the finite element solution. Then the error is bounded locally from below by

$$\eta_{\varepsilon,R,T}(u_h) \lesssim \| \|u - u_h\|_{\omega_T} + \alpha_T \cdot \|f - P_1 f\|_{\omega_T} \quad (4.12)$$

for all  $T \in \mathcal{T}_h$ . The error is bounded globally from above by

$$\| \|u - u_h\| \| \lesssim m_1(u - u_h, \mathcal{T}_h) \cdot \left( \sum_{T \in \mathcal{T}_h} \eta_{\varepsilon,R,T}^2(u_h) + \sum_{T \in \mathcal{T}_h} \alpha_T^2 \cdot \|f - P_1 f\|_T^2 \right)^{1/2}. \quad (4.13)$$

**Proof:** The proof of the first estimate (4.12) employs some standard techniques already utilized for the Poisson equation. A more detailed investigation can be found there.

We start with the norm  $\|r_T(u_h)\|_T$  of the element residual  $r_T = P_1 f + \varepsilon \cdot \Delta u_h - u_h$ . Since we use linear ansatz functions  $r_T \in \mathbb{P}^0(T)$  holds. For  $\mathbf{x} \in T$  let

$$w(\mathbf{x}) := r_T(u_h)(\mathbf{x}) \cdot b_T(\mathbf{x}) \quad \in \mathbb{P}^4(T) \cap H_o^1(T) \quad ,$$

with  $b_T$  being the usual bubble functions introduced in Section 2.3.3. Integration by parts yields

$$\begin{aligned} \int_T r_T \cdot w &= \int_T (f + \varepsilon \cdot \Delta u_h - u_h) \cdot w + \int_T (P_1 f - f) \cdot w \\ &= \int_T \varepsilon \cdot \nabla^T(u - u_h) \cdot \nabla w + (u - u_h) \cdot w + \int_T (P_1 f - f) \cdot w \\ \left| \int_T r_T \cdot w \right| &\leq \varepsilon \cdot \|\nabla(u - u_h)\|_T \cdot \|\nabla w\|_T + \|u - u_h\| \cdot \|w\|_T + \|f - P_1 f\|_T \cdot \|w\|_T \quad . \end{aligned}$$

Bounds of  $\left| \int_T r_T \cdot w \right|$ ,  $\|\nabla w\|_T$  and  $\|w\|_T$  have already been derived in (3.24). Hence one readily obtains

$$\begin{aligned} \|r_T\|_T^2 &\lesssim \varepsilon^2 \cdot h_{\min,T}^{-2} \cdot \|\nabla(u - u_h)\|_T^2 + \|u - u_h\|_T^2 + \|f - P_1 f\|_T^2 \\ \alpha_T^2 \cdot \|r_T\|_T^2 &\lesssim \min\{\varepsilon \cdot h_{\min,T}^{-2}, 1\} \cdot \varepsilon \cdot \|\nabla(u - u_h)\|_T^2 + \\ &\quad + \alpha_T^2 \cdot \|u - u_h\|_T^2 + \alpha_T^2 \cdot \|f - P_1 f\|_T^2 \\ &\leq \varepsilon \cdot \|\nabla(u - u_h)\|_T^2 + \|u - u_h\|_T^2 + \alpha_T^2 \cdot \|f - P_1 f\|_T^2 \\ &= \| \|u - u_h\|_T^2 + \alpha_T^2 \cdot \|f - P_1 f\|_T^2 \quad . \end{aligned} \quad (4.14)$$

Now we aim at a bound of the norm  $\|r_E(u_h)\|_E$  of the gradient jump across some inner face (triangle)  $E$ . Since we use linear ansatz functions  $r_E \in \mathbb{P}^0(E)$  holds. Let  $T_1$  and  $T_2$

be the two tetrahedra that  $E$  belongs to. Assume that the right hand side  $f = -\varepsilon\Delta u + u$  is in  $L_2(\Omega)$ . Integration by parts yields for any function  $w \in H_o^1(\omega_E)$

$$\begin{aligned} 0 &= \int_{\omega_E} \varepsilon \nabla^T u \nabla w + u \cdot w - f \cdot w \\ -\varepsilon \int_E r_E(u_h) \cdot w &= \varepsilon \sum_{i=1}^2 \int_{\partial T_i} w \cdot \frac{\partial u_h}{\partial n} = \varepsilon \sum_{i=1}^2 \int_{T_i} (\nabla^T u_h \nabla w + \Delta u_h \cdot w) \\ &= \sum_{i=1}^2 \int_{T_i} (\varepsilon \nabla^T u_h \nabla w + (r_{T_i} - P_1 f + u_h) \cdot w) \\ &= \sum_{i=1}^2 \int_{T_i} (\varepsilon \nabla^T (u_h - u) \nabla w + (u_h - u) \cdot w + (r_{T_i} + f - P_1 f) \cdot w) \end{aligned}$$

since  $\varepsilon \Delta u_h = r_{T_i} - P_1 f + u_h$  on  $T_i$ . Let now the function  $w$  be defined by

$$w := \begin{cases} b_{E,\delta_1} \cdot F_{ext}(r_E(u_h)) & \text{on } T_1 \\ b_{E,\delta_2} \cdot F_{ext}(r_E(u_h)) & \text{on } T_2 \end{cases},$$

with  $F_{ext}$  being the extension operator of (2.23) and  $b_{E,\delta_i}$  being the special face bubble functions defined above. The real numbers  $\delta_i$  will be chosen later.

Note that  $w \in H_o^1(\omega_E)$  since  $b_{E,\delta_1}|_E = b_{E,\delta_2}|_E = b_E|_E$ . Hence we conclude

$$\varepsilon \|b_E^{1/2} \cdot r_E\|_E^2 \leq \sum_{i=1}^2 \left( \varepsilon \|\nabla(u - u_h)\|_{T_i} \cdot \|\nabla w\|_{T_i} + (\|u - u_h\|_{T_i} + \|r_{T_i}\|_{T_i} + \|f - P_1 f\|_{T_i}) \cdot \|w\|_{T_i} \right).$$

The inverse inequalities (4.5) and (4.6) are used to bound  $\|w\|_{T_i}$  and  $\|\nabla w\|_{T_i}$ , respectively, and subsequently imply

$$\begin{aligned} \|r_E\|_E &\lesssim \sum_{i=1}^2 h_E^{1/2} \cdot \left( h_{min,T_i}^{-1} \cdot \delta_i^{-1/2} \cdot \|\nabla(u - u_h)\|_{T_i} + \right. \\ &\quad \left. + \varepsilon^{-1} \cdot \delta_i^{1/2} \cdot (\|u - u_h\|_{T_i} + \|r_{T_i}\|_{T_i} + \|f - P_1 f\|_{T_i}) \right). \end{aligned}$$

Now we choose  $\delta_i := \varepsilon^{1/2} \cdot h_{min,T_i}^{-1} \cdot \alpha_{T_i} \leq 1$  and insert estimate (4.14) which provides a bound of  $\|r_{T_i}\|_{T_i}$ . One obtains

$$\begin{aligned} \varepsilon^{3/2} \cdot \alpha_T \cdot \frac{h_{min,T}}{h_E} \cdot \|r_E(u_h)\|_E^2 &\lesssim \\ &\lesssim \sum_{i=1}^2 \varepsilon \cdot \|\nabla(u - u_h)\|_{T_i}^2 + \alpha_{T_i}^2 \cdot \|u - u_h\|_{T_i}^2 + \|u - u_h\|_{T_i}^2 + \alpha_{T_i}^2 \cdot \|f - P_1 f\|_{T_i}^2 \\ &\lesssim \|u - u_h\|_{\omega_E}^2 + \alpha_T^2 \cdot \|f - P_1 f\|_{\omega_E}^2 \end{aligned}$$

since  $h_{min,T_i}$  and  $\alpha_{T_i}$  do not change rapidly across adjacent tetrahedra, and since  $\alpha_{T_i} \leq 1$ . Summing up over all faces  $E$  of  $T$ , recalling the definition of  $\eta_{\varepsilon,R,T}(u_h)$  and applying (4.14) finishes the proof of the lower error bound (4.12).

Secondly, in order to derive (4.13) we utilize the orthogonality property of the error

$$a(u - u_h, v_h) = 0 \quad \forall v_h \in V_{o,h} \quad .$$

Integration by parts gives for all  $v \in H_o^1(\Omega)$

$$\begin{aligned}
a(u - u_h, v) &= a(u - u_h, v - R_o v) \\
&= \varepsilon(\nabla(u - u_h), \nabla(v - R_o v)) + (u - u_h, v - R_o v) \\
&= \sum_{T \in \mathcal{T}_h} (f + \varepsilon \Delta u_h - u_h, v - R_o v)_T + \varepsilon \sum_{EC\Omega \setminus \Gamma} (r_E(u_h), v - R_o v)_E \\
&= \sum_{T \in \mathcal{T}_h} \left[ (r_T(u_h) + f - P_1 f, v - R_o v)_T + \frac{1}{2} \cdot \varepsilon \sum_{EC\partial T \setminus \Gamma_D} (r_E(u_h), v - R_o v)_E \right] \\
&\leq \sum_{T \in \mathcal{T}_h} \left[ \alpha_T (\|r_T(u_h)\|_T + \|f - P_1 f\|_T) \cdot \alpha_T^{-1} \|v - R_o v\|_T + \right. \\
&\quad \left. + \frac{1}{2} \sum_{EC\partial T \setminus \Gamma_D} \varepsilon^{3/4} \alpha_T^{1/2} \frac{h_{min,T}^{1/2}}{h_{E,T}^{1/2}} \|r_E(u_h)\|_E \cdot \varepsilon^{1/4} \alpha_T^{-1/2} \frac{h_{E,T}^{1/2}}{h_{min,T}^{1/2}} \|v - R_o v\|_E \right].
\end{aligned}$$

The Cauchy-Schwarz inequality and the interpolation estimate (4.8) yield

$$\begin{aligned}
\sum_{T \in \mathcal{T}_h} \alpha_T (\|r_T(u_h)\|_T + \|f - P_1 f\|_T) \cdot \alpha_T^{-1} \|v - R_o v\|_T &\leq \\
&\leq \left( 2 \sum_{T \in \mathcal{T}_h} \alpha_T^2 (\|r_T(u_h)\|_T^2 + \|f - P_1 f\|_T^2) \right)^{1/2} \cdot \left( \sum_{T \in \mathcal{T}_h} \alpha_T^{-2} \|v - R_o v\|_T^2 \right)^{1/2} \\
&\stackrel{(4.8)}{\lesssim} \left( \sum_{T \in \mathcal{T}_h} \alpha_T^2 (\|r_T(u_h)\|_T^2 + \|f - P_1 f\|_T^2) \right)^{1/2} \cdot m_1(v, \mathcal{T}_h) \cdot \|v\|.
\end{aligned}$$

With the help of interpolation estimate (4.9) one derives analogously

$$\begin{aligned}
\sum_{T \in \mathcal{T}_h} \sum_{EC\partial T \setminus \Gamma_D} \varepsilon^{3/4} \alpha_T^{1/2} \frac{h_{min,T}^{1/2}}{h_{E,T}^{1/2}} \|r_E(u_h)\|_E \cdot \varepsilon^{1/4} \alpha_T^{-1/2} \frac{h_{E,T}^{1/2}}{h_{min,T}^{1/2}} \|v - R_o v\|_E &\leq \\
&\leq \left( \varepsilon^{3/2} \sum_{T \in \mathcal{T}_h} \sum_{EC\partial T \setminus \Gamma_D} \alpha_T \frac{h_{min,T}}{h_{E,T}} \|r_E(u_h)\|_E^2 \right)^{1/2} \cdot \\
&\quad \cdot \left( \varepsilon^{1/2} \sum_{T \in \mathcal{T}_h} \sum_{EC\partial T \setminus \Gamma_D} \alpha_T^{-1} \frac{h_{E,T}}{h_{min,T}} \|v - R_o v\|_E^2 \right)^{1/2} \\
&\stackrel{(4.9)}{\lesssim} \left( \varepsilon^{3/2} \sum_{T \in \mathcal{T}_h} \sum_{EC\partial T \setminus \Gamma_D} \alpha_T \frac{h_{min,T}}{h_{E,T}} \|r_E(u_h)\|_E^2 \right)^{1/2} \cdot m_1(v, \mathcal{T}_h) \cdot \|v\|.
\end{aligned}$$

Combining these estimates results in

$$\begin{aligned}
a(u - u_h, v) &\lesssim \left( \sum_{T \in \mathcal{T}_h} \left[ \alpha_T^2 (\|r_T(u_h)\|_T^2 + \|f - P_1 f\|_T^2) + \right. \right. \\
&\quad \left. \left. + \varepsilon^{3/2} \alpha_T \sum_{EC\partial T \setminus \Gamma_D} \frac{h_{min,T}}{h_{E,T}} \|r_E(u_h)\|_E^2 \right] \right)^{1/2} \cdot m_1(v, \mathcal{T}_h) \cdot \|v\|.
\end{aligned}$$

Substituting  $v := u - u_h \in H_o^1(\Omega)$  finishes the proof. ■

# Chapter 5

## Numerical examples

### 5.1 Scope of and introduction to the numerical experiments

#### 5.1.1 General Remarks

For clarity and optimal understanding the presentation of numerical experiments are organized as follows. In Section 5.1 we briefly recapitulate major results and topics which are to be tested, namely

- the matching functions,
- the interpolation error estimates,
- the finite element error estimates (upper and lower bounds of the error and related results).

Section 5.1.5 gives a rough overview of the computational effort of the error estimation process. To facilitate the readability of the examples, Section 5.1.6 describes the setting and the aim of each experiment. Thus one can easily find out the appropriate examples for a specific topic.

Section 5.2 is devoted to the examples. For each of them, the aim of the experiment and the boundary value problem are described, the exact (but usually unknown) solution is given, and the meshes used are listed. Then the aforementioned topics (matching function, interpolation and finite element error estimates) are numerically analysed, and conclusions are drawn. Finally, Section 5.3 summarizes the experiments.

Virtually all results are written as inequalities, most of them in the form

$$a \lesssim b \quad \text{or} \quad a \lesssim m_1(\cdot, \mathcal{T}_h) \cdot b \quad .$$

With numerical examples it is impossible to *prove* such inequalities (this has been done already), but the following steps are appropriate.

- Evaluate  $a/b$  or  $a/m_1/b$  for several meshes  $\mathcal{T}_h$ , or for a sequence of (comparable) meshes. Verify whether these ratios are bounded from above. The size and the variation of these ratios can hint at the quality and sharpness of the inequality.

A sequence of comparable meshes can suggest whether an inequality only holds in an asymptotic sense (i.e. for many degrees of freedom), or if it yields useful results also for specific examples.

- Investigate if the inequalities  $a \lesssim m_1(\cdot, \mathcal{T}_h) \cdot b$  are sharp (i.e. whether  $m_1$  is really necessary). For that purpose, evaluate  $a/m_1/b$  and  $a/b$  simultaneously for several meshes/a mesh sequence and decide whether they seem to be bounded, or growing.
- Where appropriate, observe when  $a/b$  is not bounded from below (i.e. when  $b \not\lesssim a$ ).

For all examples one requires anisotropic tetrahedral or triangular meshes. Certainly it would be desirable to obtain such meshes from an adaptive strategy, as it is common for real world problems. Such an adaptive and anisotropic strategy, however, is beyond the scope and means of this work. In particular the extraction of information for an optimal mesh, and the mesh refinement are less well understood in our anisotropic context.

As a consequence, our meshes have to be constructed in a different way, for example by using *a priori* knowledge of the (usually unknown) solution  $u$ . We may stress that usually a (quasi) optimal mesh is not known for a specific problem; even the term ‘optimal’ deserves discussion. Hence we will utilize single, particular meshes as well as a sequence of meshes. These meshes may not be optimal but will hopefully suit our demands.

Finally, for two-dimensional domains  $\Omega \subset \mathbb{R}^2$  we utilized the program FEMGP on a serial computer [53]. For three-dimensional domains  $\Omega \subset \mathbb{R}^3$  we used the package SPC-PM Po3D [4, 9]. It runs on both parallel and serial computers and demonstrates that error estimation routines can be parallelized quite effectively. We have extended both packages by error estimation modules which are suitable for error estimation on both isotropic and anisotropic meshes.

### 5.1.2 The matching functions $m_1$ and $m_2$

#### The matching function $m_1$

The matching function  $m_1(\cdot, \mathcal{T}_h)$  is important for the interpolation error estimates and for various finite element error estimates. We recapitulate the definitions of  $m_1$  from (3.5), and of our approximation  $m_1^R$  from (3.27):

$$m_1(v, \mathcal{T}_h) = \frac{\|h_{min}^{-1}(\mathbf{x}) \cdot C^T(\mathbf{x}) \nabla(u - u_h)\|}{\|\nabla(u - u_h)\|}$$

$$m_1^R(u - u_h, \mathcal{T}_h) = \frac{\left( \sum_{T \in \mathcal{T}_h} h_{min,T}^{-2} \cdot \|C_T^T (\nabla^R u_h - \nabla u_h)\|_T^2 \right)^{1/2}}{\|\nabla^R u_h - \nabla u_h\|}.$$

We compute both fractions as well as the respective numerator and denominator.

#### The matching function $m_2$

The matching function  $m_2$  plays a similar role in  $L_2$  error estimation as  $m_1$  does in  $H^1$  error estimation.

The factor  $m_2(v, \mathcal{T}_h)$  enters the interpolation error estimates of Lemma 3.12 on page 78. Similarly, the quality of the upper bound (3.59) of the  $L_2$  finite element error depends on  $m_2(v_D, \mathcal{T}_h)$ , with  $v_D \in H^2(\Omega) \cap H_o^1(\Omega)$  being the solution of the dual problem  $-\Delta v_D = u - u_h$ . Now we are no longer able to approximate (or evaluate)  $m_2(v_D, \mathcal{T}_h)$  since neither  $u - u_h$  nor the corresponding dual solution  $v_D$  is known. Therefore we do not investigate into  $m_2$  here. Moreover, the numerical experiments suggest that a mesh which yields a small finite element error in the energy norm also results in a

small  $L_2$  finite element error. This observation abates the immediate necessity to estimate  $m_2$ .

**Remark 5.1** It is conceivable that  $v_D$  (or more precisely  $D^2 v_D$ ) can be approximated to some extent by the following procedure that resembles somewhat the local problem estimation process of Section 3.3.

Imagine that the dual problem  $-\Delta v_D = u - u_h$  is replaced by a sequence of local problems (e.g. defined on  $\omega_T$ ) which are then solved approximately in a higher-dimensional local finite element space. The unknown solution  $u$  can probably be eliminated by integration by parts and/or a recovered gradient.  $\square$

### 5.1.3 Interpolation error estimates

The interpolation error estimates (Theorems 3.3 on page 41 and 3.12 on page 78) play a vital role in deriving finite element error estimates. We have chosen to test three of them, namely (3.11)–(3.13), but only for the two-dimensional examples 1 and 2. There we calculate the corresponding ratios

$$\frac{\|v - R_o v\|}{\|v\|} \quad , \quad \frac{\|h_{min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\|}{m_1(v, \mathcal{T}_h) \cdot \|v\|} \quad , \quad \frac{\|h_{min}^{-1}(\mathbf{x}) \cdot C^T(\mathbf{x}) \nabla(v - R_o v)\|}{m_1(v, \mathcal{T}_h) \cdot \|v\|}$$

which have to be bounded from above.

We note that the anisotropic interpolation error estimates (3.12)–(3.16) depend also on the matching function  $m_1$ , which is different to isotropic interpolation error estimates. In Remark 3.3 on page 44 it was shown that interpolation estimate (3.12) is sharp, i.e. that  $m_1$  is necessary. In order to quantify this result we do not only calculate the ratio

$$\frac{\|h_{min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\|}{m_1(v, \mathcal{T}_h) \cdot \|v\|} \quad \text{but also} \quad m_1(v, \mathcal{T}_h) \quad \text{and} \quad \frac{\|h_{min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\|}{\|v\|} \quad .$$

Of course, these two ratios show the same behaviour in cases where  $m_1$  is bounded. Hence we present these values only for example 2.

### 5.1.4 Finite element error estimates

#### Residual error estimator

For the Poisson problem the residual error estimator yields a global upper bound (3.23) and a local lower bound (3.22) of the finite element error:

$$\begin{aligned} \| \|u - u_h\| \| &\lesssim m_1(u - u_h, \mathcal{T}_h) \cdot \sqrt{\eta_R^2 + \zeta^2} \\ \eta_{R,T}(u_h) &\lesssim \| \|u - u_h\| \|_{\omega_T} + \zeta_T \quad \forall T \in \mathcal{T}_h \quad . \end{aligned}$$

In order to verify the upper error bound we calculate the terms

$$\| \|u - u_h\| \| \quad , \quad \eta_R \quad , \quad \zeta \quad , \quad m_1(u - u_h, \mathcal{T}_h) \quad , \quad \frac{\| \|u - u_h\| \|}{m_1(u - u_h, \mathcal{T}_h) \cdot \sqrt{\eta_R^2 + \zeta^2}} \quad .$$

To isolate the influence of the matching function  $m_1$  and its approximation  $m_1^R$ , respectively, we also present the values of

$$m_1 \quad \text{and} \quad m_1^R \quad , \quad \frac{\| \|u - u_h\| \|}{m_1 \cdot \sqrt{\eta_R^2 + \zeta^2}} \quad \text{and} \quad \frac{\| \|u - u_h\| \|}{m_1^R \cdot \sqrt{\eta_R^2 + \zeta^2}} \quad , \quad \frac{\| \|u - u_h\| \|}{\sqrt{\eta_R^2 + \zeta^2}}$$

for examples 1–3.

The lower error bound implies that the ratio

$$\frac{\eta_{R,T}(u_h)}{\| \|u - u_h\| \|_{\omega_T} + \zeta_T}$$

is bounded from above for all  $T \in \mathcal{T}_h$ . Hence we present the *maximum* of this ratio over all  $T$ . The *average* of this ratio is calculated as well and shall give some impression of the distribution. Finally, the *minimal* ratio indicates that the local error bound is only a one-sided estimate.

### Local problem error estimator

The error bounds of the local problem error estimator (Theorem 3.7 on page 64) are analogous to the residual error estimator. Hence we proceed as above. Note that the error is bounded from below with a constant 1, i.e.

$$\eta_{D,T}(u_h) \leq \| \|u - u_h\| \|_{\omega_T} + c \cdot \zeta_T \quad .$$

Furthermore, Theorem 3.6 on page 61 states the equivalence of the residual and the local problem error estimator. Therefore we calculate the ratios  $\eta_{D,T}/\eta_{R,\omega_T}$  and  $\eta_{R,T}/\eta_{D,\omega_T}$  for all  $T \in \mathcal{T}_h$  and present their maximum value (3D examples only). Finally, we calculate the maximum condition number of all local problems. Its boundedness shall confirm Theorem 3.8 on page 65.

### Three Zienkiewicz-Zhu like error estimators

In Section 3.4 three different Zienkiewicz-Zhu like error estimator are proposed, and some results are presented. The theory, however, is not complete yet. Hence we have concluded that, at the present stage, only numerical experiments can suggest (or reject) a particular error estimator. Here we investigate whether the estimators  $\eta_{Z_1}, \eta_{Z_2}, \eta_{Z_3}$  from (3.43)–(3.45) with the recovered gradient from (3.28) reliably bound the error from above (for our examples). The corresponding ratios

$$\frac{\|\nabla(u - u_h)\|}{\sqrt{\eta_{Z_1}^2 + \zeta^2}} \quad , \quad \frac{\|\nabla(u - u_h)\|}{\sqrt{\eta_{Z_2}^2 + \zeta^2}} \quad , \quad \frac{\|\nabla(u - u_h)\|}{\sqrt{\eta_{Z_3}^2 + \zeta^2}}$$

are computed. We remark that the proof for  $\eta_{Z_1}$  (cf. Theorem 3.9 on page 71) assumes a tensor product type (tetrahedral or triangular) mesh. In example 4 this is not satisfied. Nevertheless  $\eta_{Z_1}$  will be tested (cf. Section 3.4).

Finally, we restrict ourselves to three-dimensional examples. (If one is interested in two-dimensional results, the values of  $\eta_{Z_2}, \eta_{Z_3}$  and  $\|\nabla(u - u_h)\|$  are contained in the table of the matching function  $m_1$ .)

### $L_2$ residual error estimator

The residual error estimates for the  $L_2$  norm are analogous to the ones in the energy norm. The lower error bound is thus analysed as above. The upper error bound contains the matching function  $m_2(v_D, \mathcal{T}_h)$  which we could neither evaluate nor estimate. Hence we calculate only

$$\|u - u_h\| \quad , \quad \sqrt{\eta_{R,L_2}^2 + \zeta_{L_2}^2} \quad , \quad \frac{\|u - u_h\|}{\sqrt{\eta_{R,L_2}^2 + \zeta_{L_2}^2}} \quad .$$

Strictly speaking, the  $L_2$  error estimate has not been proven yet for the case of Neumann boundary conditions which occur in example 4. The tests are nevertheless carried out to obtain numerical evidence.

### 5.1.5 Computational effort

The computational expense is certainly not the most important topic of our work. Nevertheless it is interesting to have at least an overview of how expensive certain computations or estimators are. The expenditure depends on many factors like

- the degree of the ansatz and test functions used in the finite element computation (cf. Remark 3.12 on page 67),
- the kind of approximation by the mappings  $P_1, P_2, P_3$  (cf. Section 3.2.3 and Remark 3.12),
- whether the data structure meets the demands of error estimation.

Additionally several routines are interwoven or can be combined favourably which aggravates an accurate comparison. Hence we present here a sample (mesh 7e of example 4) that shall give a rough overview of the computational effort. It is scaled such that the sum of the cost of the whole estimation process sums up to 100%.

Computation of ...	Effort (approx.)
Basic routines	6%
Face residual $r_E(u_h)$ and recovered gradient $\nabla^R u_h$	5%
Matching function $m_1(u - u_h, \mathcal{T}_h)$ and three Zienkiewicz-Zhu estimators $\eta_{Z_1}, \eta_{Z_2}, \eta_{Z_3}$	4%
Residual error estimators $\eta_R$ and $\eta_{R,L_2}$	30%
Local problem error estimator $\eta_D$	55%
$\Sigma$	100%

For interior faces  $E$  the face residual  $r_E$  coincides with the gradient jump, for which  $\nabla u_h|_T$  has to be computed. The recovered gradient  $\nabla^R u_h$  of (3.28) is then obtained as a side effect, and thus very cheap.

The effort of approximating the matching function  $m_1$  is almost neglectable compared with the computation of  $\eta_R$  or  $\eta_D$ . Hence the evaluation of  $m_1^R$  yields useful information on the mesh at low cost, and is thus recommended.

### 5.1.6 Aim of the experiments

Below the setting and the aim of each experiment are briefly described. In this way one can also easily find out the appropriate examples for a specific topic.

*Example 1* is posed over a two-dimensional domain with an *anisotropic* solution. The *anisotropic* meshes are adapted to that solution and thus form the ideal environment of anisotropic error estimation. Step by step we will investigate most results numerically and explain the conclusions in detail. In particular we explore the matching function (its size and its approximation), the interpolation error estimates and, most important, the error estimators.

*Example 2* employs the same two-dimensional problem with an *anisotropic* solution. Now *anisotropic* meshes are considered which are *not adapted* to the solution. We want to know whether the matching function  $m_1$  becomes large, and if it can be approximated.

The effects of a large  $m_1$  onto interpolation error estimates and finite element error estimates are studied and compared with theoretical predictions. Furthermore this example suggests numerically that  $m_1$  cannot be neglected in the appropriate inequalities.

*Example 3* utilizes an *isotropic* solution over a two-dimensional domain. Here *anisotropic* meshes are investigated (instead of the isotropic, optimal ones). Such a situation could occur for a system of differential equations, whose solution components show different (anisotropic or isotropic) behaviour, for example. We explore  $m_1$  and the circumstances when it is large. Then the performance of the error estimators is studied.

*Example 4* is posed over a *three-dimensional* domain with an *anisotropic* solution. The *anisotropic* meshes are adapted to that solution and are thus ideal for anisotropic error estimation. Analogously to example 1 we explore our results step by step numerically. First the matching function  $m_1$  is evaluated and approximated. Then the residual error estimator, the local problem error estimator, three Zienkiewicz-Zhu like error estimators, and an  $L_2$  error estimator are investigated.

## 5.2 Numerical examples

### 5.2.1 Example 1 (2D; anisotropic solution and adapted mesh)

#### Aim of the example

Consider the following two-dimensional problem with a boundary layer. We investigate our results on several anisotropic meshes which are *adapted* to this layer. The results will show that

- the matching function  $m_1$  is small, and that it can be approximated fairly well,
- the interpolation error estimates hold,
- the finite element error estimates hold.

#### Boundary value problem

The two-dimensional Poisson problem

$$-\Delta u = f \quad \text{in } \Omega = (0, 1) \times (0, 1) \quad , \quad u = 0 \quad \text{on } \Gamma_D = \partial\Omega$$

is chosen as test problem. The exact solution  $u$  is prescribed to be

$$u(x, y) := (1 - e^{-\alpha x} - (1 - e^{-\alpha})x) \cdot 4y(1 - y)$$

with a parameter  $\alpha = 1000$ . The right-hand side  $f$  is chosen accordingly. The exact solution exhibits an exponential layer with an initial steepness of  $\alpha = 1000$  along the boundary at  $x = 0$ . Figure 5.1 shows a rough image of  $u$ .

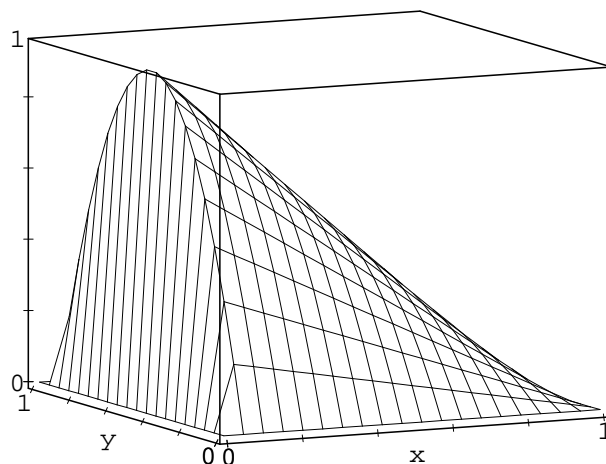


Figure 5.1: Exact solution  $u$

#### Meshes

As mentioned before, we mainly utilize meshes which were constructed based on *a priori* knowledge. These meshes are now briefly described and, additionally, most of them are visualized in figure 5.2 on page 99. Further details of the meshes are given in the table

below and in Remark 5.2 at the end of this Section. The nodal  $L_\infty$  error shall give a certain idea of the quality of the finite element solution  $u_h$ .

*Meshes 1a–c* are unstructured meshes. They were obtained by an automatic isotropic mesh generation on the unit square with a local mesh size  $h = 1/30$ ,  $h = 1/60$  and  $h = 1/120$ . A subsequent nodal coordinate transformation has been performed to resolve the boundary layer. We may stress that these meshes are far from optimal but sufficient for our test purposes. Note that the largest interior angle is always larger than  $179.5^\circ$  which confirms that a maximum angle condition is *not* necessary in our theory. The meshes are not nested.

*Mesh 2* has been constructed similarly but more sophisticated. *Meshes 3a–c* are of tensor-product type and form a mesh sequence. They have been constructed using a different nodal coordinate transformation. The meshes are not nested.

Finally, with *meshes 4a–h* we were implementing a crude adaptivity in  $x$ -direction only (in mesh 4h an additional bisection in  $y$ -direction has been inserted). To economize on space, results are presented for meshes 4a, 4c, 4e, and 4h only. We emphatically stress that meshes 4a and 4c are too coarse to resolve the boundary layer and thus yield extremely inaccurate ‘solutions’  $u_h$ . (For example,  $\|u - u_h\|_\infty \approx 7.5$  while  $\|u\|_\infty \approx 1$ , and the large error occurs not only in a singular point but in a part of  $\Omega$ .) This accounts for several unsatisfying results, and should always be kept in mind.

	# Elements	# Nodes	Max. aspect ratio	Nodal $L_\infty$ error
Mesh 1a	2194	1158	811.3	$0.658E - 1$
Mesh 1b	7904	4071	792.5	$0.332E - 1$
Mesh 1c	23444	11959	834.5	$0.240E - 1$
Mesh 2	7796	4020	1092.0	$0.777E - 2$
Mesh 3a	128	81	187.4	$0.102E + 0$
Mesh 3b	512	289	200.7	$0.237E - 1$
Mesh 3c	2048	1089	207.2	$0.603E - 2$
Mesh 3d	8192	4225	210.3	$0.161E - 2$
Mesh 4a	200	121	2.0	$0.275E + 0$
Mesh 4c	320	187	9.1	$0.742E + 1$
Mesh 4e	400	231	81.0	$0.237E + 0$
Mesh 4h	2320	1239	1111.1	$0.524E - 2$

Note that we utilize several meshes and mesh sequences which are independent from each other. Their common feature is that they are all adapted to the anisotropic solution. The advantage of using such a variety of meshes is that one can observe the deviation of certain ratios. This is usually not possible when employing only one mesh sequence.

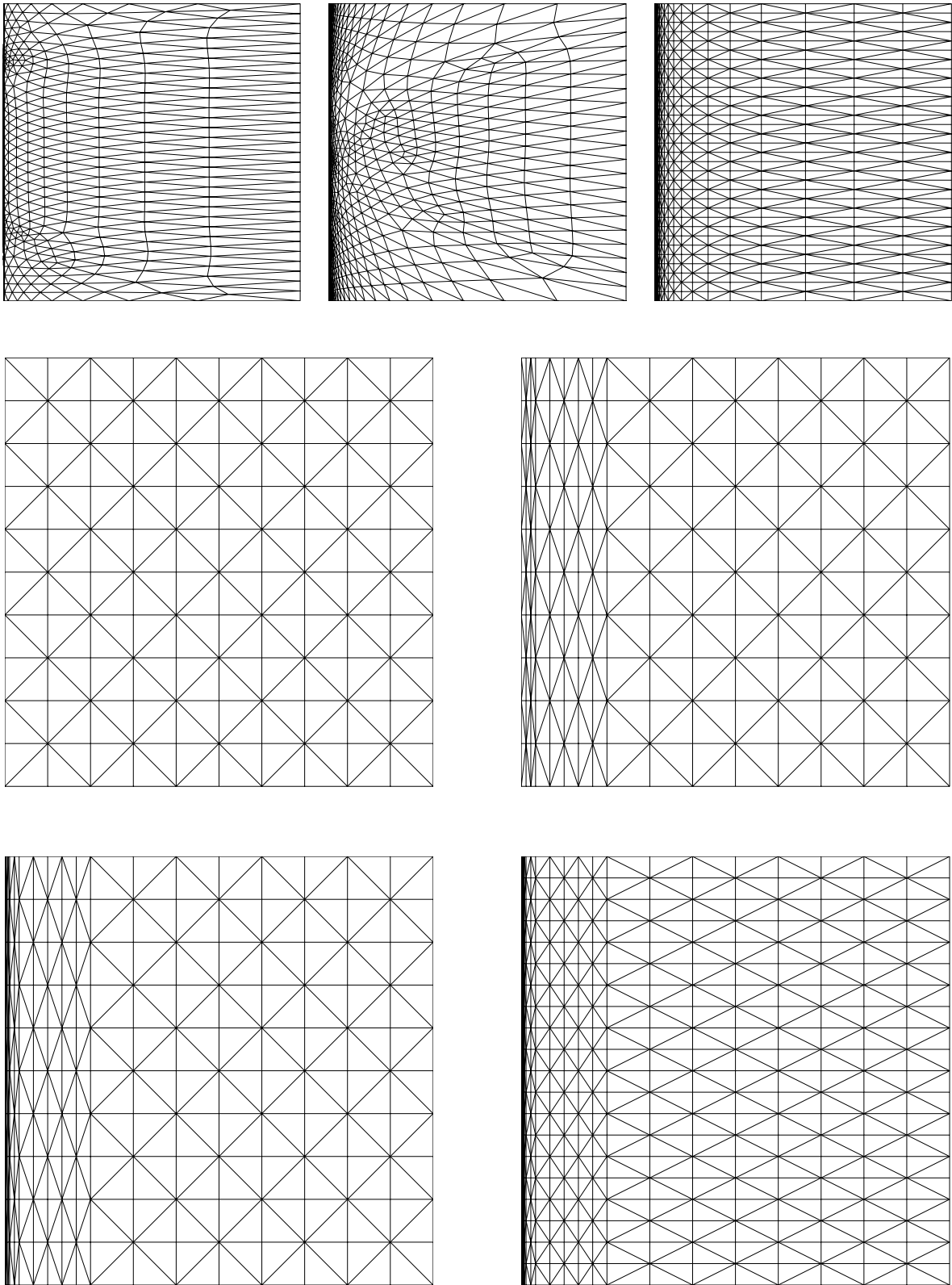


Figure 5.2: Meshes 1a, 2, 3c, 4a,c,e,h

**Matching function  $m_1$** 

We start by evaluating the matching function  $m_1(u - u_h, \mathcal{T}_h) = \|h_{min}^{-1}(\mathbf{x}) C^T(\mathbf{x}) \nabla(u - u_h)\| / \|\nabla(u - u_h)\|$ . For the numerator and the denominator the exact values are computed as well as their respective approximation by means of a recovered gradient. To obtain reliable results, high-order numerical integration is performed.

Firstly we observe that the exact and the approximate values coincide to an acceptable degree which implies  $m_1 \approx m_1^R$ , with a deviation of mostly less than 20%. Even mesh 4c (with an extremely inaccurate solution) yields an acceptable approximation. Secondly, the values of  $m_1$  and  $m_1^R$  are small enough to promise sufficiently reliable interpolation error estimates and finite element error estimates.

	$\ h_{min}^{-1}(\mathbf{x}) C^T(\mathbf{x}) \nabla(u - u_h)\ $		$\ \nabla(u - u_h)\ $		$m_1$	$m_1^R$
	exact	approx	exact	approx		
Mesh 1a	0.118E + 1	0.125E + 1	0.792E + 0	0.918E + 0	1.29	1.36
Mesh 1b	0.686E + 0	0.703E + 0	0.437E + 0	0.500E + 0	1.57	1.21
Mesh 1c	0.486E + 0	0.476E + 0	0.307E + 0	0.337E + 0	1.58	1.41
Mesh 2	0.707E + 0	0.693E + 0	0.386E + 0	0.405E + 0	1.83	1.71
Mesh 3a	0.640E + 1	0.869E + 1	0.401E + 1	0.541E + 1	1.59	1.61
Mesh 3b	0.324E + 1	0.334E + 1	0.198E + 1	0.206E + 1	1.63	1.63
Mesh 3c	0.163E + 1	0.164E + 1	0.991E + 0	0.997E + 0	1.65	1.65
Mesh 3d	0.809E + 0	0.811E + 0	0.492E + 0	0.493E + 0	1.64	1.64
Mesh 4a	0.255E + 2	0.168E + 1	0.161E + 2	0.994E + 0	1.58	1.69
Mesh 4c	0.880E + 2	0.699E + 2	0.380E + 2	0.233E + 2	2.32	3.00
Mesh 4e	0.118E + 2	0.114E + 2	0.526E + 1	0.473E + 1	2.24	2.41
Mesh 4h	0.304E + 1	0.293E + 1	0.110E + 1	0.109E + 1	2.77	2.69

**Interpolation error estimates**

We investigate the interpolation error estimates (3.11)–(3.13) for  $v = u - u_h$ . The corresponding ratios have to be bounded from above which is indeed the case. The quality (i.e. sharpness) of (3.11) apparently depends strongly on the actual mesh.

Estimate	(3.11)	(3.12)	(3.13)
	$\frac{\ v - R_o v\ }{\ v\ }$	$\frac{\ h_{min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\ }{m_1(v, \mathcal{T}_h) \cdot \ v\ }$	$\frac{\ h_{min}^{-1}(\mathbf{x}) \cdot C^T(\mathbf{x}) \nabla(v - R_o v)\ }{m_1(v, \mathcal{T}_h) \cdot \ v\ }$
Mesh 1a	0.182	0.301	1.016
Mesh 1b	0.108	0.240	0.942
Mesh 1c	0.079	0.201	0.918
Mesh 2	0.469	0.140	0.940
Mesh 3a	0.570	0.174	1.011
Mesh 3b	0.410	0.165	1.013
Mesh 3c	0.429	0.162	1.008
Mesh 3d	0.474	0.160	1.005
Mesh 4a	0.844	0.067	1.001
Mesh 4c	0.210	0.213	0.845
Mesh 4e	0.198	0.208	0.993
Mesh 4h	0.635	0.159	0.990

## Finite element error estimates

### Residual error estimator

We consider the upper error bound first and present the values of the error  $\|u - u_h\|$ , its estimation  $\eta_R$ , the approximation error  $\zeta$ , and the matching function  $m_1(u - u_h, \mathcal{T}_h)$ . Estimate (3.23) states that the corresponding ratio has to be bounded which is confirmed by the values of the last column. The approximately equal size of these ratios implies a comparable quality of the error estimation. (The coarse mesh 4a is exceptional because of its inaccurate solution).

Apart from error estimation we observe an error decrease  $\|u - u_h\| \sim N^{-1/2}$  for meshes 3a–d ( $N$  being the number of degrees of freedom). This rate is optimal at least for isotropic meshes [22]. In contrast, meshes 4a–h do not display this decrease, i.e. they are not optimal (in the sense of the convergence order with respect to the energy norm).

	$\ u - u_h\ $	$\eta_R$	$\zeta$	$m_1(u - u_h, \mathcal{T}_h)$	$\frac{\ u - u_h\ }{m_1 \cdot \sqrt{\eta_R^2 + \zeta^2}}$
Mesh 1a	$0.792E + 0$	$0.383E + 1$	$0.163E + 1$	1.29	0.128
Mesh 1b	$0.437E + 0$	$0.214E + 1$	$0.766E + 0$	1.57	0.123
Mesh 1c	$0.307E + 0$	$0.157E + 1$	$0.465E + 0$	1.58	0.119
Mesh 2	$0.386E + 0$	$0.188E + 1$	$0.180E + 0$	1.83	0.112
Mesh 3a	$0.401E + 1$	$0.192E + 2$	$0.333E + 1$	1.59	0.129
Mesh 3b	$0.198E + 1$	$0.982E + 1$	$0.800E + 0$	1.63	0.123
Mesh 3c	$0.991E + 0$	$0.498E + 1$	$0.195E + 0$	1.65	0.121
Mesh 3d	$0.492E + 0$	$0.250E + 1$	$0.487E - 1$	1.64	0.120
Mesh 4a	$0.161E + 2$	$0.497E + 1$	$0.994E + 1$	1.58	0.916
Mesh 4c	$0.380E + 2$	$0.120E + 3$	$0.118E + 3$	2.32	0.098
Mesh 4e	$0.526E + 1$	$0.253E + 2$	$0.575E + 1$	2.24	0.091
Mesh 4h	$0.110E + 1$	$0.402E + 1$	$0.155E + 0$	2.77	0.098

The influence of the approximation  $m_1^R$  of  $m_1$  is given in the next table. Since  $m_1^R \approx m_1$  with a deviation of mostly much less than 20 %, no significant changes can be observed when replacing  $m_1$  by  $m_1^R$ , or even when omitting  $m_1$  completely. On a quantitative scale we note that the error is *overestimated* by  $m_1^R \cdot \sqrt{\eta_R^2 + \zeta^2}$  by about one magnitude.

	$m_1$	$m_1^R$	$\frac{\ u - u_h\ }{m_1 \cdot \sqrt{\eta_R^2 + \zeta^2}}$	$\frac{\ u - u_h\ }{m_1^R \cdot \sqrt{\eta_R^2 + \zeta^2}}$	$\frac{\ u - u_h\ }{\sqrt{\eta_R^2 + \zeta^2}}$
Mesh 1a	1.29	1.36	0.128	0.140	0.190
Mesh 1b	1.57	1.21	0.123	0.137	0.192
Mesh 1c	1.58	1.41	0.119	0.133	0.188
Mesh 2	1.83	1.71	0.112	0.120	0.205
Mesh 3a	1.59	1.61	0.129	0.128	0.205
Mesh 3b	1.63	1.63	0.123	0.124	0.201
Mesh 3c	1.65	1.65	0.121	0.121	0.199
Mesh 3d	1.64	1.64	0.120	0.120	0.197
Mesh 4a	1.58	1.69	0.916	0.859	1.448
Mesh 4c	2.32	3.00	0.098	0.075	0.226
Mesh 4e	2.24	2.41	0.098	0.084	0.203
Mesh 4h	2.77	2.69	0.098	0.101	0.272

The lower error bound (3.22) implies that the corresponding ratio is bounded from above. The maximal values in the table below confirm this prediction. The average value of the ratios shall give some additional information about their distribution. Finally, the small minimal values indicate that the lower error bound is only a one-sided estimate.

	$\frac{\eta_{R,T}(u_h)}{\ u - u_h\ _{\omega_T} + \zeta_T}$		
	Maximum	Average	Minimum
Mesh 1a	3.148	2.044	0.129
Mesh 1b	3.341	2.204	0.136
Mesh 1c	3.390	2.169	0.134
Mesh 2	3.276	2.242	0.248
Mesh 3a	2.460	1.304	0.083
Mesh 3b	2.507	1.557	0.236
Mesh 3c	2.777	1.711	0.168
Mesh 3d	2.961	2.023	0.211
Mesh 4a	2.109	1.169	0.055
Mesh 4c	1.818	0.580	0.090
Mesh 4e	2.080	1.189	0.357
Mesh 4h	3.125	1.789	0.095

### Local problem error estimator

The local problem error estimator admits basically the same conclusions as the residual error estimator (for these meshes). Thus we shorten the explanation and present the corresponding tables, starting with the upper error bound (3.42).

	$\ u - u_h\ $	$\eta_D$	$\zeta$	$m_1(u - u_h, \mathcal{T}_h)$	$\frac{\ u - u_h\ }{m_1 \cdot \sqrt{\eta_D^2 + \zeta^2}}$
Mesh 1a	$0.792E + 0$	$0.112E + 1$	$0.163E + 1$	1.29	0.270
Mesh 1b	$0.437E + 0$	$0.605E + 0$	$0.766E + 0$	1.57	0.286
Mesh 1c	$0.307E + 0$	$0.419E + 0$	$0.465E + 0$	1.58	0.310
Mesh 2	$0.386E + 0$	$0.563E + 0$	$0.180E + 0$	1.83	0.357
Mesh 3a	$0.401E + 1$	$0.560E + 1$	$0.333E + 1$	1.59	0.386
Mesh 3b	$0.198E + 1$	$0.284E + 1$	$0.800E + 0$	1.63	0.412
Mesh 3c	$0.991E + 0$	$0.144E + 1$	$0.195E + 0$	1.65	0.415
Mesh 3d	$0.492E + 0$	$0.718E + 0$	$0.487E - 1$	1.64	0.416
Mesh 4a	$0.161E + 2$	$0.167E + 1$	$0.994E + 1$	1.58	1.010
Mesh 4c	$0.380E + 2$	$0.311E + 2$	$0.118E + 3$	2.32	0.135
Mesh 4e	$0.526E + 1$	$0.670E + 1$	$0.575E + 1$	2.24	0.266
Mesh 4h	$0.110E + 1$	$0.175E + 1$	$0.155E + 0$	2.77	0.225

The approximation  $m_1^R \approx m_1$  implies the same conclusions as before and thus the results are left out. We remark that the error is now usually overestimated by  $m_1^R \cdot \sqrt{\eta_D^2 + \zeta^2}$  by a factor of about 2.5 ... 4. When comparing the residual error estimator  $\eta_R$  with the local problem error estimator  $\eta_D$ , the latter one always gives the more realistic upper bound (i.e. a smaller overestimation).

In contrast to the residual error estimator, the lower error bound (3.41)

$$\eta_{D,T}(u_h) \leq \|u - u_h\|_{\omega_T} + c \cdot \zeta_T$$

now holds with a factor 1. Although the constant  $c$  at  $\zeta_T$  is not known, the corresponding ratio is still less than 1.

	$\frac{\eta_{D,T}(u_h)}{\ u - u_h\ _{\omega_T} + \zeta_T}$		
	Maximum	Average	Minimum
Mesh 1a	0.923	0.580	0.027
Mesh 1b	0.938	0.627	0.016
Mesh 1c	0.944	0.641	0.024
Mesh 2	0.943	0.664	0.040
Mesh 3a	0.821	0.435	0.005
Mesh 3b	0.856	0.493	0.019
Mesh 3c	0.905	0.538	0.071
Mesh 3d	0.907	0.627	0.060
Mesh 4a	0.811	0.440	0.016
Mesh 4c	0.753	0.252	0.029
Mesh 4e	0.843	0.447	0.086
Mesh 4h	0.915	0.661	0.061

### $L_2$ residual error estimator

Here we proceed similar as before, with the difference that we do not know the matching function  $m_2(v_D, \mathcal{T}_h)$ , nor can we approximate it. First we investigate the upper bound (3.59) and compute the following values. We note that the ratio in the last column varies over a quite large range now, which renders the  $L_2$  error estimator less reliable.

	$\ u - u_h\ $	$\eta_{L_2,R}$	$\zeta_{L_2}$	$\ u - u_h\  / \sqrt{\eta_{L_2,R}^2 + \zeta_{L_2}^2}$
Mesh 1a	$0.127E - 1$	$0.454E - 1$	$0.208E - 1$	0.255
Mesh 1b	$0.565E - 2$	$0.131E - 1$	$0.602E - 2$	0.392
Mesh 1c	$0.340E - 2$	$0.520E - 2$	$0.294E - 2$	0.570
Mesh 2	$0.377E - 2$	$0.423E - 1$	$0.728E - 3$	0.089
Mesh 3a	$0.194E - 1$	$0.365E + 0$	$0.108E - 1$	0.053
Mesh 3b	$0.921E - 2$	$0.524E - 1$	$0.393E - 2$	0.175
Mesh 3c	$0.247E - 2$	$0.153E - 1$	$0.264E - 3$	0.161
Mesh 3d	$0.329E - 3$	$0.257E - 2$	$0.229E - 4$	0.128
Mesh 4a	$0.143E + 0$	$0.414E + 0$	$0.703E + 0$	0.176
Mesh 4c	$0.118E + 1$	$0.285E + 1$	$0.130E + 1$	0.376
Mesh 4e	$0.330E - 1$	$0.999E - 1$	$0.777E - 2$	0.329
Mesh 4h	$0.192E - 2$	$0.216E - 1$	$0.387E - 3$	0.089

The lower error bound (3.58) yields the following values from which similar conclusions as before can be drawn. The maximum values of the second column are in accordance with the boundedness of the corresponding ratio. This ratio now varies over a rather large range, as it can be observed by the average and the minimal values. Furthermore, it is clearly seen that (3.58) is only a one-sided estimate.

	$\eta_{L_2,R,T}(u_h)$		
	Maximum	Average	Minimum
Mesh 1a	20.739	3.452	0.046
Mesh 1b	20.787	4.000	0.028
Mesh 1c	21.490	3.289	0.010
Mesh 2	21.236	4.668	0.162
Mesh 3a	8.153	2.611	0.173
Mesh 3b	11.286	3.132	0.294
Mesh 3c	15.340	4.042	0.042
Mesh 3d	15.537	5.095	0.052
Mesh 4a	6.684	1.627	0.212
Mesh 4c	3.519	0.633	0.079
Mesh 4e	6.740	2.126	0.075
Mesh 4h	14.468	3.518	0.021

**Remark 5.2 [Details of the mesh construction]**

Without going into to much detail we sketch here how the aforementioned meshes have been constructed.

In order to obtain meshes 1a–c, we first perform an automatic isotropic mesh generation on the unit square with a local mesh size  $h = 1/30$ ,  $h = 1/60$  and  $h = 1/120$ . In the subsequent nodal coordinate transformation, the  $y$  coordinate is left unchanged. The  $x$  coordinate is transformed via

$$\begin{aligned} \hat{x} \in [0, 3/4] & \quad x := -\frac{3}{2\alpha} \cdot \ln \left[ 1 - \frac{4}{3} \hat{x} \left( 1 - e^{-\frac{2}{3}\alpha z} \right) \right] \\ \hat{x} \in (3/4, 1] & \quad x := z + (1 - z) \cdot \frac{e^{4\beta(\hat{x}-1)} e^{-\beta}}{1 - e^{-\beta}} \quad \beta = 3.0 \quad . \end{aligned}$$

Here,  $z := 2/3 \cdot \ln(\alpha)/\alpha$  marks the transition from the boundary layer to the isotropic region, and  $\alpha = 1000$  is the parameter of the exponential boundary layer. The first transformation forms the mesh inside the boundary layer and aims at approximately equidistributing  $h_{\min,T} \cdot \|\partial^2 u / \partial x^2\|_T$  in order to obtain a comparatively small finite element error in the energy norm. The second transformation shall provide a smooth transition to the isotropic region.

Mesh 2 employs a similar but slightly more sophisticated transformation.

Meshes 3a–d have been constructed based on a similar principle as meshes 1a–c. Here we were aiming to approximately equidistribute  $h_{\min,T}^2 \cdot \|\partial^2 u / \partial x^2\|_T$  in order to obtain a comparatively small finite element error in the  $L_2$  norm. We utilize a tensor product type mesh with  $2m$  intervals in each direction. The  $y$  coordinates are distributed equally. The  $x$  coordinates of boundary layer nodes are given by

$$x_k := -\frac{5}{2\alpha} \cdot \ln \left[ 1 - \frac{k}{m} \left( 1 - e^{-\frac{2}{5} \ln \alpha} \right) \right] \quad \text{for } k = 0 \dots m \quad .$$

The remaining nodes shall provide a smooth transition to the isotropic region. Thus the nodal intervals (in  $x$  direction) first grow geometrically and, eventually, remain constant towards  $x = 1$  (cf. figure 5.2). The values  $m = 4, 8, 16, 32$  have been used to generate the meshes.

□

## 5.2.2 Example 2 (2D; anisotropic solution and misadapted mesh)

### Aim of the example

Consider the previous two-dimensional problem with a boundary layer. We investigate our results on several anisotropic meshes which are, in contrast to the previous example, *not adapted* to this layer.

In particular, we will study the effect of this misadaptation on the matching function  $m_1(u - u_h, \mathcal{T}_h)$ . In turn, this  $m_1$  influences the interpolation error estimates (3.12)–(3.16) as well as the upper bounds on the error, cf. (3.23) and (3.42).

### Boundary value problem

Exactly the same boundary value problem as in example 1 is solved.

### Meshes

In this example anisotropic meshes are used which are not adapted to the anisotropic solution. Such a misadaptation can be caused by several reasons, e.g.

- wrong stretching direction,
- wrong aspect ratio (too large/small),
- wrong element size.

The meshes used here are now briefly described, and some additional information are listed below. Meshes 5a and 5b are pictured in figure 5.3 on the next page, and finally, at the end of this Section, Remark 5.3 presents details of the mesh construction.

For this example, we have restricted ourselves to meshes whose stretching direction is along the  $x$  axis in the vicinity of the boundary layer. Hence the anisotropies of the solution and the mesh are mutually (almost) perpendicular. A sequence of five meshes has been constructed where the maximal aspect ration (which corresponds to the degree of the misadaptation) and the degrees of freedom are increasing, respectively.

We may stress here that all meshes fail to resolve the boundary layer (for both,  $u$  and  $f$ ). Thus, the ‘approximate solution’  $u_h$  is highly inaccurate ( $\|u\|_\infty \approx 1$  whereas  $\|u_h\|_\infty \approx 6$ , and the large error occurs not only in a singular point but in a part of  $\Omega$ ). This should always be kept in mind when interpreting the results. But even here where  $u$  and  $u_h$  have nothing in common, our results remain true.

	# Elements	# Nodes	Max. aspect ratio	Nodal $L_\infty$ error
Mesh 5a	3234	1705	5.1	$0.516E + 1$
Mesh 5b	4736	2501	8.5	$0.511E + 1$
Mesh 5c	7792	4127	21.0	$0.508E + 1$
Mesh 5d	13696	7277	37.2	$0.506E + 1$
Mesh 5e	25778	13717	108.4	$0.505E + 1$

### Matching function $m_1$

First the matching function  $m_1(u - u_h, \mathcal{T}_h) = \|h_{\min}^{-1}(\mathbf{x}) C^T(\mathbf{x}) \nabla(u - u_h)\| / \|\nabla(u - u_h)\|$  is computed. The exact and the approximated values of the numerator differ by a factor of about 2.5. The same, however, holds true for the denominator which implies a surprisingly accurate approximation  $m_1^R \approx m_1$ .

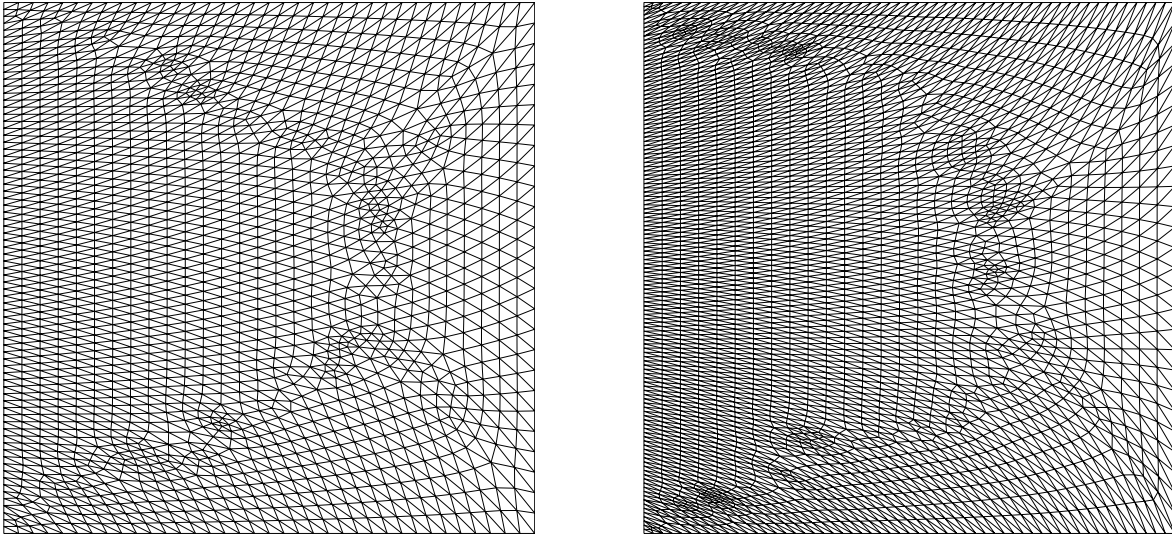


Figure 5.3: Meshes 5a and 5b

Secondly, the worse the mesh is aligned with the boundary layer the larger  $m_1(u - u_h, \mathcal{T}_h)$  becomes. The increasing values of the last two columns below show the strong dependence for our examples.

	$\ h_{min}^{-1}(\mathbf{x}) C^T(\mathbf{x}) \nabla(u - u_h)\ $		$\ \nabla(u - u_h)\ $		$m_1$	$m_1^R$
	exact	approx	exact	approx		
Mesh 5a	$0.713E + 2$	$0.287E + 2$	$0.267E + 2$	$0.108E + 2$	2.67	2.67
Mesh 5b	$0.115E + 3$	$0.475E + 2$	$0.266E + 2$	$0.110E + 2$	4.34	4.31
Mesh 5c	$0.204E + 3$	$0.850E + 2$	$0.264E + 2$	$0.111E + 2$	7.73	7.69
Mesh 5d	$0.383E + 3$	$0.159E + 3$	$0.263E + 2$	$0.111E + 2$	14.52	14.39
Mesh 5e	$0.739E + 3$	$0.306E + 3$	$0.263E + 2$	$0.110E + 2$	28.08	27.79

### Interpolation error estimates

We investigate the interpolation error estimates (3.11)–(3.13) for  $v = u - u_h$ . The corresponding ratios have to be bounded which is indeed the case.

Estimate	(3.11)	(3.12)	(3.13)
	$\frac{\ v - R_o v\ }{\ v\ }$	$\frac{\ h_{min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\ }{m_1(v, \mathcal{T}_h) \cdot \ v\ }$	$\frac{\ h_{min}^{-1}(\mathbf{x}) \cdot C^T(\mathbf{x}) \nabla(v - R_o v)\ }{m_1(v, \mathcal{T}_h) \cdot \ v\ }$
Mesh 5a	0.099	0.166	0.677
Mesh 5b	0.099	0.166	0.683
Mesh 5c	0.099	0.166	0.687
Mesh 5d	0.099	0.165	0.690
Mesh 5e	0.099	0.165	0.691

Of particular interest, however, is now the role that  $m_1$  is playing (again for  $v = u - u_h$ ). Consider, for example, interpolation error estimate (3.12). Since  $\frac{\|h_{min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\|}{m_1(v, \mathcal{T}_h) \cdot \|v\|}$  is virtually constant (middle column), the ratio  $\frac{\|h_{min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\|}{\|v\|}$  behaves exactly like  $m_1$ .

Hence this example strongly suggests that the latter ratio can be unbounded. *Conversely, this means that interpolation estimate (3.12) is valid only with the factor  $m_1$  on the right hand side.*

	$m_1(v, \mathcal{T}_h)$	$\frac{\ h_{\min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\ }{m_1(v, \mathcal{T}_h) \cdot \ v\ }$	$\frac{\ h_{\min}^{-1}(\mathbf{x}) \cdot (v - R_o v)\ }{\ v\ }$
Mesh 5a	2.67	0.166	0.442
Mesh 5b	4.34	0.166	0.720
Mesh 5c	7.73	0.166	1.281
Mesh 5d	14.52	0.165	2.400
Mesh 5e	28.08	0.165	4.638

## Finite element error estimates

### Residual error estimator

For the upper error bound, the values of  $\|u - u_h\|$ ,  $\eta_R$ ,  $\zeta$ , and  $m_1(u - u_h, \mathcal{T}_h)$  are computed as in the previous example. Estimate (3.23) states that the corresponding ratio (cf. last column) has to be bounded which is the case. The decrease of this ratio implies that the quality of the error estimation deteriorates, i.e. the error is more and more overestimated.

	$\ u - u_h\ $	$\eta_R$	$\zeta$	$m_1(u - u_h, \mathcal{T}_h)$	$\frac{\ u - u_h\ }{m_1 \cdot \sqrt{\eta_R^2 + \zeta^2}}$
Mesh 5a	$0.267E + 2$	$0.339E + 2$	$0.517E + 2$	2.67	0.162
Mesh 5b	$0.266E + 2$	$0.214E + 2$	$0.314E + 2$	4.34	0.161
Mesh 5c	$0.264E + 2$	$0.133E + 2$	$0.175E + 2$	7.73	0.156
Mesh 5d	$0.263E + 2$	$0.946E + 1$	$0.926E + 1$	14.52	0.137
Mesh 5e	$0.263E + 2$	$0.808E + 1$	$0.478E + 1$	28.08	0.100

The influence of the approximation  $m_1^R$  of  $m_1$  is given in the next table. Since  $m_1^R \approx m_1$  with a deviation of about 1%, there is virtually no difference when replacing  $m_1$  by  $m_1^R$ . If, however,  $m_1$  is left out completely then the corresponding ratio increases (cf. last column). The growth suggests that this ratio can be unbounded. *This, in turn, strongly indicates that an upper error estimate without  $m_1(u - u_h, \mathcal{T}_h)$  does not hold.*

	$m_1$	$m_1^R$	$\frac{\ u - u_h\ }{m_1 \cdot \sqrt{\eta_R^2 + \zeta^2}}$	$\frac{\ u - u_h\ }{m_1^R \cdot \sqrt{\eta_R^2 + \zeta^2}}$	$\frac{\ u - u_h\ }{\sqrt{\eta_R^2 + \zeta^2}}$
Mesh 5a	2.67	2.67	0.162	0.162	0.432
Mesh 5b	4.34	4.31	0.161	0.162	0.700
Mesh 5c	7.73	7.69	0.156	0.157	1.205
Mesh 5d	14.52	14.39	0.137	0.138	1.989
Mesh 5e	28.08	27.79	0.100	0.101	2.805

Finally, the lower error bound (3.22) is independent of  $m_1$ . The same conclusions as in example 1 can be drawn, and thus the values are omitted here.

### Local problem error estimator

The local problem error estimator admits similar conclusions as the residual error estimator (for these meshes). Thus we shorten the explanation and present the corresponding tables, starting with the upper error bound (3.42). We observe that the ratio of the last

column of the next table is bounded and, moreover, almost constant. Hence the quality of the upper error bound remains the same despite the growing  $m_1$  (note the slight difference to the residual error estimator).

	$\ u - u_h\ $	$\eta_D$	$\zeta$	$m_1(u - u_h, \mathcal{T}_h)$	$\frac{\ u - u_h\ }{m_1 \cdot \sqrt{\eta_D^2 + \zeta^2}}$
Mesh 5a	$0.267E + 2$	$0.138E + 2$	$0.517E + 2$	2.67	0.187
Mesh 5b	$0.266E + 2$	$0.924E + 1$	$0.314E + 2$	4.34	0.187
Mesh 5c	$0.264E + 2$	$0.573E + 1$	$0.175E + 2$	7.73	0.186
Mesh 5d	$0.263E + 2$	$0.386E + 1$	$0.926E + 1$	14.52	0.181
Mesh 5e	$0.263E + 2$	$0.305E + 1$	$0.478E + 1$	28.08	0.165

The approximation  $m_1^R \approx m_1$  implies the same conclusions as before. The error is now overestimated by  $m_1^R \cdot \sqrt{\eta_D^2 + \zeta^2}$  by a factor of about 6. *The increasing ratio of the last column again indicates that an upper error bound without  $m_1(u - u_h, \mathcal{T}_h)$  does not hold.*

	$m_1$	$m_1^R$	$\frac{\ u - u_h\ }{m_1 \cdot \sqrt{\eta_D^2 + \zeta^2}}$	$\frac{\ u - u_h\ }{m_1^R \cdot \sqrt{\eta_D^2 + \zeta^2}}$	$\frac{\ u - u_h\ }{\sqrt{\eta_D^2 + \zeta^2}}$
Mesh 5a	2.67	2.67	0.187	0.187	0.500
Mesh 5b	4.34	4.31	0.187	0.188	0.812
Mesh 5c	7.73	7.69	0.186	0.187	1.437
Mesh 5d	14.52	14.39	0.181	0.182	2.628
Mesh 5e	28.08	27.79	0.165	0.167	4.636

Finally, the lower error bound (3.41) admits exactly the same conclusions as for the first example. For that reason the results are omitted.

### $L_2$ residual error estimator

First we investigate the upper bound (3.59) and compute the following values. We note that the ratio in the last column varies in a small range. This is partly surprising insofar as the meshes are increasingly misaligned with the solution (in the sense of  $m_1$ ). The absolute value of the ratio, however, is much larger than in example 1 which may correspond to the poor alignment. If more details of the  $L_2$  error estimation are sought, further experiments should be carried out.

	$\ u - u_h\ $	$\eta_{L_2,R}$	$\zeta_{L_2}$	$\ u - u_h\  / \sqrt{\eta_{L_2,R}^2 + \zeta_{L_2}^2}$
Mesh 5a	$0.154E + 1$	$0.135E + 1$	$0.651E + 0$	1.025
Mesh 5b	$0.153E + 1$	$0.130E + 1$	$0.242E + 0$	1.158
Mesh 5c	$0.151E + 1$	$0.129E + 1$	$0.755E - 1$	1.176
Mesh 5d	$0.151E + 1$	$0.128E + 1$	$0.213E - 1$	1.176
Mesh 5e	$0.151E + 1$	$0.128E + 1$	$0.568E - 2$	1.176

The lower error bound (3.58) yields the following values from which similar conclusions as before can be drawn.

	Maximum	$\eta_{L_2,R,T}(u_h)$	
		$\ u - u_h\ _{\omega_T} + \zeta_{L_2,T}$	Average
Mesh 5a	2.202	0.100	0.002
Mesh 5b	3.017	0.113	0.001
Mesh 5c	3.917	0.118	0.002
Mesh 5d	4.309	0.124	0.000
Mesh 5e	5.172	0.123	0.001

**Remark 5.3 [Details of the mesh construction]**

The aforementioned meshes were constructed by a combination of an automatic meshing and a transformation. First, a trapeze with vertices  $(0, 0)$ ,  $(1, \beta)$ ,  $(1, \beta + 1)$ , and  $(0, 2\beta + 1)$  has been isotropically meshed with a mesh size  $h \approx 0.04$ . The subsequent nodal transformation

$$(\hat{x}, \hat{y}) \quad \longrightarrow \quad (x, y) := \left( \hat{x}, \frac{\hat{y} - \beta\hat{x}}{1 + 2\beta(1 - \hat{x})} \right)$$

maps the trapeze onto the unit square. For  $\beta = 1, 2, 4, 8, 16$  we thus obtain meshes 5a – 5e.  $\square$

**Remark 5.4** If one has a mesh  $\mathcal{T}_h$  whose anisotropic direction differs significantly from that of an anisotropic solution  $u$  then one may encounter great difficulties to compute even the finite element system accurately. For example, the numerical quadrature of an anisotropic right-hand side  $f$  may yield unsatisfactory results because of the poor representation of  $f$  by  $\mathcal{T}_h$ .  $\square$

**5.2.3 Example 3 (2D; isotropic solution and anisotropic mesh)****Aim of the example**

Here we investigate a problem with an *isotropic* solution which is solved on an *anisotropic* mesh. Note that such a situation constitutes another type of a misadapted mesh.

Our main focus lies on the performance of the error estimator. Thus we omit the interpolation results which are similar to examples 1 and 2.

**Boundary value problem**

The two-dimensional Poisson problem

$$-\Delta u = f \quad \text{in } \Omega = (0, 1) \times (0, 1) \quad , \quad u = 0 \quad \text{on } \Gamma_D = \partial\Omega$$

is chosen as test problem. The exact solution  $u$  is prescribed to be

$$u(x, y) := \frac{\sqrt{27}}{2} x(1-x)(2-x) \cdot \sin(2\pi y)$$

and is of isotropic character. The right-hand side  $f$  is chosen accordingly.

## Meshes

We start by considering meshes 2 and 3d from example 1 which are both anisotropically stretched along the line  $x = 0$ , cf. figure 5.2 on page 99. Since a quasi-optimal mesh is isotropic in nature, both meshes are misadapted. This misadaptation, however, occurs only in the vicinity of the line  $x = 0$ .

The other two anisotropic, triangular meshes are of tensor product type, and have equidistributed nodes along both axes. For mesh 6a, the number of intervals in  $x$  and  $y$  direction is 20 and 200, respectively. For mesh 6b, these numbers are 10 and 400. Thus, both meshes are misadapted over the whole domain  $\Omega$ .

Further details of the meshes are given in the table below.

	# Elements	# Nodes	Max. aspect ratio	Nodal $L_\infty$ error
Mesh 2	7796	4020	1092.0	$0.505E - 1$
Mesh 3d	8192	4225	210.3	$0.359E - 2$
Mesh 6a	8000	4221	10.1	$0.374E - 2$
Mesh 6b	8000	4411	40.0	$0.142E - 1$

## Matching function $m_1$

By construction, all meshes are misadapted to the solution. The values of  $m_1$  reveal that here they depend mainly on two factors:

- the maximum aspect ratio (which is a measure of the misadaptation),
- the domain where the misadaptation occurs.

For meshes 2 and 3d, the misadaptation only occurs in a small part of  $\Omega$  (i.e. in the vicinity of the line  $x = 0$ ) which accounts for the comparatively small  $m_1$ .

Meshes 6a and 6b are misadapted over the whole of  $\Omega$  implying a large  $m_1$ . The larger aspect ratio of mesh 6b also results in a larger  $m_1$ .

	$\ h_{min}^{-1}(\mathbf{x}) C^T(\mathbf{x}) \nabla(u - u_h)\ $		$\ \nabla(u - u_h)\ $		$m_1$	$m_1^R$
	exact	approx	exact	approx		
Mesh 2	$0.105E + 1$	$0.975E + 0$	$0.610E + 0$	$0.604E + 0$	1.72	1.61
Mesh 3d	$0.303E + 0$	$0.302E + 0$	$0.182E + 0$	$0.182E + 0$	1.66	1.66
Mesh 6a	$0.973E + 0$	$0.949E + 0$	$0.232E + 0$	$0.230E + 0$	4.20	4.13
Mesh 6b	$0.744E + 1$	$0.695E + 1$	$0.459E + 0$	$0.449E + 0$	16.21	15.47

## Finite element error estimates

### Residual error estimator

For the upper error bound, the values of  $\|u - u_h\|$ ,  $\eta_R$ ,  $\zeta$ , and  $m_1(u - u_h, \mathcal{T}_h)$  are computed as in the previous example. Estimate (3.23) states that the corresponding ratio (cf. last column) has to be bounded which is the case. The particularly small ratios of mesh 6a and 6b imply a decreasing quality of the upper error bound when  $m_1$  becomes

large. Hence the error is much overestimated by  $m_1 \cdot \sqrt{\eta_R^2 + \zeta^2}$ .

	$\ u - u_h\ $	$\eta_R$	$\zeta$	$m_1(u - u_h, \mathcal{T}_h)$	$\frac{\ u - u_h\ }{m_1 \cdot \sqrt{\eta_R^2 + \zeta^2}}$
Mesh 2	$0.610E + 0$	$0.260E + 1$	$0.124E + 0$	1.72	0.137
Mesh 3d	$0.182E + 0$	$0.732E + 0$	$0.116E - 1$	1.66	0.150
Mesh 6a	$0.232E + 0$	$0.749E + 0$	$0.397E - 2$	4.20	0.074
Mesh 6b	$0.459E + 0$	$0.143E + 1$	$0.387E - 2$	16.21	0.020

The last column of the next table suggests that, for these particular examples, the factor  $m_1$  can be omitted in the upper error bound. We have not found out yet whether this is by accident or not. Further investigations are necessary to clarify this phenomenon.

	$m_1$	$m_1^R$	$\frac{\ u - u_h\ }{m_1 \cdot \sqrt{\eta_R^2 + \zeta^2}}$	$\frac{\ u - u_h\ }{m_1^R \cdot \sqrt{\eta_R^2 + \zeta^2}}$	$\frac{\ u - u_h\ }{\sqrt{\eta_R^2 + \zeta^2}}$
Mesh 2	1.72	1.61	0.137	0.145	0.235
Mesh 3d	1.66	1.66	0.150	0.150	0.249
Mesh 6a	4.20	4.13	0.074	0.075	0.309
Mesh 6b	16.21	15.47	0.020	0.021	0.320

The lower error bound (3.22) yields the same conclusions as before; the results are thus omitted.

### Local problem error estimator

The results for the local problem error estimator are analogous to the residual error estimator and therefore left out.

### $L_2$ residual error estimator

First we investigate the upper bound (3.59) and compute the following values. The ratios of the last column are in a similar range as the corresponding ones of example 1. This may be surprising since our meshes here, and meshes 6a and 6b in particular, are rather misadapted (in the sense of  $m_1$ ). Nevertheless they admit a useful  $L_2$  error estimation for this example.

	$\ u - u_h\ $	$\eta_{L_2,R}$	$\zeta_{L_2}$	$\ u - u_h\  / \sqrt{\eta_{L_2,R}^2 + \zeta_{L_2}^2}$
Mesh 2	$0.282E - 1$	$0.153E + 0$	$0.581E - 2$	0.184
Mesh 3d	$0.228E - 2$	$0.147E - 1$	$0.169E - 3$	0.155
Mesh 6a	$0.313E - 2$	$0.216E - 1$	$0.197E - 4$	0.145
Mesh 6b	$0.121E - 1$	$0.815E - 1$	$0.967E - 5$	0.149

The lower error bound (3.58) yields similar conclusions as before. The table is thus omitted.

## 5.2.4 Example 4 (3D; anisotropic solution and adapted mesh)

### Aim of the example

Consider the following three-dimensional problem which exhibits an edge singularity. Mixed (Neumann and Dirichlet) boundary conditions are imposed. We investigate our

results on a sequence of anisotropic meshes which are *adapted* to this singularity. The results will show that the matching function  $m_1$  is small, that it can be approximated fairly well, and that the finite element error estimates hold. The computations are carried out on a serial as well as on a parallel machine.

### Boundary value problem

When solving a Poisson problem in three-dimensional domains  $\Omega$ , a typical occurrence of an anisotropic solution is induced by an edge with an angle  $\omega > \pi$ , and/or by a change of the boundary conditions, cf. also [5, 7, 10]. Therefore we choose the following test problem which has already been discussed in [5] and the preprint version of [10] in a slightly different form. Solve the three-dimensional Poisson problem

$$-\Delta u = f \quad \text{in } \Omega \quad , \quad u = u_0 \quad \text{on } \Gamma_D \quad , \quad \frac{\partial u}{\partial n} = g_N \quad \text{on } \Gamma_N \quad .$$

The domain  $\Omega$  consists of three quarters of a cylinder, i.e.

$$\Omega = \{(r \cos \varphi, r \sin \varphi, z) \in \mathbb{R}^3 \quad : \quad 0 < r, z < 1, 0 < \varphi < 3\pi/2\} \quad ,$$

cf. also figure 5.4. The Neumann boundary  $\Gamma_N$  shall consist of the top and bottom plane of  $\Omega$ , and of the plane described by  $x = 0$ . Let the Dirichlet boundary be  $\Gamma_D := \partial\Omega \setminus \Gamma_N$ .

The exact solution  $u$  (in cylindrical coordinates) is prescribed to be

$$\begin{aligned} u(r, \varphi, z) &:= (1 + \gamma(z)) \cdot r^\lambda \cdot \sin(\lambda\varphi) && \text{with } \lambda = 1/3 \\ \text{and } \gamma(z) &:= \begin{cases} (2z - 1) \cdot 2z & \text{when } z \in [0, 1/2] \\ (2z - 1) \cdot (3 - 4z) & \text{when } z \in (1/2, 1] \end{cases} \quad . \end{aligned}$$

The corresponding right-hand side  $f = -\Delta u$  is in  $L_p(\Omega)$  for all  $p \in [1, \infty]$ , but it has a jump at  $z = 1/2$ . The boundary conditions  $u_0$  and  $g_N$  are chosen according to  $u$ .

The exact solution  $u$  displays an edge singularity of the type  $r^\lambda$ . This implies an anisotropy of  $u$  along the  $z$ -axis.

### Meshes

Here we utilize a single sequence of meshes which are constructed as follows. First, the domain  $\Omega$  is *isotropically* and quasi-uniformly meshed, with  $h \sim 2^{-k}$ ,  $k = 0, 1, 2, \dots$  (note that the curved boundary is approximated). The final, *anisotropic* mesh is obtained by the subsequent nodal coordinate transformation (also known as mesh grading)

$$\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} \longrightarrow \begin{pmatrix} x \\ y \end{pmatrix} := \hat{r}^{\frac{1}{\mu}-1} \cdot \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} \quad \text{with } \hat{r} = \sqrt{\hat{x}^2 + \hat{y}^2} \quad . \quad (5.1)$$

This ensures the adaption of the mesh to  $u$ . The actual grading depends on a grading parameter  $\mu$ . Apel [5] has shown that  $\mu < \lambda$  guarantees the optimal rate of convergence in the energy norm, i.e.  $\|u - u_h\| \sim N^{-1/3}$ , with  $N$  being the degrees of freedom.

For our examples here we have chosen  $\mu = 0.3$ . The corresponding meshes before and after the mesh grading are depicted in figure 5.4. Some additional information on our meshes are given in the table below.

	# Elements	# Nodes	Max. aspect ratio	Nodal $L_\infty$ error
Mesh 7a	12	12	2.2	— <sup>a</sup>
Mesh 7b	96	45	9.7	$0.236E + 0$
Mesh 7c	768	225	30.6	$0.116E + 0$
Mesh 7d	6144	1377	154.0	$0.417E - 1$
Mesh 7e	49152	9537	775.8	$0.202E - 1$
Mesh 7f	393216	70785	3910.0	$0.954E - 2$
Mesh 7g	3145728	545025	19705.1	$0.444E - 2$

<sup>a</sup>All nodes are on  $\Gamma_D$

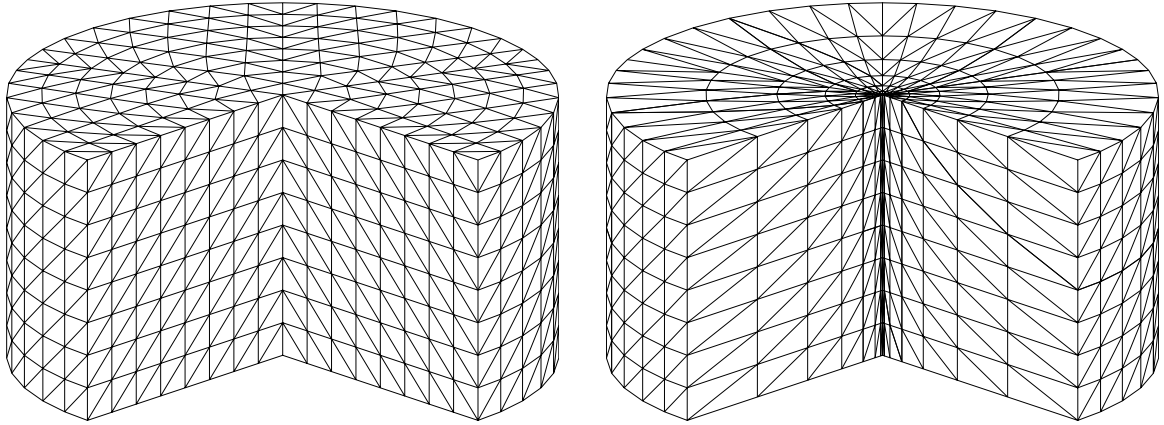


Figure 5.4: Mesh 7d before and after mesh grading

### Matching function $m_1$

We start by evaluating the matching function  $m_1(u - u_h, \mathcal{T}_h) = \|h_{min}^{-1}(\mathbf{x}) C^T(\mathbf{x}) \nabla(u - u_h)\| / \|\nabla(u - u_h)\|$ . Firstly, the values of  $m_1$  are quite small and thus promise useful and reliable upper bounds of the error. Since  $m_1$  is almost constant for the mesh sequence we anticipate a comparable quality of the error estimation.

Secondly, the numerator is computed exactly as well as approximated by means of a recovered gradient (second and third column). Both values differ by a factor of about 2.5. The same, however, holds true for the denominator. Hence the approximation  $m_1^R \approx m_1$  is nevertheless quite good. Only the two coarsest meshes yield a deviation of about 30%.

	$\ h_{min}^{-1}(\mathbf{x}) C^T(\mathbf{x}) \nabla(u - u_h)\ $		$\ \nabla(u - u_h)\ $		$m_1$	$m_1^R$
	exact	approx	exact	approx		
Mesh 7a	$0.298E + 1$	$0.328E + 0$	$0.183E + 1$	$0.281E + 0$	1.63	1.17
Mesh 7b	$0.347E + 1$	$0.459E + 0$	$0.154E + 1$	$0.295E + 0$	2.26	1.56
Mesh 7c	$0.181E + 1$	$0.651E + 0$	$0.831E + 0$	$0.334E + 0$	2.18	1.95
Mesh 7d	$0.912E + 0$	$0.351E + 0$	$0.435E + 0$	$0.180E + 0$	2.10	1.95
Mesh 7e	$0.460E + 0$	$0.181E + 0$	$0.223E + 0$	$0.924E - 1$	2.06	1.96
Mesh 7f	$0.233E + 0$	$0.925E - 1$	$0.113E + 0$	$0.467E - 1$	2.06	1.98
Mesh 7g	$0.118E + 0$	$0.470E - 1$	$0.573E - 1$	$0.235E - 1$	2.06	2.00

## Finite element error estimates

### Residual error estimator

We consider the upper error bound first and present the values of the error  $\|u - u_h\|$ , its estimation  $\eta_R$ , the approximation error  $\zeta$ , and the matching function  $m_1(u - u_h, \mathcal{T}_h)$ . Estimate (3.23) states that the corresponding ratio has to be bounded which is confirmed by the values of the last column. The small variation of these ratios yields a comparable quality of the error estimation, as promised by the matching function. Apart from mesh 7a, the error is overestimated by  $m_1 \cdot \sqrt{\eta_R^2 + \zeta^2}$  by a factor of  $8 \dots 10$ . The exception for the coarsest mesh 7a stems from the large approximation error  $\zeta$ .

Additionally, the error decreases with a rate of about  $N^{-0.32}$  which is very close to the optimal order of  $N^{-1/3}$ . Hence the meshes chosen are quasi-optimal (with respect to the convergence order of the error).

	$\ u - u_h\ $	$\eta_R$	$\zeta$	$m_1(u - u_h, \mathcal{T}_h)$	$\frac{\ u - u_h\ }{m_1 \cdot \sqrt{\eta_R^2 + \zeta^2}}$
Mesh 7a	$0.183E + 1$	$0.722E + 1$	$0.108E + 2$	1.63	0.086
Mesh 7b	$0.154E + 1$	$0.542E + 1$	$0.109E + 1$	2.26	0.123
Mesh 7c	$0.831E + 0$	$0.333E + 1$	$0.335E + 0$	2.18	0.114
Mesh 7d	$0.435E + 0$	$0.187E + 1$	$0.995E - 1$	2.10	0.111
Mesh 7e	$0.223E + 0$	$0.991E + 0$	$0.296E - 1$	2.06	0.109
Mesh 7f	$0.113E + 0$	$0.512E + 0$	$0.919E - 2$	2.06	0.108
Mesh 7g	$0.573E - 1$	$0.260E + 0$	$0.299E - 2$	2.06	0.107

The lower error bound (3.22) implies that the corresponding ratio is bounded from above. The maximal values confirm this prediction. The average value of the ratios shall give some additional information about their distribution. Finally, the small minimal values indicate that the lower error bound is only a one-sided estimate.

	$\frac{\eta_{R,T}(u_h)}{\ u - u_h\ _{\omega_T} + \zeta_T}$		
	Maximum	Average	Minimum
Mesh 7a	0.803	0.585	0.435
Mesh 7b	2.679	1.349	0.238
Mesh 7c	3.802	1.556	0.239
Mesh 7d	4.042	1.632	0.200
Mesh 7e	3.929	1.685	0.142
Mesh 7f	3.944	1.725	0.052
Mesh 7g	3.953	1.742	0.024

### Local problem error estimator

The local problem error estimator admits basically the same conclusions as the residual error estimator (for these meshes). Thus we shorten the explanation and present the corresponding tables, starting with the upper error bound (3.42). Firstly, the quality of the error estimation is almost constant, and the error is overestimated by  $m_1 \cdot \sqrt{\eta_D^2 + \zeta^2}$  by a factor of about 2.

Secondly, when comparing the residual error estimator  $\eta_R$  with the local problem error estimator  $\eta_D$ , the latter one always gives the more realistic upper bound (i.e. a smaller

overestimation).

	$\  \ u - u_h\  \ $	$\eta_D$	$\zeta$	$m_1(u - u_h, \mathcal{T}_h)$	$\frac{\  \ u - u_h\  \ }{m_1 \cdot \sqrt{\eta_D^2 + \zeta^2}}$
Mesh 7a	$0.183E + 1$	$0.147E + 1$	$0.108E + 2$	1.63	0.103
Mesh 7b	$0.154E + 1$	$0.115E + 1$	$0.109E + 1$	2.26	0.429
Mesh 7c	$0.831E + 0$	$0.721E + 0$	$0.335E + 0$	2.18	0.479
Mesh 7d	$0.435E + 0$	$0.408E + 0$	$0.995E - 1$	2.10	0.494
Mesh 7e	$0.223E + 0$	$0.218E + 0$	$0.296E - 1$	2.06	0.490
Mesh 7f	$0.113E + 0$	$0.114E + 0$	$0.919E - 2$	2.06	0.483
Mesh 7g	$0.573E - 1$	$0.583E - 1$	$0.299E - 2$	2.06	0.477

The lower error bound admits analogous conclusions as the residual error estimator, and the following table is included for completeness. (The Remarks of example 1 may deepen the understanding.)

	$\frac{\eta_{D,T}(u_h)}{\  \ u - u_h\  \ _{\omega_T} + \zeta_T}$		
	Maximum	Average	Minimum
Mesh 7a	0.145	0.132	0.119
Mesh 7b	0.697	0.322	0.088
Mesh 7c	0.760	0.383	0.047
Mesh 7d	0.784	0.409	0.041
Mesh 7e	0.791	0.427	0.041
Mesh 7f	0.791	0.437	0.019
Mesh 7g	0.788	0.442	0.009

In Theorem 3.6 on page 61 the equivalence of  $\eta_R$  and  $\eta_D$  has been shown. The maximum of the corresponding ratios has to be bounded from above which is confirmed by the following table.

Finally, we present the maximal condition number  $\kappa_{max}$  of the local problem. The boundedness of  $\kappa_{max}$  has been proven by Theorem 3.8 on page 65, and is shown below.

Maximum of	$\eta_{R,T}/\eta_{D,\omega_T}$	$\eta_{D,T}/\eta_{R,\omega_T}$		$\kappa_{max}$
Mesh 7a	4.233	0.158	Mesh 7a	13.995
Mesh 7b	4.246	0.201	Mesh 7b	23.556
Mesh 7c	4.199	0.214	Mesh 7c	24.183
Mesh 7d	4.220	0.238	Mesh 7d	24.211
Mesh 7e	4.083	0.259	Mesh 7e	24.212
Mesh 7f	3.979	0.269		
Mesh 7g	3.925	0.270		

### Three Zienkiewicz-Zhu like error estimators

The ratios in the table below suggest that all three error estimators yield an upper bound of the error for our examples. Note, however, that the theory is not wholly understood yet. The error is *underestimated* in all cases, and  $\eta_{Z_3}$  performs best here (the exceptional values of mesh 7a stem from the coarse mesh that implies a dominating approximation term  $\zeta$ ). Further conclusions are impossible; this would require an extensive coverage of all relevant situations (in particular of meshes which are not (well) adapted to an anisotropic solution).

	$\frac{\ \nabla(u - u_h)\ }{\sqrt{\eta_{Z_1}^2 + \zeta^2}}$	$\frac{\ \nabla(u - u_h)\ }{\sqrt{\eta_{Z_2}^2 + \zeta^2}}$	$\frac{\ \nabla(u - u_h)\ }{\sqrt{\eta_{Z_3}^2 + \zeta^2}}$
Mesh 7a	0.169	0.169	0.169
Mesh 7b	1.371	1.357	1.296
Mesh 7c	1.809	1.756	1.135
Mesh 7d	2.230	2.114	1.191
Mesh 7e	2.470	2.298	1.214
Mesh 7f	2.605	2.382	1.220
Mesh 7g	2.677	2.419	1.218

### $L_2$ residual error estimator

Since the domain  $\Omega$  is not convex, this example is not covered by the  $L_2$  error estimation theory. Even so it is interesting to observe the  $L_2$  error estimator for such a (practically relevant) problem.

First we investigate the upper bound (3.59). The matching function  $m_2(v_D, \mathcal{T}_h)$  is neither known nor evaluated or approximated. Nevertheless we expect  $m_2$  to be quite small (analogously to  $m_1$ ) since the meshes are adapted to the anisotropic solution  $u$ .

The last column indicates that the error  $\|u - u_h\|$  is overestimated by  $\sqrt{\eta_{L_2,R}^2 + \zeta_{L_2}^2}$  by a factor of about 8 (apart from the coarsest mesh). The quality of the error estimation is comparable for all meshes.

The second column yields a convergence rate of about  $N^{-0.64}$  for the error in the  $L_2$  norm. This is very close to the optimal rate of  $N^{-2/3}$  which further strengthens that these meshes are quasi-optimal.

	$\ u - u_h\ $	$\eta_{L_2,R}$	$\zeta_{L_2}$	$\ u - u_h\  / \sqrt{\eta_{L_2,R}^2 + \zeta_{L_2}^2}$
Mesh 7a	$0.359E + 0$	$0.485E + 1$	$0.732E + 1$	0.041
Mesh 7b	$0.248E + 0$	$0.215E + 1$	$0.448E + 0$	0.113
Mesh 7c	$0.792E - 1$	$0.654E + 0$	$0.672E - 1$	0.120
Mesh 7d	$0.230E - 1$	$0.184E + 0$	$0.100E - 1$	0.124
Mesh 7e	$0.625E - 2$	$0.486E - 1$	$0.149E - 2$	0.129
Mesh 7f	$0.165E - 2$	$0.124E - 1$	$0.232E - 3$	0.133
Mesh 7g	$0.431E - 3$	$0.314E - 2$	$0.378E - 4$	0.137

Finally, the lower error bound (3.58) yields the following values from which similar conclusions as in example 1 can be drawn.

	$\frac{\eta_{L_2,R,T}(u_h)}{\ u - u_h\ _{\omega_T} + \zeta_{L_2,T}}$		
	Maximum	Average	Minimum
Mesh 7a	0.832	0.605	0.440
Mesh 7b	9.097	2.459	0.121
Mesh 7c	19.171	2.886	0.056
Mesh 7d	23.223	3.020	0.043
Mesh 7e	28.130	3.049	0.022
Mesh 7f	29.431	3.080	0.006
Mesh 7g	29.659	3.019	0.002

## 5.3 Conclusions

Before summarizing the numerical results, recall that examples 1 and 4 have utilized meshes which were *adapted* to the anisotropic solution ( $\Omega \subset \mathbb{R}^2$  and  $\Omega \subset \mathbb{R}^3$ ). In contrast, examples 2 and 3 featured meshes which were *not adapted* to the solution (e.g. because of a different stretching direction, or since the aspect ratio has been too large).

### The matching function $m_1$

Firstly, let us consider the size of  $m_1(u - u_h, \mathcal{T}_h)$ . When using adapted meshes (examples 1 and 4) then  $m_1$  is quite small and mostly in the range of  $1.5 \dots 3$ . This promises useful and reliable estimates. For misadapted meshes  $m_1$  can still be small (example 3) but it can also become large (example 2). In the latter case one can expect the quality of certain estimates to deteriorate.

Secondly the approximation  $m_1^R$  of  $m_1$  by means of a simple gradient recovery has been investigated. Despite the simple (and not theoretically proven) principle we always obtained fairly good approximations  $m_1^R \approx m_1$ , with a deviation of mostly (much) less than 30%. We consider this as quite good and promising, in particular since this seems to hold for both, adapted and misadapted meshes.

### Interpolation error estimates

Examples 1 and 2 show the anticipated theoretical behaviour. Additionally the values of experiment 2 suggest numerically that the factor  $m_1(v, \mathcal{T}_h)$  is indeed necessary for estimate (3.12) to hold. This coincides with the analytical result of Remark 3.3 on page 44.

### Residual error estimator $\eta_R$

We firstly consider the upper error bound (3.23). In all examples the corresponding ratio has been bounded from above which is in accordance with the theory. The global error is always overestimated by  $m_1(u - u_h, \mathcal{T}_h) \cdot \sqrt{\eta_R^2 + \zeta^2}$ . For adapted meshes, this overestimation is in the range of about  $7 \dots 11$ . For misadapted meshes the overestimation can be similar (example 2), or it can become large and unsatisfactory (example 3).

Secondly, example 2 indicates that an upper error bound without  $m_1(u - u_h, \mathcal{T}_h)$  does not hold.

Lastly, all experiments show that the local lower error bound holds unconditionally, and that it is a one-sided estimate indeed.

### Local problem error estimator $\eta_D$

On the qualitative side exactly the same conclusions can be drawn as for the residual error estimator  $\eta_R$ . Quantitatively there is still an overestimation of the error by  $m_1 \cdot \sqrt{\eta_D^2 + \zeta^2}$  but this overestimation is always smaller than for  $\eta_R$ . *Hence the local problem error estimation performs better than the residual error estimator.*

The lower error bound holds here with a factor 1 which is in accordance with the theoretical prediction.

The equivalence of  $\eta_R$  and  $\eta_D$  (cf. Theorem 3.6 on page 61) can be seen in example 4. Finally, the same example exhibits the bounded condition number  $\kappa$  of the local problems.

### Three Zienkiewicz-Zhu like error estimators

The three estimators  $\eta_{Z_1}, \eta_{Z_2}, \eta_{Z_3}$  were tested in example 4. All of them seem to yield an upper bound of the error. Further conclusions require an extensive numerical investigation and are thus not possible (and intended) here.

#### $L_2$ error estimator

The upper bound (3.59) contains the matching function  $m_2(v_D, \mathcal{T}_h)$ . Since  $m_2$  is not known we had to proceed without it.

When using adapted meshes (examples 1 and 4) then the corresponding ratio  $\|u - u_h\| / \sqrt{\eta_{L_2,R}^2 + \zeta_{L_2}^2}$  is bounded. The error is overestimated by a factor of about  $2 \dots 10$ . Hence the  $L_2$  error estimate seems to be less reliable than the residual error estimator for the energy norm, for example. Furthermore we conjecture that an adapted mesh results in a relatively small matching function  $m_2$  which yields at least acceptable  $L_2$  error estimates.

The misadapted meshes of example 3 perform similar to the adapted meshes whereas in example 2 the estimate seems to be unreliable because of the much greater ratio.

All numerical experiments display an unconditional lower error bound which coincides with the theory.

Finally, we observe that generally the  $L_2$  error estimation (lower and upper bound) seems to be less reliable than the error estimation in the energy norm.

# Summary

This work has been aiming at *a posteriori* error estimators suitable for anisotropic tetrahedral or triangular grids, respectively.

Chapter 1 has given a brief introduction to that topic. In chapter 2 the notation has been introduced, and some basic tools have been derived. Amongst them, the transformation technique (here with two different transformations) is particularly important.

Chapter 3 has been devoted to the Poisson equation which serves as model problem. There we have studied the effects and difficulties of error estimation that arise from an anisotropic mesh. The main results are as follows.

A residual error estimator and a local problem error estimator have been derived. Upper and lower bounds are proven for the error in the energy norm. Both estimators are suitable for three kinds of boundary conditions. As a by-product, Robin boundary conditions can now be treated. As far as we know, this result is new also for isotropic meshes.

Three different Zienkiewicz-Zhu like error estimators have been proposed. One of them is shown to be equivalent to a modified residual error estimator. An  $L_2$  error estimator and corresponding upper and lower error bounds in the  $L_2$  norm have been presented.

It turns out that an effective error estimation requires an anisotropic mesh that corresponds (in some sense) with the anisotropic function under consideration. To measure this correspondence, a so-called matching function has been introduced. Comparing with isotropic error estimation, this feature is new, and it seems to reflect the nature of anisotropic error estimators.

In chapter 4 a singularly perturbed reaction-diffusion equation has been considered. Its solution can partially display the typical behaviour of real-world problems, but at the same time we are still able to analyse this model problem. A residual error estimator has been derived, and upper and lower bounds error bounds are proven for the energy norm.

The numerical experiments have been described in chapter 5. The different error estimators and related aspects have been investigated on numerous examples. They demonstrate that reliable and efficient error estimation is possible. Additionally it has been shown that error estimation can be parallelized effectively.

## Final Remarks and some open problems

The present state of research suggests that the following topics could be investigated in the future, or that they could enhance the understanding of anisotropic error estimation.

1. Other techniques may be suitable for anisotropic error estimation. For example, a superconvergence analysis for anisotropic meshes is certainly of interest. This

could contribute to an extension of the ‘superconvergent patch recovery’ [67] to our meshes.

2. Error estimators based on the ‘complementary energy principle’ (see e.g. [1]) could possibly yield upper bounds on the error with a constant 1.
3. Anisotropic error estimation should be extended to further problem classes. The application to real-world problems is certainly a great challenge.
4. Error estimators should be incorporated into a fully adaptive strategy. Hence the extraction of mesh information as well as the subsequent mesh refinement has to be studied more thoroughly.

# Bibliography

- [1] M. Ainsworth and J. Oden. A posteriori error estimation in finite element analysis. *Comput. Methods Appl. Mech. Eng.*, 142(1-2):1–88, 1997.
- [2] M. Ainsworth and J. T. Oden. A unified approach to a posteriori error estimation using element residual methods. *Numer. Math.*, 65:23–50, 1993.
- [3] L. Angermann. Balanced a-posteriori error estimates for finite volume type discretizations of convection-dominated elliptic problems. *Computing*, 55(4):305–323, 1995.
- [4] T. Apel. SPC-PM Po3D – User’s manual. Preprint SPC 95\_33, Technische Universität Chemnitz–Zwickau, Fakultät für Mathematik, 1995.
- [5] T. Apel. Interpolation of non-smooth functions on anisotropic finite element meshes. Preprint SFB393/97-6, TU Chemnitz-Zwickau, 1997. Submitted to *Math. Modeling Numer. Anal.*
- [6] T. Apel. Anisotropic finite elements: Local estimates and applications. Habilitationsschrift, TU Chemnitz, 1998.
- [7] T. Apel and M. Dobrowolski. Anisotropic interpolation with applications to the finite element method. *Computing*, 47:277–293, 1992.
- [8] T. Apel and G. Lube. Anisotropic mesh refinement in stabilized Galerkin methods. *Numer. Math.*, 74(3):261–282, 1996.
- [9] T. Apel, F. Milde, and M. Theß. SPC-PM Po3D – Programmer’s manual. Preprint SPC 95\_34, Technische Universität Chemnitz–Zwickau, Fakultät für Mathematik, 1995.
- [10] T. Apel and S. Nicaise. Elliptic problems in domains with edges: anisotropic regularity and anisotropic finite element meshes. In J. Cea, D. Chenais, G. Geymonat, and J. L. Lions, editors, *Partial Differential Equations and Functional Analysis (In Memory of Pierre Grisvard)*, pages 18–34. Birkhäuser, Boston, 1996. Shortened version of Preprint SPC94\_16, TU Chemnitz-Zwickau, 1994.
- [11] J. P. Aubin. Behaviour of the error of the approximate solution of boundary value problems for linear elliptic operators by Galerkin’s and finite difference methods. *Ann. Scuola Norm. Sup. Pisa*, 21:599–637, 1967.
- [12] J. P. Aubin and H. G. Burchard. Some aspects of the method of the hypercircle applied to elliptic variational problems. In *Proc. 2nd Sympos. numerical solution partial diff. equations, SYNSPADE 1970*, pages 1–67. Univ. Maryland, 1971.

- [13] I. Babuška and W. C. Rheinboldt. A-posteriori error estimates for the finite element method. *Int. J. Num. Meth. Eng.*, 12:1597–1615, 1978.
- [14] I. Babuška and W. C. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.*, 15(4):736–754, 1978.
- [15] 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 Appl. Mech. Eng.*, 114:307–378, 1994.
- [16] R. E. Bank and A. Weiser. Some a posteriori error estimators for elliptic partial differential equations. *Math. Comput.*, 44(170):283–301, 1985.
- [17] R. Becker and R. Rannacher. A feed-back approach to error control in finite element methods: Basic analysis and examples. *East-West J. Numer. Math.*, 4(4):237–264, 1996.
- [18] R. Becker and R. Rannacher. Weighted a posteriori error control in FE methods. Report 96-01, SFB 359, Heidelberg University, 1996.
- [19] R. Beinert and D. Kröner. Finite volume methods with local mesh alignment in 2-D. In *Adaptive Methods – Algorithms, Theory and Applications*, volume 46 of *Notes on Numerical Fluid Mechanics*, pages 38–53, Braunschweig, 1994. Vieweg.
- [20] F. J. Bossen and P. S. Heckbert. A pliant method for anisotropic mesh generation. In *Proceedings of the 5<sup>th</sup> Annual International Meshing Roundtable*, Pittsburgh, PA, 1996. Sandia national Laboratories.
- [21] J. U. Brackbill. An adaptive grid with directional control. *J. Comput. Physics*, 108:38–50, 1993.
- [22] D. Braess. *Finite Elemente – Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer Lehrbuch, second edition, 1997.
- [23] C. Carstensen and R. Verfürth. Edge residuals dominate a posteriori error estimates for low order finite element methods. *SIAM J. Numer. Anal.*, submitted.
- [24] M. J. Castro Díaz and F. Hecht. Anisotropic surface mesh generation. Report 2672, INRIA, 1995.
- [25] M. J. Castro-Díaz, F. Hecht, and B. Mohammadi. New progress in anisotropic grid adaptation for inviscid and viscous flow simulations. In *Proceedings of the 4<sup>th</sup> Annual International Meshing Roundtable*, pages 73–85, Albuquerque, NM, 1995. Sandia national Laboratories. Also Report 2671 at INRIA.
- [26] P. Ciarlet. *The finite element method for elliptic problems*. North-Holland Publishing Company, Amsterdam, 1978.
- [27] P. G. Ciarlet and J. L. Lions, editors. *Handbook of Numerical Analysis, Vol. II/IV*. North-Holland, Amsterdam, 1991/95.
- [28] P. Clément. Approximation by finite element functions using local regularization. *RAIRO Anal. Numér.*, 2:77–84, 1975.

- [29] K. Eriksson and C. Johnson. Adaptive finite element methods for parabolic problems I: A linear model problem. *SIAM J. Num. Anal.*, 28:43–77, 1991.
- [30] J. Fröhlich, J. Lang, and R. Roitzsch. Selfadaptive finite element computations with smooth time controller and anisotropic refinement. Report 96–16, ZIB, 1996.
- [31] W. Hackbusch. The frequency decomposition multi-grid method. Part I Application to anisotropic equations. *Numer. Math.*, 56:229–245, 1989.
- [32] C. Johnson. The characteristic streamline diffusion finite element method. *Mat. Apl. Comput.*, 10(3):229–242, 1991.
- [33] D. W. Kelly. The self-equilibration of residuals and complementary a-posteriori error estimation in the finite element method. *Int. J. Num. Meth. Eng.*, 20:1491–1506, 1984.
- [34] R. Kornhuber and R. Roitzsch. On adaptive grid refinement in the presence of internal and boundary layers. *IMPACT of Computing in Sci. and Engrg.*, 2:40–72, 1990.
- [35] G. Kunert. Error estimation for anisotropic tetrahedral and triangular finite element meshes. Preprint SFB393/97\_16, TU Chemnitz, 1997.
- [36] M. Křížek, P. Neittaanmäki, and R. Stenberg, editors. *Finite element methods. Superconvergence, post-processing, and a posteriori estimates*. Number 196 in Lecture notes in pure and applied mathematics. Dekker, New York, 1998. 1st conference at Jyväskylä, Finland.
- [37] P. Ladevèze and D. Leguillon. Error estimate procedure in the finite element method and applications. *SIAM J. Numer. Anal.*, 20(3):485–509, 1983.
- [38] G. Lube and D. Weiss. Stabilized finite element methods for singularly perturbed parabolic problems. *Appl. Numer. Math.*, 17(4):431–459, 1995.
- [39] J. J. H. Miller, E. O’Riordan, and G. Shishkin. *Fitted numerical methods for singularly perturbed problems. Error Estimates in the maximum norm for linear problems in one and two dimensions*. World Scientific Publications, Singapore, 1996.
- [40] J. A. Nitsche. Ein Kriterium für die Quasi-Optimalität des Ritzschen Verfahrens. *Numer. Math.*, 21:138–160, 1968.
- [41] R. H. Nochetto. Pointwise a posteriori error estimators for elliptic problems on highly graded meshes. *Math. Comp.*, 64:1–22, 1995.
- [42] J. Peraire, M. Vahdati, K. Morgan, and O. C. Zienkiewicz. Adaptive remeshing for compressible flow computation. *J. Comp. Phys.*, 72:449–466, 1987.
- [43] W. Rachowicz. An anisotropic  $h$ -type mesh refinement strategy. *Comput. Methods Appl. Mech. Engrg.*, 109:169–181, 1993.
- [44] E. Rank and O. C. Zienkiewicz. A simple error estimator in the finite element method. *Comm. Appl. Num. Meth.*, 3:243–249, 1987.

- [45] W. Rick, H. Greza, and W. Koschel. FCT-solution on adapted unstructured meshes for compressible high speed flow computations. In E. H. Hirschel, editor, *Flow simulation with high-performance computers I*, volume 38 of *Notes on Num. Fluid Mechanics*, pages 334–438. Vieweg, 1993.
- [46] R. Rodriguez. Some remarks on the Zienkiewicz-Zhu estimator. *Int. J. Numer. Meth. in PDE*, 10:625–635, 1994.
- [47] H.-G. Roos. Layer-adapted grids for singular perturbation problems. Report MATH-NM-03-1997, TU Dresden, 1997.
- [48] H.-G. Roos, M. Stynes, and L. Tobiska. *Numerical methods for singularly perturbed differential equations. Convection-diffusion and flow problems*. Springer, Berlin, 1996.
- [49] L. R. Scott and S. Zhang. Finite element interpolation of non-smooth functions satisfying boundary conditions. *Math. Comp.*, 54:483–493, 1990.
- [50] K. G. Siebert. An a posteriori error estimator for anisotropic refinement. *Numer. Math.*, 73(3):373–398, 1996.
- [51] R. B. Simpson. Anisotropic mesh transformation and optimal error control. *Applied Numerical Mathematics*, 14:183–198, 1994.
- [52] T. Skalický and H.-G. Roos. Anisotropic mesh refinement for problems with internal and boundary layers. Report MATH-NM-09-1997, TU Dresden, July 1997.
- [53] T. Steidten and M. Jung. Das Multigrid-Programmsystem FEMGPM zur Lösung elliptischer und parabolischer Differentialgleichungen einschließlich mechanisch-thermisch gekoppelter Probleme (Version 06.90). Programmdokumentation, Technische Universität Karl-Marx-Stadt, Sektion Mathematik, 1990.
- [54] E. Stein and S. Ohnimus. Coupled model- and solution-adaptivity in the finite-element method. *submitted to Comput. Methods Appl. Mech. Eng.*, February 1997.
- [55] J. L. Synge. *The hypercircle in mathematical physics*. Cambridge University Press, Cambridge, 1957.
- [56] R. Verfürth. A posteriori error estimators for the Stokes equation. *Numer. Math.*, 55:309–325, 1989.
- [57] R. Verfürth. A posteriori error estimation and adaptive mesh-refinement techniques. *Journal of Computational and Applied Mathematics*, 50:67–83, 1994.
- [58] R. Verfürth. *A review of a posteriori error estimation and adaptive mesh-refinement techniques*. Wiley-Teubner, Chichester; Stuttgart, 1996.
- [59] R. Verfürth. Robust a posteriori error estimators for the singularly perturbed Helmholtz equation. *Numer. Math.*, 78:479–493, 1998.
- [60] R. Vilsmeier and D. Hänel. Computational aspects of flow simulation in three dimensional, unstructured, adaptive grids. In E. H. Hirschel, editor, *Flow simulation with high-performance computers II*, volume 52 of *Notes on Num. Fluid Mechanics*, pages 431–446. Vieweg, 1996.

- [61] L. B. Wahlbin. *Superconvergence in Galerkin finite element methods*. Number 1605 in Lecture Notes in Mathematics. Springer, Berlin, Heidelberg, 1995.
- [62] S. Wang. An a posteriori error estimate for finite element approximations of a singularly perturbed advection-diffusion problem. *Comput. Appl. Math.*, 87(2):227–242, 1997.
- [63] A. Ženíšek. *Nonlinear elliptic and evolution problems and their finite element approximations*. Academic Press, London, 1990.
- [64] G. Zhou and R. Rannacher. Mesh orientation and anisotropic refinement in the streamline diffusion method. Preprint 93-57, Universität Heidelberg, IWR, SFB 359, 1993.
- [65] O. C. Zienkiewicz and J. Wu. Automatic directional refinement in adaptive analysis of compressible flows. *Int. J. Num. Meth. Eng.*, 37:2189–2210, 1994.
- [66] O. C. Zienkiewicz and J. Z. Zhu. A simple error estimator and adaptive procedure for practical engineering analysis. *Int. J. Num. Meth. Eng.*, 24:337–357, 1987.
- [67] O. C. Zienkiewicz and J. Z. Zhu. The superconvergent patch recovery (SPR) and adaptive finite element refinement. *Comput. Methods Appl. Mech. Eng.*, 101(1-3):207–224, 1992.



# Theses

## A posteriori error estimation for anisotropic tetrahedral and triangular finite element meshes

Dipl.-Math. Gerd Kunert

Chemnitz University of Technology, Faculty of Mathematics

1. When boundary value problems are solved by the finite element method, one commonly utilizes so-called *isotropic* meshes (i.e. where the ratio of the diameters of the circumscribed and inscribed sphere is bounded for each element). Some problems, however, yield anisotropic solutions, e.g. solutions with interior or boundary layers. Then *anisotropic* meshes (i.e. meshes with stretched elements) can be advantageous to reduce the computational effort substantially.

For several reasons *a posteriori* finite element error estimators are sought, e.g. in order to measure the quality of an approximate solution, or to design adaptive solution algorithms. Such error estimators are partly well understood (and proofs are available) for *isotropic* meshes and for different problem classes (see e.g. the overview work of Verfürth 1996 or Ainsworth/Oden 1997). A stringent theoretical foundation for *anisotropic* meshes is still at the beginning. First ideas go back to Siebert 1996 who considered *rectangular* or *cuboidal* anisotropic meshes.

Our theory is devoted to *a posteriori* error estimation for anisotropic *tetrahedral* or *triangular* finite element meshes. Such meshes offer a greater geometrical flexibility. We treat the Poisson equation and a singularly perturbed reaction-diffusion equation as model problems, and derive several error estimators.

To facilitate the notation, let  $u$  and  $u_h$  be the exact and the approximate solution of the boundary value problem, respectively. Denote a finite element mesh by  $\mathcal{T}_h$ .

2. We derive interpolation error estimates for anisotropic meshes. They form an important tool for anisotropic finite element error estimation.
3. For the Poisson equation we introduce a so-called *residual error estimator*. We prove that this estimator bounds the energy norm  $\|u - u_h\|$  of the error globally from above and locally from below, respectively. Our estimator is an improvement to Siebert's one because one of his assumptions can be omitted.
4. For the Poisson equation an error estimator based on a local problem is defined. This *local problem error estimator* is proven to be equivalent to the residual error estimator. Hence the local problem error estimator bounds the energy norm  $\|u - u_h\|$  of the error globally from above and locally from below, respectively. The local problem to be solved is well conditioned.
5. We introduce an error estimator that utilizes an averaged (recovered) gradient. Such estimators are known as *Zienkiewicz-Zhu like error estimators*. For the special case of tensor product type tetrahedral meshes the equivalence of this Zienkiewicz-Zhu like error estimators with a modified residual error estimator is proven.

Two further Zienkiewicz-Zhu like error estimators are introduced and motivated. The promising results of the numerical examples justify further investigations.

6. For the Poisson equation an  $L_2$  residual error estimator is proposed. We present a proof that this estimator bounds the  $L_2$  norm  $\|u - u_h\|$  of the error globally from above and locally from below, respectively. Special anisotropic  $L_2$  bubble functions are introduced and their inverse inequalities are proven.
7. We consider the singularly perturbed reaction-diffusion equation  $-\varepsilon\Delta u + u = f$  and derive a corresponding residual error estimator. It is shown that this estimator bounds the energy norm  $\|u - u_h\|$  of the error globally from above and locally from below, respectively. These estimates are uniform in  $\varepsilon$  (i.e. the constants are independent of  $\varepsilon$ ).
8. For the aforementioned error estimators it is important whether an anisotropic mesh  $\mathcal{T}_h$  is ‘adapted’ (in some sense) to the anisotropic solution. In order to measure this adaption a so-called ‘matching function’  $m_1(\cdot, \mathcal{T}_h)$  is defined. This matching function enters the interpolation error estimates as well as the upper bound of most error estimators. Analytical examples prove (and numerical examples strongly suggest) that this matching function  $m_1$  is necessary, i.e. the corresponding inequalities without  $m_1$  do not hold.

The upper bound of most error estimates contains the factor  $m_1(u - u_h, \mathcal{T}_h)$ . This matching function depends on the (unknown) solution  $u$ , and thus we can not compute or evaluate it. Hence an approximation  $m_1^R$  of this matching function  $m_1$  is introduced. The numerical examples show that  $m_1^R$  approximates  $m_1$  quite well.

9. Our theory of error estimation does *not* require a maximum angle condition on the elements.
10. We include three types of boundary conditions, namely Dirichlet, Neumann and Robin boundary conditions.

For *isotropic* meshes Robin boundary conditions have not been considered yet in *a posteriori* error estimation, as far as we know. We filled this gap and incorporated Robin boundary conditions such that isotropic and anisotropic meshes can be treated. The error estimates are uniform with respect to the parameters of the Robin boundary condition.

11. Numerous examples are investigated. The results of these numerical experiments are in accordance with our theory. The estimators yield useful and reliable error bounds not only in an asymptotic sense but also for meshes with a moderate number of elements.
12. Our work answers several important questions of the theory of *a posteriori* error estimation on anisotropic tetrahedral (or triangular) meshes. Further fruitful fields of investigation could include
  - the development of a fully adaptive strategy,
  - different problem classes and differential equations (e.g. reaction-diffusion-convection equations),
  - other error estimators (e.g. based on the ‘superconvergent patch recovery’ of Zienkiewicz-Zhu 1992, or on the ‘complementary energy principle’, cf. Ainsworth/Oden 1997).