Theory of Adaptive Finite Element Methods: An Introduction

Ricardo H. Nochetto and Kunibert G. Siebert and Andreas Veeser

Abstract This is a survey on the theory of adaptive finite element methods (AFEM), which are fundamental in modern computational science and engineering. We present a self-contained and up-to-date discussion of AFEM for linear second order elliptic partial differential equations (PDEs) and dimension $d > 1$, with emphasis on the differences and advantages of AFEM over standard FEM. The material is organized in chapters with problems that extend and complement the theory. We start with the functional framework, inf-sup theory, and Petrov-Galerkin method, which are the basis of FEM. We next address four topics of essence in the theory of AFEM that cannot be found in one single article: mesh refinement by bisection, piecewise polynomial approximation in graded meshes, a posteriori error analysis, and convergence and optimal decay rates of AFEM. The first topic is of geometric and combinatorial nature, and describes bisection as a rather simple and efficient technique to create conforming graded meshes with optimal complexity. The second topic explores the potentials of FEM to compensate singular behavior with local resolution and so reach optimal error decay. This theory, although insightful, is insufficient to deal with PDEs since it relies on knowing the exact solution. The third topic provides the missing link, namely a posteriori error estimators, which hinge exclusively on accessible data: we restrict ourselves to the simplest residual-type estimators and present a complete discussion of upper and lower bounds, along with the concept of oscillation and its critical role. The fourth topic refers to the convergence of adaptive loops and its comparison with quasi-uniform refinement. We first show, under rather modest assumptions on the problem class and AFEM, convergence in the natural norm associated to the variational formulation. We next restrict the problem class to coercive symmetric bilinear forms, and show that AFEM is a contraction for a suitable error notion involving the induced energy norm. This property is then instrumental to prove optimal cardinality of AFEM for a class of singular functions, for which the standard FEM is suboptimal.

Ricardo H. Nochetto
Department of Mathematics and Institute of Physical Science and Technology, University of Maryland, College Park, MD 20742, e-mail: rhn@math.umd.edu. Partially supported by NSF grant DMS-0807811.

Kunibert G. Siebert
Fachbereich Mathematik, Universität Duisburg-Essen, Forsthausweg 2, D-47057 Duisburg, Germany, e-mail: kg.siebert@uni-due.de

Andreas Veeser
Dipartimento di Matematica, Università degli Studi di Milano, Via C. Saldini 50, I-20133 Milano, Italy, e-mail: andreas.veeser@unimi.it

1 Introduction

Adaptive finite element methods are a fundamental numerical instrument in science and engineering to approximate partial differential equations. In the 1980s and 1990s a great deal of effort was devoted to the design of a posteriori error estimators, following the pioneering work of Babuška. These are computable quantities, depending on the discrete solution(s) and data, that can be used to assess the approximation quality and improve it adaptively. Despite their practical success, adaptive processes have been shown to converge, and to exhibit optimal complexity, only recently and for linear elliptic PDE.

This survey presents an up-to-date discussion of adaptive finite element methods encompassing its design and basic properties, convergence, and optimality.

1.1 Classical vs Adaptive Approximation in 1d

We start with a simple motivation in 1d for the use of adaptive procedures, due to DeVore [28]. Given $\Omega = (0,1)$, a partition $\mathscr{T}_N = \{x_i\}_{n=0}^N$ of Ω

$$0 = x_0 < x_1 < \cdots < x_n < \cdots < x_N = 1$$

and a continuous function $u : \Omega \to \mathbb{R}$, we consider the problem of *interpolating* u by a *piecewise constant* function U_N over $\mathscr{T}_N$. To quantify the difference between u and U_N we resort to the *maximum norm* and study two cases depending on the regularity of u.

Case 1: W^1_∞-Regularity. Suppose that u is Lipschitz in $[0,1]$. We consider the approximation

$$U_N(x) := u(x_{n-1}) \qquad \text{for all } x_{n-1} \le x < x_n.$$

Since

$$|u(x) - U_N(x)| = |u(x) - u(x_{n-1})| = \left| \int_{x_{n-1}}^{x} u'(t)dt \right| \le h_n \|u'\|_{L^\infty(x_{n-1},x_n)}$$

we conclude that

$$\|u - U_N\|_{L^\infty(\Omega)} \le \frac{1}{N} \|u'\|_{L^\infty(\Omega)}, \tag{1}$$

provided the local mesh-size h_n is about constant (*quasi-uniform* mesh), and so proportional to N^{-1} (the reciprocal of the number of degrees of freedom). Note that the same integrability is used on both sides of (1). A natural question arises: *Is it possible to achieve the same asymptotic decay rate N^{-1} with weaker regularity demands?*

Case 2: W^1_1-Regularity. To answer this question, we suppose $\|u'\|_{L^1(\Omega)} = 1$ and consider the non-decreasing function

$$\phi(x) := \int_0^x |u'(t)|dt$$

which satisfies $\phi(0) = 0$ and $\phi(1) = 1$. Let $\mathscr{T}_N = \{x_i\}_{n=0}^N$ be the partition given by

$$\int_{x_{n-1}}^{x_n} |u'(t)|dt = \phi(x_n) - \phi(x_{n-1}) = \frac{1}{N}.$$

Then, for $x \in [x_{n-1}, x_n]$,

$$|u(x) - u(x_{n-1})| = \Big| \int_{x_{n-1}}^x u'(t)dt \Big| \leq \int_{x_{n-1}}^x |u'(t)|dt \leq \int_{x_{n-1}}^{x_n} |u'(t)|dt = \frac{1}{N},$$

whence

$$\|u - U_N\|_{L^\infty(\Omega)} \leq \frac{1}{N} \|u'\|_{L^1(\Omega)}. \tag{2}$$

We thus conclude that we could achieve the same rate of convergence N^{-1} for rougher functions with just $\|u'\|_{L^1(\Omega)} < \infty$. The following comments are in order for Case 2.

Remark 1 (Equidistribution). The optimal mesh $\mathscr{T}_N$ *equidistributes* the max-error. This mesh is graded instead of uniform but, in contrast to a uniform mesh, such a partition may not be adequate for another function with the same basic regularity as u. It is instructive to consider the singular function $u(x) = x^\gamma$ with $\gamma = 0.1$ and error tolerance 10^{-2} to quantify the above computations: if N_1 and N_2 are the number of degrees of freedom with uniform and graded partitions, we obtain $N_1/N_2 = 10^{18}$.

Remark 2 (Nonlinear Approximation). The regularity of u in (2) is measured in $W_1^1(\Omega)$ instead of $W_\infty^1(\Omega)$ and, consequently, the fractional γ regularity measured in $L^\infty(\Omega)$ increases to one full derivative when expressed in $L^1(\Omega)$. This exchange of integrability between left and right-hand side of (2), and gain of differentiability, is at the heart of the matter and the very reason why suitably graded meshes achieve optimal asymptotic error decay for singular functions. By those we mean functions which are not in the usual linear Sobolev scale, say $W_\infty^1(\Omega)$ in this example, but rather in a nonlinear scale [28]. We will get back to this issue in Chap. 5.

1.2 Outline

The function U_N may be the result of a minimization process. If we wish to minimize the norm $\|u - v\|_{L^2(\Omega)}$ within the space $\mathbb{V}_N$ of piecewise constant functions over $\mathscr{T}_N$, then it is easy to see that the solution U_N satisfies the orthogonality relation

$$U_N \in \mathbb{V}_N : \qquad \langle u - U_N, v \rangle = 0 \qquad \text{for all } v \in \mathbb{V}_N \tag{3}$$

and is given by the explicit local expression

$$U_N(x) = \frac{1}{h_n}\int_{x_{n-1}}^{x_n} u \qquad \text{for all } x_{n-1} < x < x_n.$$

The previous comments apply to this U_N as well even though U_N coincides with u at an unknown point in each interval $[x_{n-1}, x_n]$.

The latter example is closer than the former to the type of approximation issues discussed in this survey. A brief summary along with an outline of this survey follows:

PDE: The function u is not directly accessible but rather it is the solution of an elliptic PDE. Its approximation properties are intimately related to its regularity. In Chap. 2 we review briefly Sobolev spaces and the variational formulation of elliptic PDE, a present a full discussion of the inf-sup theory. We show the connection between approximability and regularity in Chap. 5, when we assess constructive approximation and use this later in Chap. 9 to derive rates of convergence.

FEM: To approximate u we need a numerical method which is sufficiently flexible to handle both geometry and accuracy (local mesh refinement); the method of choice for elliptic PDEs is the finite element method. We present its basic theory in Chap. 3, with emphasis on piecewise linear elements. We discuss the refinement of simplicial meshes in any dimension by bisection in Chap. 4, and address its complexity. This allows us to shed light on the geometric aspects of FEM that make them so flexible and useful in practice. The complexity analysis of bisection turns out to be crucial to construct optimal approximations in graded meshes in Chap. 5 and to derive convergence rates in Chap. 9 for AFEM.

Approximation: We briefly recall polynomial interpolation theory in Chap. 5 as well as the principle of error equidistribution. The latter is a concept that leads to optimal graded meshes and suggests that FEM might be able to approximate singular functions with optimal rate. We conclude Chap. 5 with the construction of optimal meshes via bisection for functions in a certain regularity class relevant to elliptic PDE. We emphasize the energy norm.

A Posteriori Error Estimation: To extract the local errors incurred by FEM, and thus be able to equidistribute them, we present residual-type a posteriori error estimators in Chap. 6. These are computable quantities in terms of the discrete solution and data which encode the correct information about the error distribution. They are the simplest but not the most accurate ones. Therefore, we also present alternative estimators, which are equivalent to the residual estimators. The discussion of Chap. 6 includes the appearence of an oscillation term and a proof that it cannot be avoided for the estimator to be practical. We show both upper and lower bounds between the energy error and the residual estimator. The former is essential for convergence and the latter for optimality.

Adaptivity: This refers to the use and study of loops to the form

$$\mathsf{SOLVE} \longrightarrow \mathsf{ESTIMATE} \longrightarrow \mathsf{MARK} \longrightarrow \mathsf{REFINE} \qquad (4)$$

to iteratively improve the approximation of the solution of a PDE while keeping an optimal distribution of computational resources (degrees of freedom). The design of each module, along with some key properties, is discussed in Chap. 7 and 8. We emphasize the standard AFEM employed in practice which employs the estimator exclusively to make refinement decisions and never uses coarsening.

Convergence: This issue has been largely open until recently. In Chap. 7 we present a basic convergence theory for most linear elliptic PDEs, including saddle point problems, under rather modest assumptions and valid for all existing marking strategies. The final result is rather general but does not, and cannot, provide a convergence rate.

Optimality: We restrict ourselves to a model problem, which is symmetric and coercive, to investigate the convergence rate of AFEM. In Chap. 8 we derive a contraction property of AFEM for the so-called quasi-error, which is a scaled sum of the energy error and the estimator. In Chap. 9 we prove that AFEM converges with optimal rate as dictated by approximation theory even though the adaptive loop (4) does not use any regularity information but just the estimator. This analysis leads to approximation classes adequate for FEM, and so to the geometric restrictions caused by conforming grids, which are not the usual ones in nonlinear approximation theory.

2 Linear Boundary Value Problems

In this section we examine the variational formulation of elliptic partial differential equations (PDE). We start with a brief review of Sobolev spaces and their properties and continue with several boundary value problems with main emphasis on a model problem that plays a relevant role in the subsequent analysis. Then we present the so-called inf-sup theory that characterizes existence and uniqueness of variational problems, and conclude by reviewing the applications in light of the inf-sup theory.

2.1 Sobolev Spaces

The variational formulation of elliptic PDEs is based on Sobolev spaces. Moreover, approximability and regularity of functions are intimately related concepts. Therefore we briefly review definitions, basic concepts and properties of L^p-based Sobolev spaces for $1 \le p \le \infty$ and dimension $d \ge 1$. For convenience we restrict ourselves to bounded domains $\Omega \subset \mathbb{R}^d$ with Lipschitz boundary.

Definition 1 (Sobolev Space). Given $k \in \mathbb{N}$ and $1 \le p \le \infty$, we define

$$W_p^k(\Omega) := \{v\colon \Omega \to \mathbb{R} \mid D^\alpha v \in L^p(\Omega) \text{ for all } |\alpha| \le k\}$$

where $D^\alpha v = \partial_{x_1}^{\alpha_1} \cdots \partial_{x_d}^{\alpha_d} v$ stands for the weak derivative of order α. The corresponding norm and seminorm are for $1 \le p < \infty$

$$\|v\|_{W_p^k(\Omega)} := \Big(\sum_{|\alpha| \le k} \|D^\alpha v\|_{L^p(\Omega)}^p \Big)^{1/p}, \quad |v|_{W_p^k(\Omega)} := \Big(\sum_{|\alpha| = k} \|D^\alpha v\|_{L^p(\Omega)}^p \Big)^{1/p},$$

and for $p = \infty$

$$\|v\|_{W_p^\infty(\Omega)} := \sup_{|\alpha| \le k} \|D^\alpha v\|_{L^\infty(\Omega)}, \quad |v|_{W_p^\infty(\Omega)} := \sup_{|\alpha| = k} \|D^\alpha v\|_{L^\infty(\Omega)}.$$

For $p = 2$ the spaces $W_2^k(\Omega)$ are Hilbert spaces and we denote them by $H^k(\Omega) = W_2^k(\Omega)$. The scalar product inducing the norm $\|\cdot\|_{H^k(\Omega)} = \|\cdot\|_{W_2^k(\Omega)}$ is given by

$$\langle u, v \rangle_{H^k(\Omega)} = \sum_{|\alpha| \le k} \int_\Omega D^\alpha u D^\alpha v \qquad \text{for all } u, v \in H^k(\Omega).$$

We let $H_0^k(\Omega)$ be the completion of $C_0^\infty(\Omega)$ within $H^k(\Omega)$. The space $H_0^k(\Omega)$ is a strict subspace $H^k(\Omega)$ because $1 \in H^k(\Omega) \setminus H_0^k(\Omega)$.

There is a natural scaling of the seminorm in $W_p^k(\Omega)$. Consider for $h > 0$ the change of variables $\widehat{x} = x/h$ for all $x \in \Omega$, which transforms the domain Ω into $\widehat{\Omega}$ and functions v defined over Ω into functions $\widehat{v}$ defined over $\widehat{\Omega}$. Then

$$|\widehat{v}|_{W_p^k(\widehat{\Omega})} = h^{k-d/p}\,|v|_{W_p^k(\Omega)}.$$

This motivates the following definition, which turns out to be instrumental.

Definition 2 (Sobolev Number). The Sobolev number of $W_p^k(\Omega)$ is defined by

$$\mathrm{sob}(W_p^k) := k - d/p. \tag{5}$$

2.1.1 Properties of Sobolev Spaces

We summarize now, but not prove, several important properties of Sobolev spaces which play a key role later. We refer to [35, 38, 39] for details.

Embedding Theorem. Let $m > k \geq 0$ and assume $\mathrm{sob}(W_p^m) > \mathrm{sob}(W_q^k)$. Then the embedding

$$W_p^m(\Omega) \hookrightarrow W_q^k(\Omega)$$

is compact.

The assumption on the Sobolev number cannot be relaxed. To see this, consider Ω to be the unit ball of $\mathbb{R}^d$ for $d \geq 2$ and set $v(x) = \log\log \frac{|x|}{2}$ for $x \in \Omega \setminus \{0\}$. Then there holds $v \in W_d^1(\Omega)$ and $v \notin L^\infty(\Omega)$, but

$$\mathrm{sob}(W_d^1) = 1 - d/d = 0 = 0 - d/\infty = \mathrm{sob}(L^\infty).$$

Therefore, equality cannot be expected in the embedding theorem.

Density. The space $C^\infty(\overline{\Omega})$ is dense in $W_p^k(\Omega)$, i. e.,

$$W_p^k(\Omega) = \overline{C^\infty(\overline{\Omega})}^{\|\cdot\|_{\mathbb{V}}}.$$

Poincaré Inequality. The following inequality holds

$$\Big\| v - |\Omega|^{-1} \int_\Omega v \Big\|_{L^2(\Omega)} \leq C(\Omega)\|\nabla v\|_{L^2(\Omega)} \qquad \text{for all } v \in W_2^1(\Omega) \tag{6}$$

with a constant $C(\Omega)$ depending on the shape of Ω. The best constant within the class of convex domains is

$$C(\Omega) = \frac{1}{\pi}\,\mathrm{diam}(\Omega);$$

see [60, 11].

Poincaré-Friedrichs Inequality. There is a constant $C_d > 0$ depending only on the dimension such that [38, p. 158]

$$\|v\|_{L^2(\Omega)} \leq C_d |\Omega|^{1/d} \|\nabla v\|_{L^2(\Omega)} \qquad \text{for all } v \in H_0^1(\Omega). \tag{7}$$

Trace Theorem. Functions in $H^1(\Omega)$ have 'boundary values' in $L^2(\Omega)$, called *trace*, in that there exists a unique linear operator $T: H^1(\Omega) \to L^2(\partial\Omega)$ such that

$$\begin{aligned} \|Tv\|_{L^2(\partial\Omega)} &\le c(\Omega)\|v\|_{H^1(\Omega)} && \text{for all } v \in H^1(\Omega), \\ Tv &= v && \text{for all } v \in C^0(\overline{\Omega}) \cap H^1(\Omega). \end{aligned}$$

Since $Tv = v$ for continuous functions we write v for Tv. For a simplex we give an explicit construction of the constant $c(\Omega)$ in Sect. 6.2. The image of T is a strict subspace of $L^2(\partial\Omega)$, the so-called $H^{1/2}(\partial\Omega)$. The definition of $H^1_0(\Omega)$ can be reconciled with that of traces because

$$H^1_0(\Omega) = \{v \in H^1(\Omega) \mid v = 0 \text{ on } \partial\Omega\}.$$

The operator T is also well defined on $W^1_p(\Omega)$ for $1 \le p \le \infty$.

Green's Formula. Given functions $v, w \in H^1(\Omega)$, the following fundamental Green's formula

$$\int_\Omega \partial_i w\, v = -\int_\Omega w\, \partial_i v + \int_{\partial\Omega} w\, v\, n_i \tag{8}$$

holds for any $i = 1, \dots, d$, where $\boldsymbol{n}(x) = [n_1(x), \dots, n_d(x)]^T$ is the outer unit normal of $\partial\Omega$ at x. Equivalently, if $v \in H^1(\Omega)$ and $\boldsymbol{w} \in H^1(\Omega; \mathbb{R}^d)$ then there holds

$$\int_\Omega \operatorname{div} \boldsymbol{w}\, v = -\int_\Omega \boldsymbol{w} \cdot \nabla v + \int_{\partial\Omega} v\, \boldsymbol{w} \cdot \boldsymbol{n}. \tag{9}$$

Green's formula is a direct consequence of Gauß' Divergence Theorem

$$\int_\Omega \operatorname{div} \boldsymbol{w} = \int_{\partial\Omega} \boldsymbol{w} \cdot \boldsymbol{n} \qquad \text{for all } \boldsymbol{w} \in W^1_1(\Omega; \mathbb{R}^d).$$

2.2 Variational Formulation

We consider elliptic PDEs that can be formulated as the following variational problem: Let $(\mathbb{V}, \langle \cdot, \cdot \rangle_{\mathbb{V}})$ be an Hilbert space with induced norm $\|\cdot\|_{\mathbb{V}}$ and denote by $\mathbb{V}^*$ its dual space equipped with the norm

$$\|f\|_{\mathbb{V}^*} = \sup_{v \in \mathbb{V}} \frac{\langle f, v \rangle}{\|v\|_{\mathbb{V}}} \qquad \text{for all } f \in \mathbb{V}^*.$$

Consider a continuous bilinear form $\mathscr{B}: \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ and $f \in \mathbb{V}^*$. Then we seek a solution $u \in \mathbb{V}$ of

$$u \in \mathbb{V}: \qquad \mathscr{B}[u, v] = \langle f, v \rangle \qquad \text{for all } v \in \mathbb{V}. \tag{10}$$

We first look at several examples that are relevant for the rest of the presentation.

2.2.1 Model Problem

The model problem of this survey is the following 2nd order elliptic PDE

$$-\operatorname{div}(\boldsymbol{A}(x)\nabla u) = f \qquad \text{in } \Omega, \tag{11a}$$
$$u = 0 \qquad \text{on } \partial\Omega, \tag{11b}$$

where $f \in L^2(\Omega)$ and $\boldsymbol{A} \in L^\infty(\Omega;\mathbb{R}^{d\times d})$ is uniformly symmetric positive definite (SPD) over Ω, i. e., there exists constants $0 < \alpha_1 \le \alpha_2$ such that

$$\alpha_1|\boldsymbol{\xi}|^2 \le \boldsymbol{\xi}^T\boldsymbol{A}(x)\boldsymbol{\xi} \le \alpha_2|\boldsymbol{\xi}|^2 \qquad \text{for all } x \in \Omega, \boldsymbol{\xi} \in \mathbb{R}^d. \tag{12}$$

For the variational formulation of (11) we let $\mathbb{V} = H^1_0(\Omega)$ and denote its dual by $\mathbb{V}^* = H^{-1}(\Omega)$. Since $H^1_0(\Omega)$ is the subspace of $H^1(\Omega)$ of functions with vanishing trace, asking for $u \in \mathbb{V}$ accounts for the homogeneous Dirichlet boundary values in (11b).

We next multiply (11a) with a test function $v \in H^1_0(\Omega)$, integrate over Ω and use Green's formula (9), provided $\boldsymbol{w} = -\boldsymbol{A}\nabla u \in H^1(\Omega;\mathbb{R}^d)$, to derive the variational formulation

$$u \in \mathbb{V}: \qquad \int_\Omega \nabla v \cdot \boldsymbol{A}(x)\nabla u = \int_\Omega f v \qquad \text{for all } v \in \mathbb{V}, \tag{13}$$

because the boundary term is zero thanks to $v = 0$ on $\partial\Omega$. However, problem (13) makes sense with much less regularity of the flux $\boldsymbol{w}$. Setting

$$\mathscr{B}[w, v] := \int_\Omega \nabla v \cdot \boldsymbol{A}(x)\nabla w \qquad \text{for all } v, w \in H^1_0(\Omega),$$
$$\langle f, v\rangle := \int_\Omega f v \qquad \text{for all } v \in H^1_0(\Omega),$$

(13) formally reads as (10). In Sect. 2.5.1 we analyze further $\mathscr{B}$ and $\langle f, \cdot\rangle$.

2.2.2 Other Boundary Value Problems

We next introduce several elliptic boundary value problems that also fit within the present theory.

General 2nd Order Elliptic Operator. Let $\boldsymbol{A} \in L^\infty(\Omega;\mathbb{R}^{d\times d})$ be uniformly SPD as above, $\boldsymbol{b} \in L^\infty(\Omega;\mathbb{R}^d)$, $c \in L^\infty(\Omega)$, and $f \in L^2(\Omega)$. We now consider the general 2nd order elliptic equation

$$-\operatorname{div}(\boldsymbol{A}(x)\nabla u) + \boldsymbol{b}(x)\cdot\nabla u + c(x)u = f \qquad \text{in } \Omega,$$
$$u = 0 \qquad \text{on } \partial\Omega.$$

The variational formulation utilizes $\mathbb{V} = H^1_0(\Omega)$, as in Sect. 2.2.1. We again multiply the PDE with a test function $v \in H^1_0(\Omega)$, integrate over Ω, and use Green's formula (9) provided $\boldsymbol{A}(x)\nabla u \in H^1(\Omega;\mathbb{R}^d)$. This gives the bilinear form

$$\mathscr{B}[w,v] := \int_\Omega \nabla v \cdot \boldsymbol{A}(x)\nabla w + v\boldsymbol{b}\cdot\nabla w + c\,v\,w \qquad \text{for all } v,w \in H^1_0(\Omega)$$

and $\langle f, v\rangle = \int_\Omega f\,v$ in (10). We examine $\mathscr{B}$ further in Sect. 2.5.2.

The Biharmonic Equation. The vertical displacement u of the mid-surface $\Omega \subset \mathbb{R}^2$ of a clamped plate under a vertical acting force $f \in L^2(\Omega)$ can be modeled by the *biharmonic equation*

$$\Delta^2 u = f \qquad \text{in } \Omega, \tag{14a}$$
$$u = \partial_{\boldsymbol{n}} u = 0 \qquad \text{on } \partial\Omega, \tag{14b}$$

where $\partial_{\boldsymbol{n}} u = \nabla u \cdot \boldsymbol{n}$ is the normal derivative of u on $\partial\Omega$.

For the variational formulation we let $\mathbb{V} = H^2_0(\Omega)$, and note that

$$H^2_0(\Omega) = \left\{ v \in H^2(\Omega) \mid v = \partial_{\boldsymbol{n}} v = 0 \text{ on } \partial\Omega \right\}$$

also accounts for the boundary values (14b). Here, we use Green's formula (9) twice to deduce for all $u \in H^4(\Omega)$ and $v \in H^2(\Omega)$

$$\int_\Omega \Delta^2 u\,v = \int_\Omega \Delta u\,\Delta v + \int_{\partial\Omega} \partial_{\boldsymbol{n}}\Delta u\; v + \int_{\partial\Omega} \Delta u\;\partial_{\boldsymbol{n}} v.$$

Multiplying (14a) with $v \in H^2_0(\Omega)$, integrating over Ω, and using the above formula (without boundary terms), we derive the bilinear form of (10)

$$\mathscr{B}[w,v] := \int_\Omega \Delta v \Delta w \qquad \text{for all } v,w \in \mathbb{V},$$

and set $\langle f, v\rangle := \int_\Omega f\,v$ for $v \in \mathbb{V}$.

The 3d Eddy Current Equations. Given constant material parameters $\mu, \kappa > 0$ and $\boldsymbol{f} \in L^2(\Omega;\mathbb{R}^3)$ we next consider the *3d eddy current equations*

$$\operatorname{curl}(\mu \operatorname{curl} \boldsymbol{u}) + \kappa \boldsymbol{u} = \boldsymbol{f} \qquad \text{in } \Omega, \tag{15a}$$
$$u \wedge \boldsymbol{n} = 0 \qquad \text{on } \partial\Omega, \tag{15b}$$

with the curl operator

$$\operatorname{curl} \boldsymbol{v} := \nabla \wedge \boldsymbol{v} = \left[\frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3}, \frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1}, \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2} \right]$$

and the vector product $\wedge$ in $\mathbb{R}^3$.

The variational formulation is based on the Sobolev space

$$H(\operatorname{curl};\Omega) := \left\{ \boldsymbol{v} \in L^2(\Omega;\mathbb{R}^3) \mid \operatorname{curl} \boldsymbol{v} \in L^2(\Omega;\mathbb{R}^3) \right\}$$

equipped with the norm $\|\boldsymbol{v}\|^2_{H(\operatorname{curl};\Omega)} := \|\boldsymbol{v}\|^2_{L^2(\Omega;\mathbb{R}^3)} + \|\operatorname{curl}\boldsymbol{v}\|^2_{L^2(\Omega;\mathbb{R}^3)}$. This is a Hilbert space and is larger than $H^1(\Omega;\mathbb{R}^3)$. The weak formulation of (15) utilizes the subspace of functions with vanishing tangential trace on $\partial\Omega$

$$\mathbb{V} := H_0(\operatorname{curl};\Omega) = \left\{ \boldsymbol{v} \in H(\operatorname{curl};\Omega) \mid \boldsymbol{v} \wedge \boldsymbol{n} = 0 \text{ on } \partial\Omega \right\} = \overline{C_0^\infty(\Omega;\mathbb{R}^3)}^{\|\cdot\|_{H(\operatorname{curl};\Omega)}},$$

which thereby incorporates the boundary values of (15b). This space is a closed and proper subspace of $H(\operatorname{curl};\Omega)$.

From Green's formula (8) with proper choices of $\boldsymbol{v}$ and $\boldsymbol{w}$ it is easy to derive the following formula for all $\boldsymbol{v},\boldsymbol{w} \in H(\operatorname{curl};\Omega)$

$$\int_\Omega \operatorname{curl}\boldsymbol{w}\cdot\boldsymbol{v} = \int_\Omega \boldsymbol{w}\cdot\operatorname{curl}\boldsymbol{v} + \int_{\partial\Omega} \boldsymbol{w}\cdot(\boldsymbol{v}\wedge\boldsymbol{n}).$$

Multiplying (15a) with a test function $\boldsymbol{v} \in H_0(\operatorname{curl};\Omega)$, integrating over Ω and using the above formula with $\boldsymbol{w} = \mu\operatorname{curl}\boldsymbol{u} \in H(\operatorname{curl};\Omega)$, we end up with the bilinear form and right hand side of (10)

$$\begin{aligned} \mathscr{B}[\boldsymbol{w},\boldsymbol{v}] &:= \int_\Omega \mu\operatorname{curl}\boldsymbol{v}\cdot\operatorname{curl}\boldsymbol{w} + \kappa\boldsymbol{v}\cdot\boldsymbol{w} && \text{for all } \boldsymbol{v},\boldsymbol{w}\in\mathbb{V}, \\ \langle \boldsymbol{f},\boldsymbol{v}\rangle &:= \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v} && \text{for all } \boldsymbol{v}\in\mathbb{V}. \end{aligned}$$

The Stokes System. Given an external force $\boldsymbol{f} \in L^2(\Omega;\mathbb{R}^d)$, let the velocity-pressure pair $(\boldsymbol{u},p)$ satisfy the momemtum and incompressibility equations with no-slip boundary condition:

$$\begin{aligned} -\Delta\boldsymbol{u} + \nabla p &= \boldsymbol{f} && \text{in } \Omega, \\ \operatorname{div}\boldsymbol{u} &= 0 && \text{in } \Omega, \\ \boldsymbol{u} &= 0 && \text{on } \partial\Omega. \end{aligned}$$

For the variational formulation we consider two Hilbert spaces $\mathbb{V} = H^1_0(\Omega;\mathbb{R}^d)$ and $\mathbb{Q} = L^2_0(\Omega)$, where $L^2_0(\Omega)$ is the space of L^2 functions with zero mean value. The space $H^1_0(\Omega;\mathbb{R}^d)$ takes care of the no-slip boundary values of the velocity. Proceeding as in Sect. 2.2.1, this time using component-wise integration by parts for $\int_\Omega v_i\Delta w_i$ and assuming $\boldsymbol{w} \in H^2(\Omega;\mathbb{R}^d)$, we obtain the bilinear form $a\colon \mathbb{V}\times\mathbb{V}\to\mathbb{R}$

$$a[\boldsymbol{w},\boldsymbol{v}] := \int_\Omega \nabla\boldsymbol{v} : \nabla\boldsymbol{w} = \sum_{i=1}^d \int_\Omega \nabla v_i\cdot\nabla w_i \qquad \text{for all } \boldsymbol{v},\boldsymbol{w}\in\mathbb{V}.$$

Likewise, integration by parts of $\int_\Omega \boldsymbol{v}\nabla q$ yields the bilinear form $b\colon \mathbb{Q}\times\mathbb{V}\to\mathbb{R}$

$$b[q,\boldsymbol{v}] := -\int_\Omega q \operatorname{div} \boldsymbol{v} \qquad \text{for all } q \in \mathbb{Q},\ \boldsymbol{v} \in \mathbb{V}.$$

The variational formulation then reads: find $(\boldsymbol{u}, p) \in \mathbb{V} \times \mathbb{Q}$ such that

$$\begin{aligned} a[\boldsymbol{u},\boldsymbol{v}] + b[p,\boldsymbol{v}] &= \langle \boldsymbol{f}, \boldsymbol{v}\rangle \qquad && \text{for all } \boldsymbol{v} \in \mathbb{V}, \\ b[q,\boldsymbol{u}] &= 0 && \text{for all } q \in \mathbb{Q}. \end{aligned}$$

We will see in Sect. 2.4.2 how this problem can be formulated in the form (10).

2.3 The Inf-Sup Theory

In this subsection we present a functional analytic theory, the so-called inf-sup theory, that characterizes existence, uniqueness, and continuous dependence on data of the variational problem (10).

Throughout this section we let $(\mathbb{V}, \langle\cdot,\cdot\rangle_\mathbb{V})$ and $(\mathbb{W}, \langle\cdot,\cdot\rangle_\mathbb{W})$ be a pair of Hilbert spaces with induced norms $\|\cdot\|_\mathbb{V}$ and $\|\cdot\|_\mathbb{W}$. We denote by $\mathbb{V}^*$ and $\mathbb{W}^*$ their respective dual spaces equipped with norms

$$\|f\|_{\mathbb{V}^*} = \sup_{v\in\mathbb{V}} \frac{\langle f, v\rangle}{\|v\|_\mathbb{V}} \qquad \text{and} \qquad \|g\|_{\mathbb{W}^*} = \sup_{v\in\mathbb{W}} \frac{\langle g, v\rangle}{\|v\|_\mathbb{W}}.$$

We write $L(\mathbb{V};\mathbb{W})$ for the space of all linear and continuous operators from $\mathbb{V}$ into $\mathbb{W}$ with operator norm

$$\|B\|_{L(\mathbb{V};\mathbb{W})} = \sup_{v\in\mathbb{V}} \frac{\|Bv\|_\mathbb{W}}{\|v\|_\mathbb{V}}.$$

The following result relates a continuous bilinear form $\mathscr{B}\colon \mathbb{V}\times\mathbb{W} \to \mathbb{R}$ with an operator $B \in L(\mathbb{V};\mathbb{W})$.

Theorem 1 (Banach-Nečas). *Let $\mathscr{B} : \mathbb{V}\times\mathbb{W} \to \mathbb{R}$ be a continuous bilinear form with norm*

$$\|\mathscr{B}\| := \sup_{v\in\mathbb{V}} \sup_{w\in\mathbb{W}} \frac{\mathscr{B}[v,w]}{\|v\|_\mathbb{V}\|w\|_\mathbb{W}}. \tag{16}$$

Then there exists a unique linear operator $B \in L(\mathbb{V},\mathbb{W})$ such that

$$\langle Bv, w\rangle_\mathbb{W} = \mathscr{B}[v,w] \qquad \textit{for all } v \in \mathbb{V},\ w \in \mathbb{W}$$

with operator norm

$$\|B\|_{L(\mathbb{V};\mathbb{W})} = \|\mathscr{B}\|.$$

Moreover, the bilinear form $\mathscr{B}$ satisfies

$$\textit{there exists } \alpha > 0 \ \textit{ such that } \ \alpha \|v\|_{\mathbb{V}} \leq \sup_{w \in \mathbb{W}} \frac{\mathscr{B}[v,w]}{\|w\|_{\mathbb{W}}} \quad \textit{for all } v \in \mathbb{V}, \tag{17a}$$

$$\textit{for every } \ 0 \neq w \in \mathbb{W} \ \textit{ there exists } \ v \in \mathbb{V} \ \textit{ such that } \ \mathscr{B}[v,w] \neq 0, \tag{17b}$$

if and only if $B : \mathbb{V} \to \mathbb{W}$ *is an isomorphism with*

$$\|B^{-1}\|_{L(\mathbb{W},\mathbb{V})} \leq \alpha^{-1}. \tag{18}$$

Proof. 1 *Existence of B.* For fixed $v \in \mathbb{V}$, the mapping $\mathscr{B}[v,\cdot]$ belongs to $\mathbb{W}^*$ by linearity of $\mathscr{B}$ in the second component and continuity of $\mathscr{B}$. Applying the Riesz Representation Theorem (see for instance [16, (2.4.2) Theorem], [38, Theorem 5.7]), we deduce the existence of an element $Bv \in \mathbb{W}$ such that

$$\langle Bv, w\rangle_{\mathbb{W}} = \mathscr{B}[v,w] \quad \text{for all } w \in \mathbb{W}.$$

Linearity of $\mathscr{B}$ in the first argument and continuity of $\mathscr{B}$ imply $B \in L(\mathbb{V};\mathbb{W})$. In view of (16), we get

$$\|B\|_{L(\mathbb{V};\mathbb{W})} = \sup_{v\in\mathbb{V}} \frac{\|Bv\|_{\mathbb{W}}}{\|v\|_{\mathbb{V}}} = \sup_{v\in\mathbb{V}} \sup_{w\in\mathbb{W}} \frac{\langle Bv, w\rangle}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}} = \sup_{v\in\mathbb{V}} \sup_{w\in\mathbb{W}} \frac{\mathscr{B}[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}} = \|\mathscr{B}\|.$$

2 *Closed Range of B.* The inf-sup condition (17a) implies

$$\alpha\,\|v\|_{\mathbb{V}} \leq \sup_{w\in\mathbb{W}} \frac{\langle Bv, w\rangle}{\|w\|_{\mathbb{W}}} = \|Bv\|_{\mathbb{W}} \quad \text{for all } v \in \mathbb{V}, \tag{19}$$

whence B is *injective*. To prove that the range $B(\mathbb{V})$ of B is closed in $\mathbb{W}$, we let $w_k = Bv_k$ be a sequence such that $w_k \to w \in \mathbb{W}$ as $k \to \infty$. We need to show that $w \in B(\mathbb{V})$. Invoking (19), we have

$$\alpha\,\|v_k - v_j\|_{\mathbb{V}} \leq \|B(v_k - v_j)\|_{\mathbb{W}} = \|w_k - w_j\|_{\mathbb{W}} \to 0$$

as $k, j \to \infty$. Thus $\{v_k\}_{k=0}^{\infty}$ is a Cauchy sequence in $\mathbb{V}$ and so it converges $v_k \to v \in \mathbb{V}$ as $k \to \infty$. Continuity of B yields

$$Bv = \lim_{k\to\infty} Bv_k = w \in B(\mathbb{V}),$$

which shows that $B(\mathbb{V})$ is closed.

3 *Surjectivity of B.* We argue by contradiction, i. e., assume $B(\mathbb{V}) \neq \mathbb{W}$. Since $B(\mathbb{V})$ is closed we can decompose $\mathbb{W} = B(\mathbb{V}) \oplus B(\mathbb{V})^{\perp}$, where $B(\mathbb{V})^{\perp}$ is the orthogonal complement of $B(\mathbb{V})$ in $\mathbb{W}$ (see for instance [16, (2.3.5) Proposition], [38, Theorem 5.6]). By assumption $B(\mathbb{V})^{\perp}$ is non-trivial, i. e., there exists $0 \neq w_0 \in B(\mathbb{V})^{\perp}$. This is equivalent to

$$w_0 \neq 0 \quad \text{and} \quad \langle w, w_0\rangle = 0 \quad \text{for all } w \in B(\mathbb{V}),$$

or

$$w_0 \neq 0 \quad \text{and} \quad 0 = \langle Bv, w_0 \rangle = \mathscr{B}[v, w_0] \quad \text{for all } v \in \mathbb{V}.$$

This in turn contradicts (17b) and shows that $B(\mathbb{V}) = \mathbb{W}$. Therefore, we conclude that B is an isomorphism from $\mathbb{V}$ onto $\mathbb{W}$.

[4] *Property* (18). We rewrite (19) as follows:

$$\alpha \, \|B^{-1} w\|_{\mathbb{V}} \leq \|w\|_{\mathbb{W}} \quad \text{for all } w \in \mathbb{W},$$

which is (18) in disguise.

[5] *Property* (18) *implies* (17a) *and* (17b). Compute

$$\begin{aligned} \inf_{v \in \mathbb{V}} \sup_{w \in \mathbb{W}} \frac{\mathscr{B}[v, w]}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}} &= \inf_{v \in \mathbb{V}} \sup_{w \in \mathbb{W}} \frac{\langle Bv, w \rangle}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}} = \inf_{v \in \mathbb{V}} \frac{\|Bv\|_{\mathbb{W}}}{\|v\|_{\mathbb{V}}} \\ &= \inf_{w \in \mathbb{W}} \frac{\|w\|_{\mathbb{W}}}{\|B^{-1} w\|_{\mathbb{V}}} = \frac{1}{\sup_{w \in \mathbb{W}} \frac{\|B^{-1} w\|_{\mathbb{V}}}{\|w\|_{\mathbb{W}}}} = \frac{1}{\|B^{-1}\|} \geq \alpha \end{aligned}$$

which shows (17a). Property (17b) is a consequence of B being an isomorphism: there exists $0 \neq v \in \mathbb{V}$ such that $Bv = w$ and

$$\mathscr{B}[v, w] = \langle Bv, w \rangle = \|w\|_{\mathbb{W}}^2 \neq 0.$$

This concludes the theorem. □

We are now in the position to characterize properties of the bilinear form $\mathscr{B}$ in (10) that imply that the variational problem (10) is well-posed. This result from 1962 is due to Nečas [56, Theorem 3.3].

Theorem 2 (Nečas Theorem). *Let $\mathscr{B}\colon \mathbb{V} \times \mathbb{W} \to \mathbb{R}$ be a continuous bilinear form. Then the variational problem*

$$u \in \mathbb{V}: \qquad \mathscr{B}[u, v] = \langle f, v \rangle \qquad \textit{for all } v \in \mathbb{W}, \tag{20}$$

admits a unique solution $u \in \mathbb{V}$ for all $f \in \mathbb{W}^*$, which depends continuously on f, if and only if the bilinear form $\mathscr{B}$ satisfies one of the equivalent inf-sup conditions*:*

(1) *There exists $\alpha > 0$ such that*

$$\sup_{w \in \mathbb{W}} \frac{\mathscr{B}[v, w]}{\|w\|_{\mathbb{W}}} \geq \alpha \|v\|_{\mathbb{V}} \quad \textit{for some } \ \alpha > 0; \tag{21a}$$

$$\textit{for every } 0 \neq w \in \mathbb{W} \textit{ there exists } v \in \mathbb{V} \textit{ such that } \mathscr{B}[v, w] \neq 0. \tag{21b}$$

(2) *There holds*

$$\inf_{v \in \mathbb{V}} \sup_{w \in \mathbb{W}} \frac{\mathscr{B}[v, w]}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}} > 0, \qquad \inf_{w \in \mathbb{W}} \sup_{v \in \mathbb{V}} \frac{\mathscr{B}[v, w]}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}} > 0. \tag{22}$$

(3) *There exists $\alpha > 0$ such that*

$$\inf_{v\in\mathbb{V}}\sup_{w\in\mathbb{W}}\frac{\mathscr{B}[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}}=\inf_{w\in\mathbb{W}}\sup_{v\in\mathbb{V}}\frac{\mathscr{B}[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}}=\alpha. \tag{23}$$

In addition, the solution u of (20) *satisfies the stability estimate*

$$\|u\|_{\mathbb{V}}\le\alpha^{-1}\|f\|_{\mathbb{W}^*}. \tag{24}$$

Proof. 1 Denote by $J\colon\mathbb{W}\to\mathbb{W}^*$ the isometric Riesz isomorphism between $\mathbb{W}$ and $\mathbb{W}^*$; see [16, (2.4.2) Theorem], [38, Theorem 5.7]. Let $B\in L(\mathbb{V};\mathbb{W})$ be the linear operator corresponding to $\mathscr{B}$ introduced in Theorem 1. Then (20) is equivalent to

$$u\in\mathbb{V}:\qquad Bu=J^{-1}f\qquad\text{in }\mathbb{W}.$$

Assume that (21) is satisfied. Then, according to Theorem 1, the operator B is invertible. For any $f\in\mathbb{W}^*$ the unique solution $u\in\mathbb{V}$ is given by $u=B^{-1}J^{-1}f$ and u depends continuously on f with

$$\|u\|_{\mathbb{V}}\le\|B^{-1}\|_{L(\mathbb{W};\mathbb{V})}\|J^{-1}f\|_{\mathbb{W}}=\|B^{-1}\|_{L(\mathbb{W};\mathbb{V})}\|f\|_{\mathbb{W}^*}\le\alpha^{-1}\|f\|_{\mathbb{W}^*}.$$

Conversely, if (20) admits a unique solution u for any $f\in\mathbb{W}^*$, then B has to be invertible, which implies (21) by Theorem 1.
2 To show the equivalence of the inf-sup conditions (21), (22), and (23) we rewrite Step 5 of the proof of Theorem 1:

$$\inf_{v\in\mathbb{V}}\sup_{w\in\mathbb{W}}\frac{\mathscr{B}[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}}=\|B^{-1}\|^{-1}_{L(\mathbb{W};\mathbb{V})}.$$

Furthermore,

$$\inf_{w\in\mathbb{W}}\sup_{v\in\mathbb{V}}\frac{\mathscr{B}[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}}=\inf_{w\in\mathbb{W}}\sup_{v\in\mathbb{V}}\frac{\langle Bv,w\rangle_{\mathbb{W}}}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}}=\inf_{w\in\mathbb{W}}\sup_{v\in\mathbb{V}}\frac{\langle v,B^*w\rangle_{\mathbb{V}}}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}}=\|B^{-*}\|^{-1}_{L(\mathbb{V};\mathbb{W})},$$

where $B^*\colon\mathbb{W}\to\mathbb{V}$ is the adjoint operator of B and $B^{-*}\colon\mathbb{V}\to\mathbb{W}$ is its inverse. Recalling that $\|B^*\|_{L(\mathbb{W};\mathbb{V})}=\|B\|_{L(\mathbb{V};\mathbb{W})}$ and $\|B^{-*}\|_{L(\mathbb{V};\mathbb{W})}=\|B^{-1}\|_{L(\mathbb{W};\mathbb{V})}$ we deduce the desired expression

$$\inf_{w\in\mathbb{W}}\sup_{v\in\mathbb{V}}\frac{\mathscr{B}[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}}=\|B^{-1}\|^{-1}_{L(\mathbb{W};\mathbb{V})}.$$

and conclude the proof. □

The equality in (23) might seem at first surprising but is just a consequence of $\|B^{-*}\|_{L(\mathbb{V};\mathbb{W})}=\|B^{-1}\|_{L(\mathbb{W};\mathbb{V})}$. In general, (21) is simpler to verify than (23) and α of (23) is the largest possible α in (21a). Moreover, the above proof readily gives the following result.

Corollary 1 (Well Posedness vs. Inf-Sup). *Assume that the variational problem* (20) *admits a unique solution $u\in\mathbb{V}$ for all $f\in\mathbb{W}^*$ so that*

$$\|u\|_{\mathbb{V}} \le C\|f\|_{\mathbb{W}^*}.$$

Then $\mathscr{B}$ satisfies the inf-sup condition (23) *with $\alpha \ge C^{-1}$.*

Proof. Since (20) admits a unique solution u for all f, we conclude that the operator $B \in L(\mathbb{V};\mathbb{W})$ of Theorem 1 is invertible and the solution operator $B^{-1} \in L(\mathbb{W};\mathbb{V})$ is bounded with norm $\|B^{-1}\|_{L(\mathbb{W};\mathbb{W})} \le C$, thanks to $\|u\|_{\mathbb{V}} \le C\|f\|_{\mathbb{W}^*}$. On the other hand, Step 2 in the proof of Theorem 2 shows that $\|B^{-1}\|^{-1}_{L(\mathbb{W};\mathbb{V})}$ is the optimal inf-sup constant α for $\mathscr{B}$, which yields $\alpha \ge C^{-1}$. □

2.4 Two Special Problem Classes

We next study two special cases included in the inf-sup theory. The first class are problems with coercive bilinear form and the second one comprises problems of saddle point type.

2.4.1 Coercive Bilinear Forms

An existence and uniqueness result for coercive bilinear forms was established by Lax and Milgram eight years prior to the result by Nečas [45]. Coercivity of $\mathscr{B}$ is a sufficient condition for existence and uniqueness but it is not necessary.

Corollary 2 (Lax-Milgram Theorem). *Let $\mathscr{B}\colon \mathbb{V}\times\mathbb{V} \to \mathbb{R}$ be a continuous bilinear form that is* coercive*, namely there exists $\alpha > 0$ such that*

$$\mathscr{B}[v, v] \ge \alpha\|v\|^2_{\mathbb{V}} \qquad \text{for all } v \in \mathbb{V}. \tag{25}$$

Then (10) *has a unique solution that satisfies* (24)*.*

Proof. Since (25) implies $\sup_{w\in\mathbb{V}} \mathscr{B}[v, w] \ge \mathscr{B}[v, v] \ge \alpha\|v\|^2_{\mathbb{V}}$ for all $0 \ne v \in \mathbb{V}$, both (21a) and (21b) follow immediately, whence Theorem 2 implies the assertion. □

If the bilinear form $\mathscr{B}$ is also symmetric, i. e.,

$$\mathscr{B}[v, w] = \mathscr{B}[w, v] \qquad \text{for all } v, w \in \mathbb{V},$$

then $\mathscr{B}$ is a scalar product on $\mathbb{V}$. The norm induced by $\mathscr{B}$ is the so-called *energy norm*

$$|||v|||_{\Omega} := \mathscr{B}[v, v]^{1/2}.$$

Coercivity and continuity of $\mathscr{B}$ in turn imply that $|||\cdot|||_{\Omega}$ is equivalent to the natural norm $\|\cdot\|_{\mathbb{V}}$ in $\mathbb{V}$ since

$$\alpha\|v\|^2_{\mathbb{V}} \le |||v|||^2_{\Omega} \le \|\mathscr{B}\|\|v\|^2_{\mathbb{V}} \qquad \text{for all } v \in \mathbb{V}. \tag{26}$$

Moreover, it is rather easy to show that for symmetric and coercive $\mathscr{B}$ the solution u of (10) is the unique minimizer of the quadratic energy

$$J[v] := \frac{1}{2}\mathscr{B}[v,v] - \langle f, v\rangle \qquad \text{for all } v \in \mathbb{V},$$

i. e., $u = \operatorname{argmin}_{v\in\mathbb{V}} J[v]$. The energy norm and the quadratic energy play a relevant role in both Chap. 8 and Chap. 9.

2.4.2 Saddle Point Problems

Given a pair of Hilbert spaces $(\mathbb{V},\mathbb{Q})$, we consider two continuous bilinear forms $a\colon \mathbb{V}\times\mathbb{V}\to\mathbb{R}$ and $b\colon \mathbb{Q}\times\mathbb{V}\to\mathbb{R}$. If $f\in\mathbb{V}^*$ and $g\in\mathbb{Q}^*$, then we seek a pair $(u,p)\in\mathbb{V}\times\mathbb{Q}$ solving the *saddle point problem*

$$a[u,v] + b[p,v] = \langle f, v\rangle \qquad \text{for all } v\in\mathbb{V}, \tag{27a}$$
$$b[q,u] = \langle g, q\rangle \qquad \text{for all } q\in\mathbb{Q}. \tag{27b}$$

Problem (27) is a variational problem which can of course be stated in the form (10). In doing so we define the product space $\mathbb{W} := \mathbb{V}\times\mathbb{Q}$, which is a Hilbert space with scalar product

$$\langle (v,q),(w,r)\rangle_{\mathbb{W}} := \langle v,w\rangle_{\mathbb{V}} + \langle q,r\rangle_{\mathbb{Q}} \qquad \text{for all } (v,q),(w,r)\in\mathbb{W}$$

and induced norm $\|(v,q)\|_{\mathbb{W}} := (\|v\|_{\mathbb{V}}^2 + \|q\|_{\mathbb{Q}}^2)^{1/2}$. From the bilinear forms a and b we define the bilinear form $\mathscr{B}\colon \mathbb{W}\times\mathbb{W}\to\mathbb{R}$ by

$$\mathscr{B}[(v,q),(w,r)] := a[v,w] + b[q,w] + b[r,v] \qquad \text{for all } (v,q),(w,r)\in\mathbb{W}.$$

Then, (27) is equivalent to the problem

$$(u,p)\in\mathbb{W}: \qquad \mathscr{B}[(u,p),(v,q)] = \langle f,v\rangle + \langle g,q\rangle \qquad \text{for all } (v,q)\in\mathbb{W}. \tag{28}$$

To see this, test (28) first with $(v,0)$, which gives (27a), and then utilizing $(0,q)$ yields (27b). Obviously, a solution (u,p) to (27) is a solution to (28) and vice versa.

Therefore, the saddle point problem (27) is well-posed if and only if $\mathscr{B}$ satisfies the inf-sup condition (23). Since $\mathscr{B}$ is defined via the bilinear forms a and b and due to the degenerate structure of (27) it is not that simple to show (23). However it is a direct consequence of the inf-sup theorem for saddle point problems given by Brezzi in 1974 [17].

Theorem 3 (Brezzi Theorem). *The saddle point problem* (27) *has a unique solution* $(u,p)\in\mathbb{V}\times\mathbb{Q}$ *for all data* $(f,g)\in\mathbb{V}^*\times\mathbb{Q}^*$*, that depends continuously on data, if and only if there exist constants* $\alpha,\beta>0$ *such that*

$$\inf_{v\in\mathbb{V}_0}\sup_{w\in\mathbb{V}_0}\frac{a[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{V}}}=\inf_{w\in\mathbb{V}_0}\sup_{v\in\mathbb{V}_0}\frac{a[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{V}}}=\alpha>0, \tag{29a}$$

$$\inf_{q\in\mathbb{Q}}\sup_{v\in\mathbb{V}}\frac{b[q,v]}{\|q\|_{\mathbb{Q}}\|v\|_{\mathbb{V}}}=\beta>0, \tag{29b}$$

where

$$\mathbb{V}_0:=\{v\in\mathbb{V}\mid b[q,v]=0 \text{ for all } q\in\mathbb{Q}\}.$$

In addition, there exists $\gamma=\gamma(\alpha,\beta,\|a\|)$ *such that the solution* (u,p) *is bounded by*

$$\big(\|u\|_{\mathbb{V}}^2+\|p\|_{\mathbb{Q}}^2\big)^{1/2}\le\gamma\big(\|f\|_{\mathbb{V}^*}^2+\|g\|_{\mathbb{Q}^*}^2\big)^{1/2}. \tag{30}$$

Proof. 1 Continuity of b implies that the subspace $\mathbb{V}_0$ of $\mathbb{V}$ is closed. We therefore can decompose $\mathbb{V}=\mathbb{V}_0\oplus\mathbb{V}_\perp$ where $\mathbb{V}_\perp$ is the orthogonal complement of $\mathbb{V}_0$ in $\mathbb{V}$; see [16, (2.3.5) Proposition], [38, Theorem 5.6]. Both $\mathbb{V}_0$ and $\mathbb{V}_\perp$ are Hilbert spaces. 2 The inf-sup condition (29b) is (21a) for $\mathscr{B}=b$. On the other hand, by definition of $\mathbb{V}_0$, for every $v\in\mathbb{V}_\perp$ there exists a $q\in\mathbb{Q}$ with $b[q,v]\neq 0$, which is (21b). Hence, the equivalence of (21) and (23) implies that the operators $B:\mathbb{Q}\to\mathbb{V}_\perp$ and $B^*:\mathbb{V}_\perp\to\mathbb{Q}$ defined by

$$\langle Bq,v\rangle_{\mathbb{V}}=\langle B^*v,q\rangle_{\mathbb{Q}}=b[q,v]\qquad\text{for all } q\in\mathbb{Q},v\in\mathbb{V}_\perp,$$

are isomorphisms.
3 We write the solution $u=u_0+u_\perp$ with $u_0\in\mathbb{V}_0$ and $u_\perp\in\mathbb{V}_\perp$ to be determined as follows. Since B^* is an isomorphism, the problem

$$u_\perp\in\mathbb{V}_\perp:\qquad b[q,u_\perp]=\langle B^*v,q\rangle_{\mathbb{Q}}=\langle g,q\rangle\qquad\text{for all } q\in\mathbb{Q} \tag{31}$$

is well-posed for all $g\in\mathbb{Q}^*$, and selects $u_\perp$ uniquely. We next consider

$$u_0\in\mathbb{V}_0:\qquad a[u_0,v]=\langle f,v\rangle-a[u_\perp,v]\qquad\text{for all } v\in\mathbb{V}_0. \tag{32}$$

This problem admits a unique solution u_0 thanks to (29b), which is (23) with $\mathscr{B}=a$.
4 Upon setting

$$\langle F,v\rangle:=\langle f,v\rangle-a[u,v]\qquad\text{for all } v\in\mathbb{V}$$

we see that $F\in\mathbb{V}_\perp^*$ because $\langle F,v\rangle=0$ for all $v\in\mathbb{V}_0$ by (32). Since B is an isomorphism, there is a unique solution of

$$p\in\mathbb{Q}:\qquad b[p,v]=\langle Bp,v\rangle_{\mathbb{V}}=\langle F,v\rangle\qquad\text{for all } v\in\mathbb{V}_\perp. \tag{33}$$

This construction yields the desired pair (u,p) and shows that problems (31), (32), and (33) are well-posed if and only if b satisfies (29b) and a fulfills (29a).
5 We conclude by estimating (u,p). In view of (29b), $u_\perp$ is bounded by

$$\|u_\perp\|_{\mathbb{V}}\le\beta^{-1}\|g\|_{\mathbb{Q}^*}$$

which, in conjunction with (29a), implies for u_0

$$\|u_0\|_{\mathbb{V}} \le \alpha^{-1}\big(\|f\|_{\mathbb{V}^*} + \|a\|\|u_\perp\|_{\mathbb{V}}\big) \le \alpha^{-1}\|f\|_{\mathbb{V}^*} + \|a\|(\alpha\beta)^{-1}\|g\|_{\mathbb{Q}^*}.$$

Hence,

$$\|u\|_{\mathbb{V}} \le \|u_0\|_{\mathbb{V}} + \|u_\perp\|_{\mathbb{V}} \le \alpha^{-1}\|f\|_{\mathbb{V}^*} + \big(1+\alpha^{-1}\|a\|\big)\beta^{-1}\|g\|_{\mathbb{Q}^*}.$$

Finally, using $\|F\|_{\mathbb{V}_\perp^*} = \|F\|_{\mathbb{V}^*} \le \|f\|_{\mathbb{V}^*} + \|a\|\|u\|_{\mathbb{V}}$, (29b) gives the bound for p

$$\|p\|_{\mathbb{Q}} \le \beta^{-1}\|F\|_{\mathbb{V}^*} \le \beta^{-1}\big(1+\alpha^{-1}\|a\|\big)\ \big(\|f\|_{\mathbb{V}^*} + \beta^{-1}\|a\|\|g\|_{\mathbb{Q}^*}\big).$$

Adding the two estimates gives the stability bound (30) with $\gamma = \gamma(\alpha,\beta,\|a\|)$. □

Remark 3 (Optimal constant). A better bound of the stability constant γ in terms of α,β and $\|a\|$ is available. Setting

$$\kappa := \frac{\|a\|}{\beta}, \quad \kappa_{11} := \frac{1+\kappa^2}{\alpha^2}, \quad \kappa_{22} := \kappa^2\kappa_{11} + \frac{1}{\beta^2}, \quad \kappa_{12} := \kappa\kappa_{11},$$

Xu and Zikatanov have derived the bound [78]

$$\gamma \le \kappa_{12} + \max(\kappa_{11},\kappa_{22}).$$

For establishing this improved bound one has to make better use of the orthogonal decomposition $\mathbb{V} = \mathbb{V}_0 \oplus \mathbb{V}_\perp$ when estimating $u = u_0 + u_\perp$ and one has to resort to a result of Kato for non-trivial idempotent operators [42].

Combining the Brezzi theorem with Corollary 1 we infer the inf-sup condition for the bilinear form $\mathscr{B}$ in (28).

Corollary 3 (Inf-Sup of $\mathscr{B}$). *Let the bilinear form $\mathscr{B}\colon \mathbb{W} \to \mathbb{W}$ be defined by* (28). *Then there holds*

$$\inf_{(v,q)\in\mathbb{W}} \sup_{(w,r)\in\mathbb{W}} \frac{\mathscr{B}[(v,q),(w,r)]}{\|(v,q)\|_{\mathbb{W}}\|(w,r)\|_{\mathbb{W}}} = \inf_{(w,r)\in\mathbb{W}} \sup_{(v,q)\in\mathbb{W}} \frac{\mathscr{B}[(v,q),(w,r)]}{\|(v,q)\|_{\mathbb{W}}\|(w,r)\|_{\mathbb{W}}} \ge \gamma^{-1},$$

where γ is the stability constant from Theorem 3.

Assume that $a\colon \mathbb{V}\times\mathbb{V} \to \mathbb{R}$ is symmetric and let (u,p) be the solution to (27). Then u is the unique minimizer of the energy $J[v] := \frac{1}{2}a[v,v] - \langle f, v\rangle$ under the constraint $b[\cdot,u] = g$ in $\mathbb{Q}^*$. In view of this, p is the corresponding Lagrange multiplier and the pair (u,p) is the unique saddle point of the Lagrangian

$$L[v,q] := J[v] + b[q,v] - \langle g, q\rangle \qquad \text{for all } v\in\mathbb{V}, q\in\mathbb{Q}.$$

The Brezzi theorem also applies to non-symmetric a, in which case the pair (u,p) is no longer a saddle point.

2.5 Applications

We now review the examples introduced in Sect. 2.2 in light of the inf-sup theory.

2.5.1 Model Problem

Since $\boldsymbol{A}$ is symmetric, the variational formulation of the model problem in Sect. 2.2.1 leads to the symmetric bilinear form $\mathscr{B}\colon H_0^1(\Omega)\times H_0^1(\Omega)\to\mathbb{R}$ defined by

$$\mathscr{B}[w,v] := \int_\Omega \nabla v\cdot\boldsymbol{A}(x)\nabla w, \qquad \text{for all } v,w\in H_0^1(\Omega).$$

We have to decide which norm to use on $H_0^1(\Omega)$. The Poincaré-Friedrichs inequality (7) implies the equivalence of $\|\cdot\|_{H^1(\Omega)}$ and $|\cdot|_{H^1(\Omega)}$ on $H_0^1(\Omega)$ because

$$|v|_{H^1(\Omega)} \le \|v\|_{H^1(\Omega)} \le \left(1+C_d^2\,|\Omega|^{2/d}\right)^{1/2}|v|_{H^1(\Omega)} \qquad \text{for all } v\in H_0^1(\Omega). \quad (34)$$

On the other hand, assumption (12) on the eigenvalues of $\boldsymbol{A}$ directly leads to

$$\alpha_1\,|v|_{H^1(\Omega)}^2 \le \mathscr{B}[v,v] \le \alpha_2\,|v|_{H^1(\Omega)}^2 \qquad \text{for all } v\in H_0^1(\Omega).$$

Therefore, $|\cdot|_{H_0^1(\Omega)}$ is a convenient norm on $\mathbb{V}=H_0^1(\Omega)$ for the model problem, for which $\mathscr{B}$ is coercive with constant $\alpha=\alpha_1$ and continuous with norm $\|\mathscr{B}\|=\alpha_2$.

To apply the Lax-Milgram theorem it remains to show that $f\in L^2(\Omega)$ implies $f\in\mathbb{V}^*=H^{-1}(\Omega)$, in the sense that $v\mapsto\int_\Omega f\,v$ belongs to $H^1{-}1(\Omega)$. Recalling

$$\|f\|_{H^{-1}(\Omega)} = \sup_{v\in H_0^1(\Omega)}\frac{\langle f,v\rangle}{|v|_{H^1(\Omega)}}.$$

and using the Poincaré-Friedrichs inequality (7) once more we estimate

$$|\langle f,v\rangle| = \left|\int_\Omega f\,v\right| \le \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \le C_d\,|\Omega|^{1/d}\,\|f\|_{L^2(\Omega)}\,|v|_{H^1(\Omega)},$$

and therefore $\|f\|_{H^{-1}(\Omega)}\le C_d\,|\Omega|^{1/d}\,\|f\|_{L^2(\Omega)}$. In view of Corollary 2, we have the stability bound

$$|u|_{H^1(\Omega)} \le \frac{C_d|\Omega|^{1/d}}{\alpha_1^{1/2}}\|f\|_{L^2(\Omega)}.$$

Since $\mathscr{B}$ is symmetric and coercive, it defines a scalar product in $H_0^1(\Omega)$. Consequently, an even more convenient choice of norm on $\mathbb{V}$ is the energy norm $|||\cdot|||_\Omega=\mathscr{B}[\cdot,\cdot]^{1/2}$. In this case we have $\alpha=\|\mathscr{B}\|=1$ and

$$\|f\|_{H^{-1}(\Omega)} = \sup_{v\in H_0^1(\Omega)} \frac{\langle f, v\rangle}{|||v|||_\Omega} \le \frac{C_d\,|\Omega|^{1/d}}{\alpha_1^{1/2}} \|f\|_{L^2(\Omega)}$$

whence we obtain the same stability estimate as above.

2.5.2 Other Boundary Value Problems

We now review the examples from Sect. 2.2.2.

General 2nd Order Elliptic Operator. We take $\mathbb{V} = H_0^1(\Omega)$ and the bilinear form

$$\mathscr{B}[w, v] := \int_\Omega \nabla v \cdot \boldsymbol{A}(x)\nabla w + v\boldsymbol{b}\cdot\nabla w + c\,v\,w \qquad \text{for all } v, w \in H_0^1(\Omega).$$

A straightforward estimate shows continuity of $\mathscr{B}$ with respect to the norm $\|\cdot\|_{H^1(\Omega)}$

$$|\mathscr{B}[w, v]| \le \|\mathscr{B}\|\,\|v\|_{H^1(\Omega)}\,\|w\|_{H^1(\Omega)} \qquad \text{for all } v, w \in H^1(\Omega)$$

with operator norm $\|\mathscr{B}\| \le \alpha_2 + \|\boldsymbol{b}\|_{L^\infty(\Omega;\mathbb{R}^d)} + \|c\|_{L^\infty(\Omega)}$.

Assume now that $\operatorname{div}\boldsymbol{b}$ is bounded and $c - \frac{1}{2}\operatorname{div}\boldsymbol{b} \ge 0$ in Ω. In light of Green's formula (9) we get the identity

$$\int_\Omega v\boldsymbol{b}\cdot\nabla w = -\int_\Omega \nabla v\cdot\boldsymbol{b}\,w - \int_\Omega \operatorname{div}\boldsymbol{b}\,v\,w \qquad \text{for all } v, w \in H_0^1(\Omega),$$

whence $\int_\Omega v\boldsymbol{b}\cdot\nabla v = -\frac{1}{2}\int_\Omega \operatorname{div}\boldsymbol{b}\,v^2$. If $C = C_d\,|\Omega|^{1/d}$ is the Poincarè-Friedrichs constant for Ω, then we deduce as in Sect. 2.2.1 for any $v \in H_0^1(\Omega)$

$$\mathscr{B}[v, v] \ge \alpha_1\,|v|^2_{H^1(\Omega)} + \int_\Omega \big(c - \tfrac{1}{2}\operatorname{div}\boldsymbol{b}\big)v^2 \ge \alpha_1\,|v|^2_{H^1(\Omega)} \ge \frac{\alpha_1}{1+C^2}\|v\|^2_{H^1(\Omega)},$$

thanks to the norm equivalence (34). Using $\|\cdot\|_{H^1(\Omega)}$ as norm on $\mathbb{V}$ we have

$$\|f\|_{H^{-1}(\Omega)} = \sup_{v\in H_0^1(\Omega)} \frac{\langle f, v\rangle}{\|v\|_{H^1(\Omega)}} \le \|f\|_{L^2(\Omega)}.$$

Assuming only $c \ge 0$ the bilinear form $\mathscr{B}$ is no longer coercive. Nevertheless, for any bounded $\boldsymbol{b}$ and $c \ge 0$ it can be shown that $\mathscr{B}$ satisfies the inf-sup condition (23) but the proof is not elementary; see for instance [9].

The Biharmonic Equation. For the variational formulation of the biharmonic equation we use the Hilbert space $\mathbb{V} = H_0^2(\Omega)$ and claim that $\|\Delta\cdot\|_{L^2(\Omega)}$ is a norm on $H_0^2(\Omega)$ that is equivalent to $\|\cdot\|_{H^2(\Omega)}$. From Green's formula we deduce for $v \in C_0^\infty(\Omega)$

$$|v|^2_{H^2(\Omega)} = \sum_{i,j=1}^d \int_\Omega \left(\partial^2_{ij} v\right)^2 = -\sum_{i,j=1}^d \int_\Omega \partial_i v\, \partial^3_{ijj} v = \sum_{i,j=1}^d \int_\Omega \partial^2_{ii} v\, \partial^2_{jj} v = \|\Delta v\|^2_{L^2(\Omega)}.$$

Using density we thus conclude $|v|_{H^2(\Omega)} = \|\Delta v\|_{L^2(\Omega)}$ for all $v \in H^2_0(\Omega)$. For those functions v the Poincaré-Friedrichs inequality (7) implies $|v|_{H^1(\Omega)} \le c(\Omega)\, |v|_{H^2(\Omega)}$ which, in conjunction with the norm equivalence (34), yields

$$\|\Delta v\|_{L^2(\Omega)} \le \|v\|_{2,\Omega} \le C(\Omega)\ |v|_{H^2(\Omega)} = C(\Omega)\, \|\Delta v\|_{L^2(\Omega)}. \tag{35}$$

The bilinear form $\mathscr{B}$ given by

$$\mathscr{B}[w, v] = \int_\Omega \Delta v \Delta w$$

is symmetric and the energy norm $|||\cdot|||_\Omega$ coincides with the norm $\|\Delta \cdot\|_{L^2(\Omega)}$. Therefore, $\mathscr{B}$ is continuous and coercive on $H^2_0(\Omega)$ with constants $\|B\| = \alpha = 1$.

We denote by $H^{-2}(\Omega)$ the dual space of $H^2_0(\Omega)$. The norm equivalence (35) implies $\|f\|_{H^{-2}(\Omega)} \le C(\Omega)\|f\|_{L^2(\Omega)}$ for $f \in L^2(\Omega)$.

The 3d Eddy Current Equations. We take $\mathbb{V} = H_0(\mathrm{curl};\Omega)$ along with the symmetric bilinear form

$$\mathscr{B}[\boldsymbol{w}, \boldsymbol{v}] := \int_\Omega \mu\, \mathrm{curl}\, \boldsymbol{v} \cdot \mathrm{curl}\, \boldsymbol{w} + \kappa\, \boldsymbol{v} \cdot \boldsymbol{w} \qquad \text{for all } \boldsymbol{v}, \boldsymbol{w} \in \mathbb{V}$$

and subordinate energy norm

$$|||\boldsymbol{v}|||^2_\Omega = \|\mu^{1/2}\, \mathrm{curl}\, \boldsymbol{v}\|^2_{L^2(\Omega;\mathbb{R}^3)} + \|\kappa^{1/2} \boldsymbol{v}\|^2_{L^2(\Omega;\mathbb{R}^3)}.$$

Since $\mu, \kappa > 0$, this norm and the corresponding $H(\mathrm{curl};\Omega)$ norm (i.e. $\mu = \kappa = 1$) are equivalent. Accordingly, $\mathscr{B}$ is continuous and coercive with respect to $|||\cdot|||_\Omega$ with $\|\mathscr{B}\| = \alpha = 1$.

Furthermore, any $\boldsymbol{f} \in L^2(\Omega;\mathbb{R}^3)$ belongs to the dual space $\mathbb{V}^* = (H_0(\mathrm{curl};\Omega))^*$ and $\|\boldsymbol{f}\|_{\mathbb{V}^*} \le \kappa^{-1/2} \|\boldsymbol{f}\|_{L^2(\Omega;\mathbb{R}^3)}$.

The Stokes System. We use the Hilbert spaces $\mathbb{V} = H^1_0(\Omega;\mathbb{R}^d)$ equipped with the norm $|\cdot|_{H^1_0(\Omega;\mathbb{R}^d)}$ and $\mathbb{Q} = L^2_0(\Omega)$ equipped with $\|\cdot\|_{L^2(\Omega)}$. With this choice, $\|\cdot\|_{\mathbb{V}}$ is the energy norm associated with the bilinear form $a[\boldsymbol{w}, \boldsymbol{v}] = \int_\Omega \nabla \boldsymbol{v} : \nabla \boldsymbol{w}$. Therefore, a is continuous and coercive on $\mathbb{V}$ with $\|a\| = \alpha = 1$. This implies the inf-sup condition (29a).

Using integration by parts one can show $\|\operatorname{div} \boldsymbol{v}\|_{L^2(\Omega)} \le |\boldsymbol{v}|_{H^1_0(\Omega;\mathbb{R}^d)}$, whence the bilinear form $b[q, \boldsymbol{v}] = -\int_\Omega \operatorname{div} \boldsymbol{v}\, q$ is continuous with norm $\|b\| = 1$. In addition, for any $q \in L^2_0(\Omega)$ there exists a $\boldsymbol{w} \in H^1_0(\Omega;\mathbb{R}^d)$ such that

$$-\operatorname{div} \boldsymbol{w} = q \quad \text{in } \Omega \qquad \text{and} \qquad |\boldsymbol{w}|_{H^1(\Omega;\mathbb{R}^d)} \le C(\Omega) \|q\|_{L^2(\Omega)}.$$

This non-trivial result goes back to Nečas [19] and a proof can for instance be found in [36, Theorem III.3.1]. This implies

$$\sup_{\boldsymbol{v}\in H^1_0(\Omega;\mathbb{R}^d)} \frac{b[q,\boldsymbol{v}]}{|\boldsymbol{v}|_{H^1(\Omega;\mathbb{R}^d)}} \geq \frac{b[q,\boldsymbol{w}]}{|\boldsymbol{w}|_{H^1(\Omega;\mathbb{R}^d)}} = \frac{\|q\|^2_{L^2(\Omega)}}{|\boldsymbol{w}|_{H^1(\Omega;\mathbb{R}^d)}} \geq C(\Omega)^{-1}\|q\|_{L^2(\Omega)}.$$

Therefore, (29b) holds with $\beta \geq C(\Omega)^{-1}$ and Theorem 3 applies for all $\boldsymbol{f} \in L^2(\Omega;\mathbb{R}^d)$ and gives existence, uniqueness and stability of the solution $(\boldsymbol{u},p) \in \mathbb{V}\times\mathbb{Q}$ of the Stokes system.

2.6 Problems

Problem 1. Let $\Omega = (0,1)$ and $u \in W^1_p(\Omega)$ with $1 < p \leq \infty$. Prove that the function u is $(p-1)/p$-Hölder continuous, namely

$$|u(x)-u(y)| \leq |x-y|^{(p-1)/p}\|u'\|_{L^p(\Omega)} \qquad \text{for all } x,y \in \Omega.$$

If $p = 1$, then $u \in W^1_1(\Omega)$ is uniformly continuous in $\overline{\Omega}$ because of the absolute continuity of the integral.

Problem 2. Find the weak gradient of $v(x) = \log\log(|x|/2)$ in the unit ball Ω, and show that $v \in W^1_d(\Omega)$ for $d \geq 2$. This shows that functions in $W^1_d(\Omega)$, and in particular in $H^1(\Omega)$, may not be continuous, and even bounded, in dimension $d \geq 2$.

Problem 3. Prove the following simplified version of the Poincaré-Friedrichs inequality (7): let Ω be contained in the strip $\{x \in \mathbb{R}^d \mid 0 < x_d < h\}$; then

$$\|v\|_{L^2(\Omega)} \lesssim h\|\nabla v\|_{L^2(\Omega)} \quad \text{for all } v \in H^1_0(\Omega).$$

To this end, take $v \in C^\infty_0(\Omega)$, write for $0 < s < h$

$$v^2(x,s) = v^2(x,0) + 2\int_0^s \partial_d v \cdot v, \tag{36}$$

integrate, and use Cauchy-Schwarz inequality to prove (7). Next use a density argument, based on the definition of $H^1_0(\Omega)$, to extend the inequality to $H^1_0(\Omega)$.

Problem 4. Let $\Omega_h = \{(x',x_d) \mid |x| < h, x_d > 0\}$ be the upper half ball in $\mathbb{R}^d$ of radius $h > 0$ centered at the origin. Let Γ_h be the flat part of $\partial\Omega_h$.

(a) Let $\zeta \geq 0$ be a C^∞_0 cut-off function in the unit ball that equals 1 in the ball of radius $1/2$. Use the identity (36) for $v\zeta$, followed by a density argument, to derive the trace inequality

$$\|v\|^2_{L^2(\Gamma_{1/2})} \lesssim \|v\|^2_{L^2(\Omega_1)} + \|\nabla v\|^2_{L^2(\Omega_1)} \qquad \text{for all } v \in H^1(\Omega_1).$$

(b) Use a scaling argument to Ω_h to deduce the scaled trace inequality

$$\|v\|^2_{L^2(\Gamma_{h/2})} \lesssim h^{-1}\|v\|^2_{L^2(\Omega_h)} + h\|\nabla v\|^2_{L^2(\Omega_h)} \qquad \text{for all } v \in H^1(\Omega_h).$$

Problem 5. Show that $\operatorname{div} \boldsymbol{q} \in H^{-1}(\Omega)$ for $\boldsymbol{q} \in L^2(\Omega;\mathbb{R}^d)$. Compute the corresponding H^{-1}-norm.

Problem 6. (a) Find a variational formulation which amounts to solving

$$-\Delta u = f \quad \text{in } \Omega, \qquad \partial_\nu u + pu = g \quad \text{on } \partial\Omega,$$

where $f \in L^2(\Omega)$, $g \in L^2(\partial\Omega)$, $0 < p_1 \le p \le p_2$ on $\partial\Omega$. Show that the bilinear form is coercive in $H^1(\Omega)$.

(b) Suppose that $p = \varepsilon^{-1} \to \infty$ and denote the corresponding solution by u_ε. Determine the boundary value problem satisfied by $u_0 = \lim_{\varepsilon \downarrow 0} u_\varepsilon$.

(c) Derive an error estimate for $\|u_0 - u_\varepsilon\|_{H^1(\Omega)}$.

Problem 7. Let $\boldsymbol{A}$ be uniformly SPD and $c \in L^\infty(\Omega)$ satisfy $c \ge 0$. Consider the quadratic functional

$$I[v] = \frac{1}{2}\int_\Omega \nabla v \cdot \boldsymbol{A}(x)\nabla v + c(x)v^2 - \langle f, v\rangle \quad \text{for all } v \in H^1_0(\Omega),$$

where $f \in H^{-1}(\Omega)$. Show that $u \in H^1_0(\Omega)$ is a minimizer of $I[v]$ if and only if u satisfies the Euler-Lagrange equation

$$\mathscr{B}[u, v] = \int_\Omega \nabla v \cdot \boldsymbol{A}\nabla u + c\,uv = \langle f, v\rangle \quad \text{for all } v \in H^1_0(\Omega).$$

Problem 8. Consider the model problem with Neumann boundary condition

$$-\operatorname{div}(\boldsymbol{A}\nabla u) = f \quad \text{in } \Omega, \qquad \boldsymbol{n}\cdot\boldsymbol{A}\nabla u = g \quad \text{on } \partial\Omega$$

(a) Derive the variational formulation in $\mathbb{V} = H^1(\Omega)$ and show that the bilinear form $\mathscr{B}$ is continuous and symmetric but not coercive.

(b) Let $\mathbb{V}$ be the subspace of $H^1(\Omega)$ of functions with vanishing mean value. Show that $\mathscr{B}$ is coercive.

(c) Derive a compatibility condition between f and g for existence of a weak solution.

Problem 9. Consider the space $\mathbb{V} = H(\operatorname{div};\Omega) = \{\boldsymbol{q} \in L^2(\Omega;\mathbb{R}^d) \mid \operatorname{div}\boldsymbol{q} \in L^2(\Omega)\}$, and the bilinear form

$$\mathscr{B}[\boldsymbol{p}, \boldsymbol{q}] = \int_\Omega \operatorname{div}\boldsymbol{p}\operatorname{div}\boldsymbol{q} + \boldsymbol{p}\cdot\boldsymbol{q} \qquad \text{for all } \boldsymbol{p}, \boldsymbol{q} \in \mathbb{V}.$$

(a) Show that $\mathbb{V}$ is a Hilbert space and that $\mathscr{B}$ is symmetric, continuous and coercive in $H(\operatorname{div};\Omega)$.

(b) Determine the strong form of the PDE and implicit boundary condition corresponding to the variational formulation

$$\boldsymbol{p} \in \mathbb{V}: \qquad \mathscr{B}[\boldsymbol{p}, \boldsymbol{q}] = \langle \mathbf{f}, \boldsymbol{q} \rangle \qquad \text{for all } \boldsymbol{q} \in \mathbb{V}.$$

Problem 10. Let $\boldsymbol{\sigma} := -\boldsymbol{A}\nabla u$ be the flux of the model problem, which can be written equivalently as

$$\boldsymbol{A}^{-1}\boldsymbol{\sigma} + \nabla u = 0, \qquad \operatorname{div} \boldsymbol{\sigma} = -f.$$

(a) Let $\mathbb{V} = H(\operatorname{div}; \Omega)$ and $\mathbb{Q} = L^2_0(\Omega)$. Multiply the first equation by $\boldsymbol{\tau} \in \mathbb{V}$ and integrate by parts using Green's formula (9). Multiply the second equation by $v \in \mathbb{Q}$. Write the resulting variational formulation in the form (27) and show that (29) is satisfied.

(b) Apply Theorem 3 to deduce existence, uniqueness, and stability.

3 The Petrov-Galerkin Method and Finite Element Bases

The numerical approximation of boundary value problems is typically an effective way, and often the only one available, to extract quantitive information about their solutions. In this chapter we introduce the finite element method (FEM) which, due to its geometric flexibility, practical implementation, and powerful and elegant theory, is one of the most successful discretization methods for this task.

Roughly speaking, a finite element method consists in computing the Petrov-Galerkin solution with respect to a finite-dimensional space and that space is constructed from local function spaces (finite elements), which are glued together by some continuity condition.

We first analyze Petrov-Galerkin approximations and then review Lagrange elements, the most basic and common finite element spaces; for other finite element spaces, we refer to the standard finite element literature, e.g. [15, 16, 18, 25, 51].

3.1 Petrov-Galerkin Solutions

The solution of a boundary value problem cannot be computed, since the solution is characterized by an infinite number of (linearly-independent) conditions. To overcome this principal obstacle, we replace the boundary value problem by its Petrov-Galerkin discretization.

3.1.1 Definition, Existence and Uniqueness

To obtain a computable approximation to a solution to the variational problem (10) we simply restrict the continuous spaces $\mathbb{V}, \mathbb{W}$ in (10) to finite dimensional subspaces of equal dimension $N < \infty$. As we shall see, this leads to a linear system in $\mathbb{R}^{N\times N}$ which can be solved by standard methods.

Definition 3 (Discrete Solution). For $N \in \mathbb{N}$ let $\mathbb{V}_N \subset \mathbb{V}$ and $\mathbb{W}_N \subset \mathbb{W}$ be subspaces of equal dimension N. Then a solution U_N to

$$U_N \in \mathbb{V}_N : \qquad \mathscr{B}[U_N, W] = \langle f, W\rangle \qquad \text{for all } W \in \mathbb{W}_N \tag{37}$$

is called *Petrov-Galerkin Solution*.

Remark 4. For $\mathbb{V} \neq \mathbb{W}$ the test functions $W \in \mathbb{W}_N$ in (37) are different from the ansatz functions $V \in \mathbb{V}_N$ which results in the naming *Petrov-Galerkin discretization*. If the continuous spaces $\mathbb{V} = \mathbb{W}$ are equal, then we will choose also the same discrete space $\mathbb{V}_N = \mathbb{W}_N$. In this case, (37) is called *Galerkin discretization* and, if additionally $\mathscr{B}$ is symmetric and coercive, it is called *Ritz-Galerkin discretization*. In any case, the discrete spaces are subsets of the continuous ones, and thus all dis-

crete functions belong to the continuous function spaces. For this reason, the method is called a *conforming discretization* of (10).

For any conforming discretization, the bilinear form $\mathscr{B}$ is well defined and continuous on the discrete pair $\mathbb{V}_N \times \mathbb{W}_N$. The continuity constant is bounded by $\|\mathscr{B}\|$. This can easily be seen, since all discrete functions $V \in \mathbb{V}_N$ and $W \in \mathbb{W}_N$ are admissible in (16). In the same vain, for a coercive form $\mathscr{B}\colon \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ we are allowed to use any discrete function $V \in \mathbb{V}_N$ in (25) yielding

$$\mathscr{B}[V,V] \geq c_{\mathscr{B}} \|V\|_{\mathbb{V}}^2 \qquad \text{for all } V \in \mathbb{V}_N.$$

Therefore coercivity of $\mathscr{B}$ is inherited for conforming discretizations from the continuous space to the discrete one with the same coercivity constant $c_{\mathscr{B}} > 0$. This in turn implies the existence and uniqueness of the Galerkin solution $U_N \in \mathbb{V}_N$.

Recalling the theorem of Lax-Milgram, stated as Corollary 2, we know that a coercive form $\mathscr{B}$ satisfies the inf-sup condition (23). Since coercivity is inherited to subspaces we can conclude in this case the discrete counterpart of (23), namely

$$\inf_{V \in \mathbb{V}_N} \sup_{W \in \mathbb{W}_N} \frac{\mathscr{B}[V,W]}{\|V\|_{\mathbb{V}} \|W\|_{\mathbb{W}}} = \inf_{W \in \mathbb{W}_N} \sup_{V \in \mathbb{V}_N} \frac{\mathscr{B}[V,W]}{\|V\|_{\mathbb{V}} \|W\|_{\mathbb{W}}} = \beta_N \tag{38}$$

with a constant $\beta_N \geq c_{\mathscr{B}}$.

For general $\mathscr{B}$, the continuous inf-sup (23) does not imply the discrete one. In order to state a simple as possible criterion for the existence and uniqueness or a discrete solution, we consider the discrete operators $B_N \in L(\mathbb{V}_N; \mathbb{W}_N^*)$ and $B_N^* \in L(\mathbb{W}_N; \mathbb{V}_N^*)$, defined in the same way as B and B^* in Sect. 2.3 by

$$\langle B_N V, W \rangle = \langle B_N^* W, V \rangle = \mathscr{B}[V,W] \qquad \text{for all } V \in \mathbb{V}_N, W \in \mathbb{W}_N.$$

The discrete problem (37) is well-posed if and only if the operator B_N is an isomorphism from $\mathbb{V}_N$ to $\mathbb{W}_N^*$. Since we deal with finite dimensional spaces, a necessary condition for B_N being invertible is $\dim \mathbb{V}_N = \dim \mathbb{W}_N^* = \dim \mathbb{W}_N$, which we assume in the definition of the Petrov-Galerkin solution. Hence a neccessary and sufficient condition for invertibility of B_N is injectivity of B_N, which can be characterized by

$$\textit{for every } 0 \neq V \in \mathbb{V}_N \textit{ there exists } W \in \mathbb{W}_N \textit{ such that } \mathscr{B}[V,W] \neq 0. \tag{39}$$

As a direct consequence we can characterize the existence and uniqueness of the discrete solution.

Theorem 4 (Existence and Uniqueness of the Petrov-Galerkin Solution). *Let $\mathbb{V}_N \subset \mathbb{V}$ and $\mathbb{W}_N \subset \mathbb{W}$ be subspaces of equal dimension.*

Then for any $f \in \mathbb{W}_N^*$ there exists a unique Petrov-Galerkin solution $U_N \in \mathbb{V}_N$, i. e.,

$$U_N \in \mathbb{V}_N : \qquad \mathscr{B}[U_N, W] = \langle f, W \rangle \qquad \textit{for all } W \in \mathbb{W}_N,$$

if and only if (39) *is satisfied.*

Proof. As for the continuous problem (10) the existence and uniqueness of a discrete solution U_N for any $f \in \mathbb{W}_N^*$ is equivalent to the invertibility of the operator $B_N : \mathbb{V}_N \to \mathbb{W}_N^*$. The latter is equivalent to (39). □

Proposition 1. *Let $\mathbb{V}_N \subset \mathbb{V}$ and $\mathbb{W}_N \subset \mathbb{W}$ be subspaces of equal dimension.*
Then the following statements are equivalent:

(1) *The discrete inf-sup condition* (38) *holds for some $\beta_N > 0$;*
(2) $\inf\limits_{V\in\mathbb{V}_N} \sup\limits_{W\in\mathbb{W}_N} \dfrac{\mathscr{B}[V,W]}{\|V\|_{\mathbb{V}}\|W\|_{\mathbb{W}}} > 0;$
(3) $\inf\limits_{W\in\mathbb{W}_N} \sup\limits_{V\in\mathbb{V}_N} \dfrac{\mathscr{B}[V,W]}{\|V\|_{\mathbb{V}}\|W\|_{\mathbb{W}}} > 0;$
(4) *condition* (39) *is satisfied;*
(5) *for every $0 \neq W \in \mathbb{W}_N$ there exists $V \in \mathbb{V}_N$ such that $\mathscr{B}[V,W] \neq 0$.*

Proof. Obviously, (1) implies (2) and (3). The inf-sup condition (2) implies (4) and (3) yields (5). Statement (4) is equivalent to invertibility of $B_N \in L(\mathbb{V}_N, \mathbb{W}_N^*)$ and in the same way (5) is equivalent to invertibility of $B_N^* \in L(\mathbb{W}_N, \mathbb{V}_N^*)$, whence (4) and (5) are equivalent. Recalling Theorem 4, statement (4) is equivalent to existence and uniqueness of a discrete solution for any $f \in \mathbb{W}_N^*$. Applying Theorem 2 with $\mathbb{V}, \mathbb{W}$ replaced by $\mathbb{V}_N, \mathbb{W}_N$ the latter is equivalent with the inf-sup condition on $\mathbb{V}_N, \mathbb{W}_N$, i. e., (4) is equivalent to (1). □

This proposition allows for different conditions that imply existence and uniqueness of a discrete solution. Conditions (2)–(5) of Proposition 1 seem to be more convenient than (1) since we do not have to specify the discrete inf-sup constant β_N. However, the *value* of this constant is critical, as we shall see from the following section.

3.1.2 Stability and Quasi-Best Approximation

In this section we investigate the stability and approximation properties of Petrov-Galerkin solutions. In doing so, we explore properties that are uniform in the dimension N of the discrete spaces.

We start with the stability properties.

Corollary 4 (Stability of the Discrete Solution). *If* (38) *holds, then the Petrov-Galerkin solution U_N satisfies*

$$\|U_N\|_{\mathbb{V}} \leq \frac{1}{\beta_N} \|f\|_{\mathbb{W}^*}. \tag{40}$$

Proof. Use the same arguments as in the proof of Theorem 2 for the stability estimate of the true solution. □

We next relate the Petrov-Galerkin solution to the best possible approximation to the true solution u in $\mathbb{V}_N$ and show that U_N is up to a constant as close to u as the

best approximation. For coercive forms this is Cea's Lemma [22]. For general $\mathscr{B}$ this follows from the theories of Babuška [8, 9] and Brezzi [17].

The key for the best approximation property of the Petrov-Galerkin solution is the following relationship, which holds for all conforming discretizations and is usually referred to as *Galerkin orthogonality*:

$$\mathscr{B}[u-U_N, W] = 0 \qquad \text{for all } W \in \mathbb{W}_N. \tag{41}$$

If $\mathbb{V} = \mathbb{W}$, $\mathscr{B}$ symmetric and coercive, then this means that the error $u - U_N$ is orthogonal to $\mathbb{V}_N = \mathbb{W}_N$ in the energy norm $|||\cdot|||_\Omega$. To prove (41), simply observe that we are allowed to use any $W \in \mathbb{W}_N$ as a test function in the definition of the continuous solution (10), which gives

$$\mathscr{B}[u, W] = \langle f, W\rangle \qquad \text{for all } W \in \mathbb{W}_N.$$

Then recalling the definition of the Petrov-Galerkin solution and taking the difference yields (41).

Theorem 5 (Quasi-Best-Approximation Property). *Let $\mathscr{B}\colon \mathbb{V}\times\mathbb{V}\to\mathbb{R}$ be continuous and assume* (38) *is satisfied. Let u be the solution to* (10) *and let $U_N \in \mathbb{V}_N$ be the Petrov-Galerkin solution.*

Then the error $u - U_N$ satisfies the bound

$$\|u-U_N\|_{\mathbb{V}} \le \frac{\|\mathscr{B}\|}{\beta_N} \min_{V\in\mathbb{V}_N} \|u-V\|_{\mathbb{V}}.$$

Proof. We give a simplified proof, which follows Babuška [8, 9] and yields the constant $1+\frac{\|\mathscr{B}\|}{\beta_N}$. The asserted constant is due to Xu and Zikatanov [78].

Combining (38), (41), and the continuity of $\mathscr{B}$, we derive for all $V \in \mathbb{V}_N$

$$\beta_N\|U_N - V\|_{\mathbb{V}} \le \sup_{W\in\mathbb{W}_N} \frac{\mathscr{B}[U_N-V, W]}{\|W\|_{\mathbb{W}}} = \sup_{W\in\mathbb{W}_N} \frac{\mathscr{B}[u-V, W]}{\|W\|_{\mathbb{W}}} \le \|\mathscr{B}\|\|u-V\|_{\mathbb{V}},$$

whence

$$\|U_N - V\|_{\mathbb{V}} \le \frac{\|\mathscr{B}\|}{\beta_N}\|u-V\|_{\mathbb{V}}.$$

Using the triangle inequality yields

$$\|u-U_N\|_{\mathbb{V}} \le \|u-V\|_{\mathbb{V}} + \|V-U_N\|_{\mathbb{V}} \le \left(1+\frac{\|\mathscr{B}\|}{\beta_N}\right)\|u-V\|_{\mathbb{V}}$$

for all $V \in \mathbb{V}_N$. It just remains to minimize in $\mathbb{V}_N$. □

The last two results reveal the critical role of the discrete inf-sup constant β_N. If a sequence of spaces $\{(\mathbb{V}_N, \mathbb{W}_N)\}_{N\ge1}$ approximates the pair $(\mathbb{V}, \mathbb{W})$ with deteriorating $\beta_N \to 0$ as $N \to \infty$, then the sequence of discrete solutions $\{U_N\}_{N\ge1}$ is not guaranteed to be uniformly bounded. Furthermore, the discrete solutions in general

approximate the true solution with a reduce rate as compared to the best approximation within $\mathbb{V}_N$. For these reasons a lower bound for the discrete inf-sup constants becomes highly desirable.

Definition 4 (Stable Discretization). We call a sequence $\{(\mathbb{V}_N, \mathbb{W}_N)\}_{N\geq 1}$ of discrete spaces with inf-sup constants $\{\beta_N\}_{N\geq 1}$ stable if and only if there exists $\beta > 0$ such that

$$\inf_{N\geq 1} \beta_N \geq \beta > 0.$$

In contrast to the continuous inf-sup condition where one has to prove

$$\inf_{v\in\mathbb{V}} \sup_{w\in\mathbb{W}} \frac{\mathscr{B}[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}} > 0 \qquad \textit{and} \qquad \inf_{w\in\mathbb{W}} \sup_{v\in\mathbb{V}} \frac{\mathscr{B}[v,w]}{\|v\|_{\mathbb{V}}\|w\|_{\mathbb{W}}} > 0$$

it suffices in the discrete setting to show one

$$\inf_{V\in\mathbb{V}_N} \sup_{W\in\mathbb{W}_N} \frac{\mathscr{B}[V,W]}{\|V\|_{\mathbb{V}}\|W\|_{\mathbb{W}}} \geq \beta \qquad \textit{or} \qquad \inf_{W\in\mathbb{W}_N} \sup_{V\in\mathbb{V}_N} \frac{\mathscr{B}[V,W]}{\|V\|_{\mathbb{V}}\|W\|_{\mathbb{W}}} \geq \beta$$

in order to furnish a uniform lower bound for the discrete inf-sup constant $\beta_N \geq \beta$. This simplification stems from the assumption $\dim \mathbb{V}_N = \dim \mathbb{W}_N < \infty$. Allowing for infinite dimensional spaces $\mathbb{V}_N$ and $\mathbb{W}_N$ gives rise to both inf-sup conditions as in the continuous case.

3.1.3 Computation

In view of Theorem 5, the *quality* of the Petrov-Galerkin solution depends in particular on the approximation properties of the discrete spaces. Before embarking on the construction of suitable spaces, it is useful to see how a Petrov-Galerkin solution can be computed. This will reveal that the real task is the construction of a suitable basis and it will give hints towards what affects the *cost* of a Petrov-Galerkin solution.

Let $\phi_1, \ldots, \phi_N$ and $\psi_1, \ldots, \psi_N$ be bases of $\mathbb{V}_N$ and $\mathbb{W}_N$, respectively. Writing

$$U_N = \sum_{j=1}^{N} \alpha_j \phi_j,$$
$$K = (k_{ij})_{i,j=1,\ldots,N} \quad \text{with} \quad k_{ij} = \mathscr{B}[\phi_j, \psi_i],$$
$$F = (F_1, \ldots, F_N) \quad \text{with} \quad F_i = \langle f, \psi_i \rangle$$

the definition of the Petrov-Galerkin solution (37) is equivalent to the linear system

$$\alpha \in \mathbb{R}^N : \qquad K\alpha = F. \tag{42}$$

Its solution can be computed by various methods from numerical linear algebra. The method of choice as well as the cost is affected by the properties of the matrix. Of

course these properties depend on the bilinear form $\mathscr{B}[\cdot,\cdot]$ and on the chosen bases $\phi_1,\ldots,\phi_N$ and $\psi_1,\ldots,\psi_N$.

For example, in the case of the model problem of Sect. 2.2.1, $\mathbb{V}_N = \mathbb{W}_N$ and $\phi_i = \psi_i$ for $i = 1,\ldots,N$, the matrix K is symmetric positive definite, irrespective of the choice of $\phi_1,\ldots,\phi_N$. The linear system (42) gets trivial if we take $\phi_1,\ldots,\phi_N$ to be the eigenvectors of K. However, finding the eigenvectors of K is a nonlinear problem and typically more expensive than solving linear systems. On the other hand, taking the easily available polynomials for $\phi_1,\ldots,\phi_N$ will lead to full and ill-conditioned matrices in general.

Finite element bases provide a compromise between these two extremes. The basis functions can be relatively easily constructed and are locally supported. The latter leads to sparse matrices for bilinear forms associated with boundary value problems.

3.2 Finite Element Spaces

The choice, or better the construction, of suitable finite element spaces in the Petrov-Galerkin discretization is the subject of this section. We shall discuss here only the spaces of Lagrange elements, emphasizing the case of polynomial degree $n = 1$. These spaces are appropriate for our model problem of Sect. 2.2.1.

3.2.1 Simplices and Triangulations

As already mentioned, a key property of finite element bases is that there are locally supported. This is achieved with the help of a decomposition of the domain of the boundary value problem. Here we consider triangulations, which are build from simplices.

Definition 5 (Simplex and Subsimplices). Let $d \in \mathbb{N}$. A subset T of $\mathbb{R}^d$ is an *n-simplex in* $\mathbb{R}^d$ if there exist $n+1$ points $z_0,\ldots,z_n \in \mathbb{R}^d$ such that

$$T = \text{conv hull}\{z_0,\ldots,z_n\} = \left\{ \sum_{i=1}^{n} \lambda_i z_i \mid \lambda_i \geq 0 \text{ for } i = 0,\ldots,d, \sum_{i=0}^{n} \lambda_i = 1 \right\}$$

and $z_1 - z_0,\ldots,z_n - z_0$ are linearly independent vectors in $\mathbb{R}^d$. By convention, we refer to points as 0-simplices. A subset T' of T is a (proper) *k-subsimplex* of T if T' is a k-simplex such that

$$T' = \text{conv hull}\{z'_0,\ldots,z'_k\} \subset \partial T$$

with $k < n$ and $z'_0,\ldots,z'_k \in \{z_0,\ldots,z_d\}$.

The 0-subsimplices are the vertices of a simplex. Moreover, 1-subsimplices are edges and 2-subsimplices of 3-simplices are faces. We shall refer to $(n-1)$-subsimplices of n-simplices as sides.

Two d-simplices in $\mathbb{R}^d$ are always affine equivalent, meaning that one can be mapped onto the other by an affine bijection. This fact is useful for implementation and also for the theory that follows. The following lemma fixes a reference simplex and controls the affine bijection in terms of geometric quantities of the generic simplex.

Lemma 1 (Reference and Generic Simplex). *Let the* reference simplex *in $\mathbb{R}^d$ be defined as*

$$\hat{T} = \mathit{conv\ hull}\{0, e_1, \dots, e_d\},$$

where $e_1, \dots, e_d$ denotes the canonical basis in $\mathbb{R}^d$. For any d-simplex T in $\mathbb{R}^d$, there exists a bijective affine map

$$F_T : \hat{T} \to T, \quad \hat{x} \mapsto A_T \hat{x} + b_T$$

where $A_T \in \mathbb{R}^{d\times d}$ and $b_T \in \mathbb{R}^d$. If we define

$$\begin{aligned}
\overline{h}_T &:= \sup\{|x-y| \mid x, y \in T\},\\
\underline{h}_T &:= \sup\{2r \mid B_r \subset T \text{ is a ball of radius } r\},\\
h_T &:= |T|^{1/d},
\end{aligned}$$

there holds

$$\|A_T\| \le \overline{h}_T, \qquad \|A_T^{-1}\| \le \frac{C_d}{\underline{h}_T}, \qquad |\det A_T| = \frac{h_T^d}{d!}. \tag{43}$$

Proof. See Problem 14.

All three quantities in (43) measure somehow the size of the given simplex. In view of

$$\underline{h}_T \le h_T \le \overline{h}_T$$

they are equivalent up to the following quantity.

Definition 6 (Shape Coefficient). The *shape coefficient* of a d-simplex T in $\mathbb{R}^d$ is the ratio of the diameter and the inball diameter of T,

$$\sigma_T := \frac{\overline{h}_T}{\underline{h}_T}.$$

Of course this notion becomes useful when it refers to many simplices. This brings us to the notion of triangulation.

Definition 7 (Triangulation). Let $\Omega \subset \mathbb{R}^d$ be a bounded, polyhedral domain. A finite set $\mathscr{T}$ of d-simplices in $\mathbb{R}^d$ with

$$\overline{\Omega} = \bigcup_{T\in\mathscr{T}} T \qquad \text{and} \qquad |\Omega| = \sum_{T\in\mathscr{T}} |T| \tag{44}$$

is a *triangulation of* Ω. We denote the set of all vertices of $\mathscr{T}$ by $\mathscr{V}_\mathscr{T}$ and the set of all sides by $\mathscr{S}_\mathscr{T}$. The *shape coefficient* of a triangulation $\mathscr{T}$ is the quantity $\sigma_\mathscr{T} := \max_{T\in\mathscr{T}} \sigma_T$. A triangulation $\mathscr{T}$ is *conforming* if it satisfies the following property: if any two simplices $T_1, T_2 \in \mathscr{T}$ have a nonempty intersection $S = T_1 \cap T_2 \neq \emptyset$, then S is a k-subsimplex of both T_1 and T_2 with $k \in \{0,\dots,d\}$.

A sequence of triangulations $\{\mathscr{T}_k\}_{k\geq 0}$ is *shape regular* if $\sup_{T\in\mathscr{T}_k} \sigma_\mathscr{T} \leq C$. It is called *quasi-uniform* if there exists a constant C such that, for all k, there holds $\max_{T\in\mathscr{T}_k} \overline{h}_T \leq C \min_{T\in\mathscr{T}_k} \underline{h}_T$. In both cases we assume tacitly that the constant C is of moderate size.

The first condition in (44) ensures that $\mathscr{T}$ is a covering of the closure of Ω, while the second requires that there is no overlapping. Notice that the latter is required not in a set-theoretic but in a measure-theoretic manner. Conformity will turn out to be a very useful property when constructing bases that are regular across simplex boundaries.

3.2.2 Lagrange Elements

The purpose of this section is to show that the following finite-dimensional space is appropriate for our model problem in Sect. 2.2.1:

$$\mathbb{V}(\mathscr{T}) := \{v \in C(\overline{\Omega}) \mid v_{|T} \in \mathbb{P}_n(T) \quad \text{for all } T \in \mathscr{T} \ \text{ and } \ v_{|\partial\Omega} = 0\}$$

where $\mathscr{T}$ is a conforming triangulation of $\Omega \subset \mathbb{R}^d$ and $\mathbb{P}_n(T)$ stands for the space of polynomials with degree $\leq n$ over T. More precisely, we will show that $\mathbb{V}(\mathscr{T}) \subset H^1_0(\Omega)$ possesses a basis which is locally supported and easy to implement, and conclude with approximation properties of $\mathbb{V}(\mathscr{T})$. In what follows, this will be called the *standard discretization* of the model problem.

Lemma 2 (H^1_0-Conformity). *If $\mathscr{T}$ is a conforming triangulation of a bounded, polyhedral Lipschitz domain $\Omega \subset \mathbb{R}^d$, then $\mathbb{V}(\mathscr{T}) \subset H^1_0(\Omega)$.*

Proof. Let $v \in \mathbb{V}(\mathscr{T})$. We start by checking that v has a weak derivative. For any test function $\eta \in C^\infty_0(\Omega)$ and $i \in \{1,\dots,d\}$ there holds

$$\int_\Omega v\partial_i\eta = \sum_{T\in\mathscr{T}} \int_T v\partial_i\eta = \sum_{T\in\mathscr{T}} \int_T (\partial_i v)\eta + \sum_{T\in\mathscr{T}} \sum_{S\subset\partial T} \int_S v\eta n_{T,i},$$

where $n_{T,i}$ is the i-th coordinate of the exterior normal to ∂T. The second sum on the right hand side vansishes for the following reasons: if $S \subset \partial\Omega$, then there holds $\eta_{|S} = 0$; otherwise there exists a unique simplex $T' \in \mathscr{T}$ such that $S = T \cap T'$ and $n_{T',i} = -n_{T,i}$. Consequently, $w \in L^\infty(\Omega)$ given by $w_{|T} = \partial_i v_{|T}$ for all $T \in \mathscr{T}$ is the i-th weak derivate of v. In particular, we have $v \in H^1(\Omega)$. In view of the characterization

$$H^1_0(\Omega) = \{v \in H^1(\Omega) \mid v_{|\partial\Omega} = 0\}$$

and the definition of $\mathbb{V}(\mathscr{T})$, we conclude that $v \in H^1_0(\Omega)$. □

Next, we construct a suitable basis of

$$S^{n,0}(\mathscr{T}) := \{v \in C(\overline{\Omega}) \mid \forall T \in \mathscr{T}\ v_{|T} \in \mathbb{P}_n(T)\},$$

which yields immediately one for $\mathbb{V}(\mathscr{T})$. We first consider the case $n = 1$ and, in view of the piecewise structure, start with the following result on $\mathbb{P}_1(T)$.

Lemma 3 (Local $\mathbb{P}_1$-Basis). *Let $T =$ conv hull $\{z_0, \dots, z_d\}$ be a d-simplex in $\mathbb{R}^d$. The barycentric coordinates $\lambda_0, \dots, \lambda_d : T \to \mathbb{R}$ on T defined by*

$$T \ni x = \sum_{i=0}^{d} \lambda_i(x) z_i, \quad \text{and} \quad \sum_{i=0}^{d} \lambda_i(x) = 1, \tag{45}$$

are a basis of $\mathbb{P}_1(T)$ such that

$$\lambda_i(z_j) = \delta_{ij} \qquad \text{for all } i, j \in \{1, \dots, d\}. \tag{46}$$

For each $p \in \mathbb{P}_1(T)$, there holds the representation formula

$$p = \sum_{i=0}^{d} p(z_i) \lambda_i. \tag{47}$$

Proof. We first check that the barycentric coordinates $\lambda_0, \dots, \lambda_d$ are well-defined. To this end, fix $x \in T$ for a moment and observe that (45) for $\lambda_i = \lambda_i(x)$ can be rewritten as

$$\begin{bmatrix} | & & | \\ z_0 & \cdots & z_d \\ | & & | \\ 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_d \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ 1 \end{bmatrix}.$$

If we choose F_T in Lemma 1 such that $F_T(0) = z_0$, $F_T(e_i) = F(z_i)$ for $i = 1, \dots, d$, we easily see that the above matrix has the same determinant as A_T, which is different from 0.

Consequently, the functions $\lambda_0, \dots, \lambda_d$ are well-defined and, varying x, we see that $\lambda_i \in \mathbb{P}_1(T)$ for $i = 0, \dots, d$. Property (46) is now readily verified and ensures that the $(d+1)$ functions $\lambda_0, \dots, \lambda_d$ are linearly independent. From the definition of $\mathbb{P}_1(T)$ it is immediate that $\dim \mathbb{P}_1(T) = d+1$, whence $\lambda_0, \dots, \lambda_d$ has to be a basis. Writing $p = \sum_{i=0}^{d} \alpha_i \lambda_i$ for $p \in \mathbb{P}_1(T)$, and using (46), yields (47) and finishes the proof. □

Property (46) means that $\lambda_0, \dots, \lambda_d$ is the basis in $\mathbb{P}_1(T)$ that is dual to the basis $\mathscr{N}_1(T) = \{N_1, \dots, N_d\}$ of $\mathbb{P}_1(T)^*$ given by $p \mapsto p(z_i)$ for $i = 0, \dots, d$. By the Riesz representation theorem in $L^2(T)$, we can associate a function $\lambda_i^* \in \mathbb{P}_1(T)$ to each functional N_i such that

$$\int_T \lambda_i \lambda_j^* = \delta_{ij} \qquad \text{for all } i,j \in \{1,\dots,d\}. \tag{48}$$

A simple computation using [25, Exercise 4.1.1] reveals that

$$\lambda_i^* = \frac{(1+d)^2}{|T|}\lambda_i - \frac{1+d}{|T|}\sum_{j\neq i}\lambda_j \qquad \text{for all } i \in \{1,\dots,d\}.$$

Since $\mathscr{N}_1(T)$ is a basis of $\mathbb{P}_1(T)^*$, the triple

$$(T, \mathbb{P}_1(T), \mathscr{N}_1(T))$$

is a finite element; for the definition of a finite element see, e.g., [16, Ch. 3]. The elements of $\mathscr{N}_1(T)$ are its *nodal variables* and $\lambda_0,\dots,\lambda_d$ its *nodal basis*.

Theorem 6 (Courant Basis). *A function $v \in S^{1,0}(\mathscr{T})$ is characterized by its values at the nodes $\mathscr{N}_1(\mathscr{T}) := \mathscr{V}_{\mathscr{T}}$. The functions ϕ_z, $z \in \mathscr{N}_1(\mathscr{T})$, defined by*

$$\phi_z \in S^{1,0}(\mathscr{T}) \qquad \text{and} \qquad \phi_z(y) = \delta_{yz} \qquad \text{for all } y \in \mathscr{N}_1(\mathscr{T})$$

are a basis of $S^{1,0}(\mathscr{T})$ such that, for every $v \in S^{1,0}(\mathscr{T})$,

$$v = \sum_{z\in\mathscr{N}_1(\mathscr{T})} v(z)\phi_z.$$

In particular, $\{\phi_z\}_{z\in\mathscr{N}_1(\mathscr{T})\cap\Omega}$ is a basis of $S^{1,0}(\Omega)\cap H_0^1(\Omega)$.

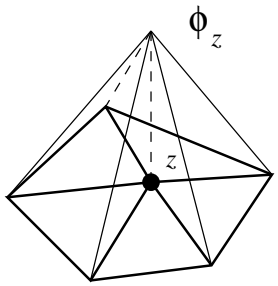

Fig. 1 Courant basis function ϕ_z for an interior vertex $z \in \mathscr{N}_1(\mathscr{T})$.

Proof. Let $T_1, T_2 \in \mathscr{T}$ be two distinct simplices such that $T_1 \cap T_2 \neq \emptyset$; then $S := T_1 \cap T_2$ is a k-subsimplex with $0 \le k < d$ because $\mathscr{T}$ is conforming. Let $w_i \in \mathbb{P}_1(T_i)$, for $i = 1,2$, be two affine functions with the same nodal values $w_1(z) = w_2(z)$ at all vertices $z \in S$. We assert that $w_1 = w_2$ on S. Since this is obvious for $k = 0$, we consider $k > 0$, recall that S is isomorphic to the reference simplex $\hat{T}_k$ in $\mathbb{R}^k$ and apply Lemma 3 to deduce $w_1 = w_2$ on S. This shows that any continuous piecewise

affine function $v \in S^{1,0}(\mathscr{T})$ can be built by pasting together local affine functions with the restriction of having the same nodal values, or equivalently to coincide at all vertices $z \in \mathscr{N}_1(\mathscr{T})$. Moreover, v is characterized by its nodal values $\{v(z)\}_{z\in\mathscr{N}_1(\mathscr{T})}$.

Therefore, the functions ϕ_z are well-defined for all $z \in \mathscr{N}_1(\mathscr{T})$. In addition, for all $v \in S^{1,0}(\mathscr{T})$ the function $\sum_{z\in\mathscr{N}_1(\mathscr{T})} v(z)\phi_z$ equals v at the nodes, whence they coincide everywhere and $S^{1,0}(\mathscr{T}) = \operatorname{span}\{\phi_z\}_{z\in\mathscr{N}_1(\mathscr{T})}$. Since $\{\phi_z\}_{z\in\mathscr{N}_1(\mathscr{T})}$ are linearly independent, they form a basis of $S^{1,0}(\mathscr{T})$.

Finally, to prove that $\{\phi_z\}_{z\in\mathscr{N}_1(\mathscr{T})\cap\Omega}$ is a basis of $S^{1,0}(\Omega)\cap H^1_0(\Omega)$ we observe that if $v \in S^{1,0}(\Omega)$ vanishes at the vertices of a side $S \in \mathscr{S}$ contained in $\partial\Omega$ then v vanishes in S, again as a consequence of Lemma 3. Therefore, $v \in S^{1,0}(\Omega)\cap H^1_0(\Omega)$ if and only if the nodal values $v(z) = 0$ for all $z \in \mathscr{N}_1(\mathscr{T})\cap\partial\Omega$. □

Remark 5 (Representation of Courant Basis). The proof of Theorem 6 shows that the global basis functions are given in terms of local basis functions. More precisely, if λ_z^T denotes the barycentric coordinate of $T \in \mathscr{T}$ associated with the vertex $z \in T$, there holds

$$\phi_z = \begin{cases} \lambda_z^T & \text{if } z \in T, \\ 0 & \text{otherwise} \end{cases}$$

for any node $z \in \mathscr{N}$.

We thus now have a basis of $\mathbb{V}(\mathscr{T}) = S^{1,0}(\mathscr{T})\cap H^1_0(\Omega)$ that can be implemented relatively easily. Its basis functions are locally supported and the corresponding matrix in (42) is sparse in the case of our model problem in Sect. 2.2.1; see Problem 17.

Remark 6 (Dual of Courant Basis). Let $v_z \in \mathbb{N}$ be the valence of z for each node $z \in \mathscr{N}_1(\mathscr{T})$, namely the number of elements $T \in \mathscr{T}$ containing z as a vertex. The *discontinuous* piecewise linear functions $\phi_z^* \in S^{1,-1}(\mathscr{T})$ defined by

$$\phi_z^* = \frac{1}{v_z}\sum_{T\ni z} (\lambda_z^T)^*\,\chi_T \qquad \text{for all } z \in \mathscr{N}_1(\mathscr{T}), \tag{49}$$

with χ_T being the characteristic function of T, are (global) dual functions to the Courant basis $\{\phi_z\}_{z\in\mathscr{N}}$ in that they satisfy

$$\int_\Omega \phi_z\phi_y^* = \delta_{yz} \qquad \text{for all } y,z \in \mathscr{N}_1(\mathscr{T}). \tag{50}$$

We briefly comment on the generalization to arbitray polynomial degree $n \in \mathbb{N}$. Given a d-simplex $T = \text{ conv hull } \{z_0,\dots,z_d\}$ and identifying nodal variables and nodes, we set

$$\mathscr{N}_n(T) := \left\{ z_\alpha = \sum_{i=0}^{d+1} \frac{\alpha_i}{n} z_i \mid \alpha \in \mathbb{N}_0^{d+1}, \sum_{i=0}^{d+1} \alpha_i = n \right\} \tag{51}$$

The number of elements in $\mathscr{N}_n(T)$ coincides with the number of coefficients of polynomial in $\mathbb{P}_n(T)$. This is necessary for the existence of the corresponding nodal

basis. The construction, see e.g. [16, Chapt. 3], reveals that also the location of the nodes plays some role. The latter implies also that restricting $\mathscr{N}_n(T)$ to a k-subsimplex and transforming to $\hat{T}_k$ yields $\mathscr{N}_n(\hat{T}_k)$. Consequently, the following theorem can be proven in the same way as Theorem 6.

Theorem 7 (Lagrange Basis). *A function $v \in S^{n,0}(\mathscr{T})$ is characterized by its values at the nodes $\mathscr{N}_n(\mathscr{T}) := \cup_{T\in\mathscr{T}} \mathscr{N}_n(T)$. The functions ϕ_z, $z \in \mathscr{N}_n(\mathscr{T})$, defined by*

$$\phi_z \in S^{n,0}(\mathscr{T}) \qquad \text{and} \qquad \phi_z(y) = \delta_{yz} \qquad \text{for all } y \in \mathscr{N}_n(\mathscr{T})$$

are a basis of $S^{n,0}(\mathscr{T})$ such that, for every $v \in S^{n,0}(\mathscr{T})$,

$$v = \sum_{z\in\mathscr{N}_n(\mathscr{T})} v(z)\phi_z.$$

In particular, $(\phi_z)_{z\in\mathscr{N}_n(\mathscr{T})\cap\Omega}$ is a basis of $S^{n,0}(\Omega)\cap H_0^1(\Omega)$.

Remark 7 (Dual of Lagrange Basis). The construction of local and global piecewise linear dual functions extends to any polynomial degree $n \geq 1$; see Problem 19 for $k = 2$. Consequently, there exist discontinuous functions $\phi_z^* \in S^{n,-1}(\mathscr{T})$ such that $\operatorname{supp}\phi_z^* = \operatorname{supp}\phi_z$ and

$$\int_\Omega \phi_z\phi_y^* = \delta_{yz} \qquad \text{for all } y,z \in \mathscr{N}_n(\mathscr{T}). \tag{52}$$

Remark 8 (Barycentric Coordinates). For linear finite elements the basis functions on a single element T are the barycentric coordinates on T. The barycentric coordinates play also an important role for higher degree. First we observe that any point $z_\alpha \in \mathscr{N}_n(T)$ is determined from the barycentric coordinates $\frac{1}{n}(\alpha_1,\ldots,\alpha_d)$. Secondly, using the $(d+1)$ barycentric coordinates as a local coordinate system on T is a rather convenient choice for the explicit construction of a local basis on T; compare with Problem 18 as well as [63, Sect. 1.4.1] for a more detailed description. This is one reason that local basis functions are defined in the finite element toolbox ALBERTA in terms of the barycentric coordinates [63, Sect. 3.5].

3.2.3 Looking Ahead

We close this section with a few comments about fundamental issues of finite elements that will be addressed later in this survey.

Mesh Construction. The formalism above relies on a conforming mesh $\mathscr{T}$. Its practical construction is a rather delicate matter, especially if it will be successively refined as part of an adaptive loop. We study mesh refinement by *bisection* in Chap. 4 in any dimension and assess the complexity of such process. This study involves basic geometry and graph theory as well as combinatorics.

Piecewise Polynomial Interpolation. As established in Theorem 5, the performance of the FEM hinges on the quality of piecewise polynomial approxima-

tion. We discuss this topic in Chap. 5, where we construct a *quasi interpolation* operator to approximate rough functions and introduce the concept of mesh optimality; Remark 7 will be crucial in this respect. We present an algorithm that builds quasi-optimal meshes by thresholding for a rather large class of rough functions. This hints at the potentials of FEM to approximate singular solutions.

A Posteriori Error Analysis. Thresholding assumes to have full access to the function in question, which is not realistic when dealing with PDE. The missing item is the design of a posteriori error estimators that extract the desired information from the discrete solution rather than the exact one. We present *residual estimators* in Chap. 6 and discuss their basic properties. They are instrumental.

Adaptivity. The fact that we learn about the approximation quality via a posteriori error estimators rather than directly from the function being approximated makes the study of AFEM quite different from classical approximation theory. This interplay between discrete and continuum will permeate the subsequent discussion in Chap. 7–Chap. 9.

In this survey, particularly when studying a posteriori error estimators and adaptivity, we assume that we have the *exact* Petrov-Galerkin solution U at hand. In doing this we ignore two important aspects of a practical finite element method: numerical integration and inexact solution of the resulting linear system. We close this chapter with two remarks concerning these issues.

Remark 9 (Numerical Integration). In contrast to the a priori error analysis of quadrature [25, Chapter 4.1], its treatment within an a posteriori context is a delicate matter, especially if one is not willing to assume regularity a priori and accept asymptotic results as the mesh size goes to zero. This seems to be largely open.

Remark 10 (Multilevel Solvers). For a hierarchy of quasi-uniform meshes, V-cycle multigrid and BPX-preconditioned conjugate gradient methods can approximate the Ritz-Galerkin solution U of our model problem (13) to a desired accuracy with a number of operations proportional to $\#\mathscr{T}$ [15, 16]. This, however, entails some restrictions on the coefficient matrix $\boldsymbol{A}$. Much less is known for graded meshes such as those generated by an adaptive method. For graded bisection meshes, we quote the results of Wu and Chen [77] for the V-cycle multigrid for $d = 2, n = 1$, and the recent results of Chen et al. [23] for multigrid methods and multilevel preconditioners for $d \geq 2, n \geq 1$: they both show linear complexity in terms of $\#\mathscr{T}$. The latter exploits the geometric properties of bisection grids explained in Chap. 4.

3.3 Problems

Problem 11. Prove *Cea's Lemma:* Let $\mathscr{B}\colon \mathbb{V} \times \mathbb{V}$ be a continuous and coercive form. Let u be the true solution and $U_N \in \mathbb{V}_N$ be the Galerkin solution. Then U_N is a quasi-best approximation to u in $\mathbb{V}_N$, i. e.,

$$\|u-U_N\|_{\mathbb{V}} \le \frac{\|\mathscr{B}\|}{c_{\mathscr{B}}} \min_{V\in\mathbb{V}_N} \|u-V\|_{\mathbb{V}}.$$

If, in addition, $\mathscr{B}$ is symmetric, then U_N is the best approximation to u in $\mathbb{V}_N$ with respect to the energy norm $|||\cdot|||_\Omega$, i. e.,

$$|||u-U_N|||_\Omega = \min_{V\in\mathbb{V}_N} |||u-V|||_\Omega$$

and the error in the $\mathbb{V}$-norm can be estimated by

$$\|u-U_N\|_{\mathbb{V}} \le \sqrt{\frac{\|\mathscr{B}\|}{c_{\mathscr{B}}}} \min_{V\in\mathbb{V}_N} \|u-V\|_{\mathbb{V}}.$$

Problem 12. Let $\{\mathbb{V}_N,\mathbb{W}_N\}_{N\in\mathbb{N}}$ be a sequence of nested subspaces of $\mathbb{V},\mathbb{W}$ of equal dimension N, i. e.,

$$\mathbb{V}_M \subset \mathbb{V}_N \qquad \text{and} \qquad \mathbb{W}_M \subset \mathbb{W}_N \qquad \text{for all } M \le N,$$

such that

$$\overline{\bigcup_{N\in\mathbb{N}} \mathbb{V}_N}^{\|\cdot\|_{\mathbb{V}}} = \mathbb{V} \qquad \text{and} \qquad \overline{\bigcup_{N\in\mathbb{N}} \mathbb{W}_N}^{\|\cdot\|_{\mathbb{W}}} = \mathbb{W}.$$

Suppose that, for every $f\in\mathbb{W}^*$, the sequence of discrete Petrov-Galerkin solutions $\{U_N\}_{\mathbb{N}}$ defined by

$$U_N \in \mathbb{V}_N: \qquad \mathscr{B}[U_N,W] = \langle f,W\rangle \qquad \text{for all } W\in\mathbb{W}_N$$

satisfies

$$\lim_{N\to\infty} \|u-U_N\|_{\mathbb{V}} = 0.$$

Show that there holds

$$\inf_{N\in\mathbb{N}} \inf_{V\in\mathbb{V}_N} \sup_{W\in\mathbb{W}_N} \frac{\mathscr{B}[V,W]}{\|V\|_{\mathbb{V}}\|W\|_{\mathbb{W}}} > 0.$$

Problem 13. Verify that the matrix K in (42) is symmetric positive definite for the model problem of Sect. 2.2.1, $\mathbb{V}_N=\mathbb{W}_N$ and $\phi_i=\psi_i$ for $i=1,\dots,N$, irrespective of the choice of $\phi_1,\dots,\phi_N$.

Problem 14. Prove Lemma 1. Start by expressing A_T and b_T in terms of the vertices of T.

Problem 15. Prove Lemma 2 for a not necessarily conforming triangulation.

Problem 16. Given a d-simplex $T=$ conv hull $\{z_0,\dots,z_d\}$ in $\mathbb{R}^d$, construct a basis $\bar\lambda_0,\dots,\bar\lambda_d$ of $\mathbb{P}_1(T)$ such that

$$\bar\lambda_i(\bar z_j) = \delta_{ij} \qquad \text{for all } i,j\in\{1,\dots,d\},$$

where $\bar{z}_j$ denotes the barycenter of the face opposite to the vertex z_j. Does this local basis also lead to a global one in $S^{1,0}(\mathscr{T})$?

Problem 17. Determine the support of a basis function ϕ_z, $z \in \mathscr{N}$, in Theorem 6. Show that, with this basis, the matrix K in (42) is sparse for the model problem in Sect. 2.2.1.

Problem 18. Express the nodal basis of $(T, \mathbb{P}_2(T), \mathscr{N}_2(T))$ in terms of barycentric coordinates.

Problem 19. Derive expressions for the dual functions of the quadratic local Lagrange basis of $P_2(T)$ for each element $T \in T$. Construct a global discontinuous dual basis $\phi_z^* \in S^{2,-1}(\mathscr{T})$ of the global Lagrange basis $\phi_z \in S^{2,0}(\mathscr{T})$ for all $z \in \mathscr{N}_2(\mathscr{T})$.

4 Mesh Refinement by Bisection

In this section we discuss refinement of a given initial triangulation consisting of d simplices using bisection, i. e., any selected simplex is divided into two subelements of same size. Refinement by bisection in 2d can be traced back to Sewell in the early 1970s [66]. In the mid of the 1980s Rivara introduced the longest edge bisection [61] and Mitchell formulated a recursive algorithm for the newest vertex bisection [49, 50]. In the beginning of the 1990s Bänsch was the first to present a generalization of the newest vertex bisection to 3d [10]. A similar approach was published by Liu and Joe [46] and later on by Arnold et al. [2]. A recursive variant of the algorithm by Bänsch was derived by Kossaczký [44]. He formulated the bisection rule for tetrahedra using a local order of their vertices and their element type. This concept is very convenient for implementation. In addition, it can be generalized to any space dimension which was done independently by Maubach [47] and Traxler [72].

Asking for conformity of locally refined meshes has the unalterable consequence that refinement propagates, i. e., besides the selected elements additional simplices have to be refined in order to maintain conformity. Although practical experience clearly suggests that local refinement stays local, the first theoretical foundation was given by Binev, Dahmen, and DeVore [13] in 2d in 2004. We summarize in this chapter the generalization to any space dimension by Stevenson [70].

4.1 Subdivision of a Single Simplex

We first describe how a single d-simplex is bisected, along with the concepts of vertex order and type. We then turn to recurrent bisection of a given initial element and the problem of shape regularity.

Bisection Rule based on Vertex Order and Type. We identify a simplex T with the set of its *ordered vertices* and its *type* t by

$$T = \{z_0, \ldots, z_d\}_t, \quad t \in \{0, \ldots, d-1\}.$$

Given such a d-simplex T we use the following bisection rule to split it in a unique fashion and to impose both vertex order and type to its children. The edge $\overline{z_0 z_d}$ connecting the first and last vertex of T is the *refinement edge* of T and its midpoint $\bar{z} = \frac{z_0+z_d}{2}$ becomes the new vertex. Connecting the new vertex $\bar{z}$ with the vertices of T other than z_0, z_d determines the common side $S = \{\bar{z}, z_1, \ldots, z_{d-1}\}$ shared by the two children T_1, T_2 of T. The *bisection rule* dictates the following vertex order and type for T_1, T_2

$$
\begin{aligned}
T_1 &:= \{z_0, \bar{z}, \underbrace{z_1, \dots, z_t}_{\rightarrow}, \underbrace{z_{t+1}, \dots, z_{d-1}}_{\rightarrow}\}_{(t+1) \bmod d}, \\
T_2 &:= \{z_d, \bar{z}, \underbrace{z_1, \dots, z_t}_{\rightarrow}, \underbrace{z_{d-1}, \dots, z_{t+1}}_{\leftarrow}\}_{(t+1) \bmod d},
\end{aligned}
\tag{53}
$$

with the convention that arrows point in the direction of increasing indices and $\{z_1, \dots, z_0\} = \emptyset$, $\{z_d, \dots, z_{d-1}\} = \emptyset$.

In 2d the bisection rule does not depend on the element type and we get for $T = \{z_0, z_1, z_2\}$ the two children

$$T_1 = \{z_0, \bar{z}, z_1\} \qquad \text{and} \qquad T_2 = \{z_2, \bar{z}, z_1\}.$$

As depicted in Fig. 2, the refinement edge of the two children is opposite to the

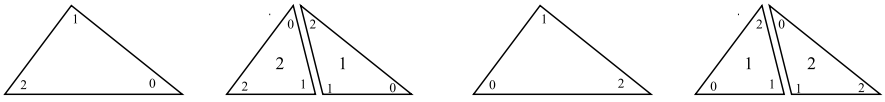

Fig. 2 Refinement of a single triangle $T = \{z_0, z_1, z_2\}$ and its reflected triangle $T_R = \{z_2, z_1, z_0\}$.

new vertex $\bar{z}$, whence this procedure coincides with *the newest vertex bisection* for $d = 2$. For $d \geq 3$ the bisection of an element does depend on its type, and, as we shall see below, this is important for preserving shape regularity. For instance, in 3d the children of $T = \{z_0, z_1, z_2, z_3\}_t$ are (see Fig. 3)

$$
\begin{aligned}
t = 0: &\qquad T_1 = \{z_0, \bar{z}, z_1, z_2\}_1 \quad \text{and} \quad T_2 = \{z_3, \bar{z}, z_2, z_1\}_1, \\
t = 1: &\qquad T_1 = \{z_0, \bar{z}, z_1, z_2\}_2 \quad \text{and} \quad T_2 = \{z_3, \bar{z}, z_1, z_2\}_2, \\
t = 2: &\qquad T_1 = \{z_0, \bar{z}, z_1, z_2\}_0 \quad \text{and} \quad T_2 = \{z_3, \bar{z}, z_1, z_2\}_0.
\end{aligned}
$$

Note that the vertex labeling of T_1 is type-independent, whereas that of T_2 is the

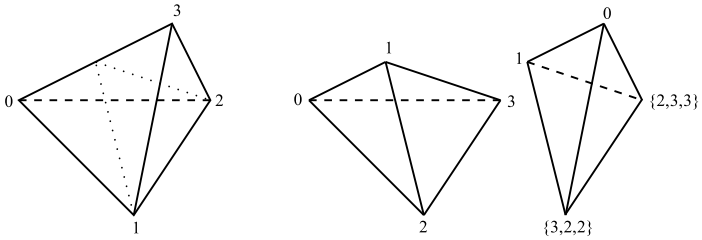

Fig. 3 Refinement of a single tetrahedron T of type t. The child T_1 in the middle has the same node ordering regardless of type. In contrast, for the child T_2 on the right a triple is appended to two nodes. The local vertex index is given for these nodes by the t-th component of the triple.

same for type 1 and 2. To account for this fact the vertices z_1 and z_2 of T are tagged

$(3,2,2)$ and $(2,3,3)$ in Fig. 3. The type of T then dictates which component of the triple is used to label the vertex.

Any different labeling of an element's vertices does not change its geometric shape but applying the above bisection rule it does change the shape and vertex order of its two children. This holds true for any relabeling except one. An element with this special relabeling of vertices is called *reflected element*. We state next its precise definition.

Definition 8 (Reflected Element). Given an element $T = \{z_0, \cdots, z_d\}_t$, the *reflected element* is given by

$$T_R := \{z_d, \underbrace{z_1, \ldots, z_t}_{\rightarrow}, \underbrace{z_{d-1}, \ldots, z_{t+1}}_{\leftarrow}, z_0\}_t.$$

Fig. 2 depicts for 2d $T = \{z_0, z_1, z_2\}$ and $T_R = \{z_2, z_1, z_0\}$. It shows that the children of T and T_R are the same. This property extends to $d \geq 3$; compare with Problem 21. Any other relabeling of vertices leads to different shapes of the children, in fact as many as $\frac{1}{2}(d+1)!$

Recurrent Bisection and Binary Tree. We next turn towards the recurrent bisection of a given *initial* simplex $T_0 = \{z_0, \ldots, z_d\}_{t_0}$. We let $\{T_1, T_2\} = \mathsf{BISECT}(T)$ be a function that implements the above bisection rule and outputs the two children of T. The input of BISECT can be T_0 or any element of the output from a previous application of BISECT.

This procedure of recurrent bisection of T_0 is associated with an *infinite binary tree* $\mathbb{F}(T_0)$. The nodes $T \in \mathbb{F}(T_0)$ correspond to simplices generated by repeated application of BISECT. The two successors of a node T are the two children $\{T_1, T_2\} = \mathsf{BISECT}(T)$. Note that $\mathbb{F}(T_0)$ strongly depends on the vertex order of T_0 and its type t_0. Once this is set for T_0 the associated binary tree is completely determined by the bisection rule. Recalling that the children of an element and its reflected element are the same this gives in total $\frac{d(d+1)!}{2}$ different binary trees that can be associated with T_0 by the bisection procedure.

The binary tree $\mathbb{F}(T_0)$ holds full information about the shape, ordering of vertices, type, etc. of any element T that can be generated by recurrent bisection of T_0. Important in this context is the distance of T to T_0 within $\mathbb{F}(T_0)$, which we call *generation*.

Definition 9 (Generation). The *generation* $g(T)$ of a node/element $T \in \mathbb{F}(T_0)$ is the number of its ancestors in the tree, or, equivalently, the number of bisections needed to create T from T_0.

Using the notion of generation, some information about T can uniquely be deduced from $g(T)$. For instance, for an element $T \in \mathbb{F}(T_0)$, its type is $(g(T) + t_0) \bmod d$, and, in view of the definition $h_T = |T|^{1/d}$, its size is

$$h_T = 2^{-g(T)/d} h_{T_0}. \tag{54}$$

Shape Regularity. We next analyse the shape coefficients of descendants of a given simplex T_0. A uniform bound on the shape coefficients σ_T for all $T \in \mathbb{F}(T_0)$ plays a crucial role in the interpolation estimates derived in Sect. 5.1. When turning towards shape regularity the dependence of the bisection rule on the element type for $d \geq 3$ becomes indispensable. The fact that the type t increases by 1 and the vertex ordering changes with t implies that after d recurrent bisections of T all its edges are bisected; compare with Problem 20.

We first consider a so-called *Kuhn-simplex*, i. e., a simplex with (ordered) vertices

$$z_0^\pi = 0, \quad z_i^\pi := \sum_{j=1}^{i} e_{\pi(j)} \quad \text{for all } i = 1, \ldots, d,$$

where π is a permutation of $\{1, \ldots, d\}$. Note, that $z_d^\pi = (1, \ldots, 1)^T$ for any permutation π. Therefore, the refinement edge $\overline{z_0^\pi, z_d^\pi}$ of any Kuhn-simplex is always the longest edge. If T_0 is a type 0 Kuhn-simplex, recurrent bisection always cuts the longest edge. This is the key property for obtaining uniform bound on the shape coefficients [47, 72].

Theorem 8 (Shape Regularity for a Kuhn-Simplex). *All 2^g descendants of generation g of a Kuhn-simplex $T_\pi = \{z_0^\pi, \ldots, z_d^\pi\}_0$ are mutually congruent with at most d different shapes. Moreover, the descendants of generation d are congruent to T_0 up to a scaling with factor $\frac{1}{2}$.*

In two dimensions, all descendants of a Kuhn-triangle belong to one similarity class; see Figure 4. Using an affine transformation we conclude from Theorem 8 shape

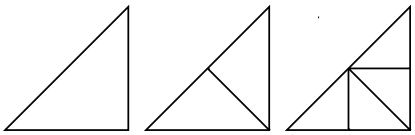

Fig. 4 Recurrent bisection of a Kuhn-triangle generates only one similarity class.

regularity for all descendants of an arbitrary simplex.

Corollary 5 (Shape Regularity). *Let $T_0 = \{z_0, \ldots, z_d\}_t$ be an arbitrary d-simplex. Then all descendants of T generated by bisection are shape regular, i. e.,*

$$\sup_{T \in \mathbb{F}(T_0)} \sigma_T = \sup_{T \in \mathbb{F}(T_0)} \frac{\overline{h}_T}{\underline{h}_T} \leq C(T_0) < \infty.$$

Proof. Consider first a simplex T_0 of type 0 and let $\hat{T}_0 := \{\hat{z}_0, \ldots, \hat{z}_d\}_0$ be the a Kuhn-simplex of type 0. From Lemma 1 we know that there exists a bijective affine mapping $F : \hat{T}_0 \to T_0$.

Recurrent refinement by bisection implies that for any $T \in \mathbb{F}(T_0)$ there exists a unique $\hat{T} \in \mathbb{F}(\hat{T}_0)$ such that $T = F(\hat{T})$. Since all descendants of $\hat{T}_0$ belong to at most d similarity classes, this implies that the minimal angle of all descendants of T_0 is uniformly bounded from below by a constant solely depending on the shape of T_0.

The same is valid for a simplex T_0 of type $t \in \{1,\dots,d-1\}$ because its 2^{d-t} descendants of generation $d-t$ are all of type 0. □

Note, that for a general d-simplex, the number of similarity classes for the descendants is larger than for a Kuhn d-simplex. This number is 4 for $d = 2$; compare Figures 4 and 5.

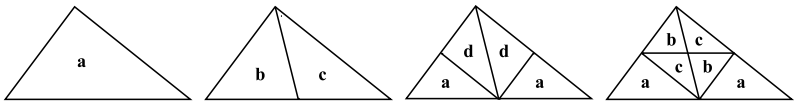

Fig. 5 Bisection produces at most 4 similarity classes for any initial triangle.

4.2 Mesh Refinement by Bisection

After discussing the refinement of a single simplex, we next turn to the refinement of a given initial conforming triangulation $\mathscr{T}_0$ by bisection. For recurrent refinement of a single element T_0 we are free to choose any order of its vertices and element type. The requirement to produce conforming refinements of $\mathscr{T}_0$ results in restrictions on local vertex order and type of the elements in $\mathscr{T}_0$. We first introduce the binary forest associated to triangulations generated by bisection and then elaborate on conformity and basic properties of the refined triangulations.

Master Forest and Forest. We recall that recurrent bisection of an element $T_0 \in \mathscr{T}_0$ is uniquely associated with an infinite binary tree $\mathbb{F}(T_0)$; see Sect. 4.1. In the same way we can identify all possible refinements of $\mathscr{T}_0$ with a *master forest* of binary trees.

Definition 10 (Forest and Refinement). Let $\mathscr{T}_0$ be an initial conforming triangulation. Then

$$\mathbb{F} = \mathbb{F}(\mathscr{T}_0) := \bigcup_{T_0 \in \mathscr{T}_0} \mathbb{F}(T_0).$$

is the associated *master forest* of binary trees. For a node $T \in \mathbb{F}$ so that $T \in \mathbb{F}(T_0)$ with $T_0 \in \mathscr{T}_0$, the generation $g(T)$ is the generation of T within $\mathbb{F}(T_0)$.

A subset $\mathscr{F} \subset \mathbb{F}$ is called *forest* iff

(1) $\mathscr{T}_0 \subset \mathscr{F}$;
(2) all nodes of $\mathscr{F} \setminus \mathscr{T}_0$ have a predecessor;
(3) all nodes of $\mathscr{F}$ have either two successors or none.

A forest $\mathscr{F}$ is called *finite*, if $\max_{T\in\mathscr{F}} g(T) < \infty$. The nodes with no successors are called *leaves* of $\mathscr{F}$.

Any finite forest $\mathscr{F}$ is uniquely associated with a triangulation $\mathscr{T} = \mathscr{T}(\mathscr{F})$ of Ω by defining $\mathscr{T}$ to be the set of all leaves in $\mathscr{F}$. Given two finite forests $\mathscr{F}, \mathscr{F}_* \in \mathbb{F}$ with associated triangulations $\mathscr{T}, \mathscr{T}_*$ we call $\mathscr{T}_*$ *refinement* of $\mathscr{T}$ iff $\mathscr{F} \subset \mathscr{F}_*$ and we denote this by $\mathscr{T} \leq \mathscr{T}_*$ or, equivalently, $\mathscr{T}_* \geq \mathscr{T}$.

Note that the definition of a finite forest $\mathscr{F}$ implies that the leaf nodes cover Ω, whence the associated triangulation $\mathscr{T}(\mathscr{F})$ is a partition of Ω. In general, this triangulation is not conforming and it is a priori not clear that conforming refinements of $\mathscr{T}_0$ exist.

Conforming Refinements. We next wonder about the properties of $\mathscr{T}_0$ that allow for conforming refinements. This brings us to the notion of *neighboring elements*.

Definition 11 (Neighboring Elements). Two elements $T_1, T_2 \in \mathscr{T}$ are called *neighboring elements* if they share a common side, namely a full $(d-1)$-simplex.

In 2d new vertices are always midpoints of edges. Generating the descendants of generation 2 for all elements of a given conforming triangulation $\mathscr{T}$ bisects all edges of $\mathscr{T}$ exactly once and all midpoints of the edges are vertices of the grandchildren. This implies conformity for $d = 2$. For $d > 2$ the situation is completely different.

Assume $d = 3$ and let $T_1, T_2 \in \mathscr{T}$ be two neighboring elements with common side $S = T_1 \cap T_2$. Denote by E_1, E_2 their respective refinement edges and assume that they belong to S. The 3d bisection of T_1 leads to a 2d bisection of S with E_1 being the refinement edge of S induced by T_1. The same holds true for T_2. If $E_1 \neq E_2$ the new edges in S created by refinement of T_1 and T_2 are not identical but do intersect. This leads to a non-conformity that cannot be cured by any further bisection of S. The same holds true for $d > 3$ upon replacing the newly created edge by the newly created $(d-2)$-simplex inside the common side. This yields for $d \geq 3$ a *necessary* condition for constructing a conforming refinement:

> *Whenever the refinement edges of two neighboring elements are both on the common side they have to coincide.*

For $d = 3$ this condition has been shown to also be *sufficient* for obtaining conforming refinements. It is also known that for any initial conforming triangulation $\mathscr{T}_0$ there exists a local labeling of the vertices satisfying this condition [10, 46, 2].

For $d > 3$ the above condition is not known to be sufficient. In addition, for proving the complexity result in Sect. 4.5 we need stronger assumptions on the distribution of refinement edges on $\mathscr{T}_0$. For the general case $d \geq 2$, we therefore formulate an assumption on the labeling of $\mathscr{T}_0$ given by Stevenson that ensures conformity of any *uniform* refinement of $\mathscr{T}_0$. This condition relies on the notion of reflected neighbor.

Definition 12 (Reflected Neighbors). Two neighboring elements $T = \{z_0, \ldots, z_d\}_t$ and $T' = \{z'_0, \ldots, z'_d\}_t$ are called *reflected neighbors* iff the ordered vertices of T or T_R coincide exactly with those of T' at all but one position.

We are now in the position to pose the assumptions on the initial triangulation $\mathscr{T}_0$.

Assumption 1 (Admissibility of the Initial Grid). *Let $\mathscr{T}_0$ be a conforming triangulation that fulfills*

(1) *all elements are of the same type $t \in \{0,\dots,d-1\}$;*
(2) *all neighboring elements $T = \{z_0,\dots,z_d\}_t$ and $T' = \{z'_0,\dots,z'_d\}_t$ with common side S are* matching neighbors *in the following sense: if $\overline{z_0 z_d} \subset S$ or $\overline{z'_0 z'_d} \subset S$ then T and T' are reflected neighbors; otherwise the pair of neighboring children of T and T' are reflected neighbors.*

For instance, the set of the $d!$ Kuhn-simplices of type 0 is a conforming triangulation of the unit cube in $\mathbb{R}^d$ satisfying Assumption 1; see Problem 22. We also refer to Fig. 6 and Problem 23 to explore this concept for $d = 2$.

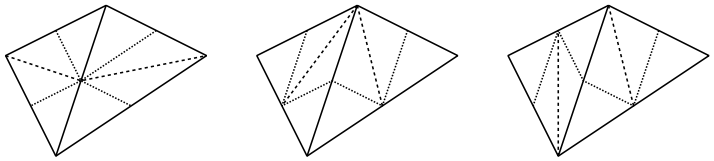

Fig. 6 Matching neighbors in 2d and their descendants of generation 1 and 2. The elements in the left and middle picture are reflected neighbors. The elements in the rightmost picture are not reflected neighbours, but the pair of their neighboring children are.

Uniform Refinements. We next state the following important implication of this structural assumption on $\mathscr{T}_0$. The proof is a combination of [72, Sect. 4] and [70, Theorem 4.3].

Theorem 9 (Uniform Refinement). *Let $\mathscr{T}_0$ be a conforming triangulation and for $g \in \mathbb{N}_0$ denote by*

$$\mathscr{T}_g := \{T \in \mathbb{F}(\mathscr{T}_0) \mid g(T) = g\}$$

the uniform refinement of $\mathscr{T}_0$ with elements of generation exactly g.

If Assumption 1 is satisfied, then $\mathscr{T}_g$ is conforming for any $g \in \mathbb{N}_0$. In addition, if all elements in $\mathscr{T}_0$ are of the same type, then condition (2) is necessary for $\mathscr{T}_g$ to be conforming for all g.

To interpret Theorem 9 we introduce the following useful definition.

Definition 13 (Compatible Bisection). We say that two elements $T, T' \in \mathbb{F}$ are *compatibly divisible* if they have the same refinement edge. If all elements sharing an edge are compatibly divisible, then they form a *bisection patch*.

Using this notion, Theorem 9 states that two elements $T, T' \in \mathbb{F}$ of the same generation sharing a common edge are either compatibly divisible, or the refinement of T does not affect T' and vice versa. In the latter case any common edge is neither the refinement edge of T nor of T'.

Let $d = 2$ and $T = \{z_0, z_1, z_2\}_t$ and $T' = \{z'_0, z'_1, z'_2\}_t$ be neighboring elements with common side S. If $\overline{z_0 z_d} = \overline{z'_0 z'_d}$ then T and T' are compatibly divisible and thus form a bisection patch: they can be refined without affecting any other element. If $z_1, z'_1 \in S$, then the pair of neighboring children of T and T' are compatibly divisible and thus form a bisection patch; compare with Fig. 6 and Problem 23.

Remark 11 (Discussion of Assumption 1). Assumption 1, given by Stevenson [70], is weaker than the condition required by Maubach [47] and Traxler [72]: they asked that all neighboring elements are reflected neighbors. It is an important open question whether for any conforming triangulation $\mathscr{T}_0$ there exists a suitable labeling of the element's vertices such that Assumption 1 is satisfied.

For dimension $d = 2$ such a result has been shown by Mitchell [49, Theorem 2.9] as well as Binev et al. [13, Lemma 2.1]. Both proofs are based on graph theory and they are not constructive. It can be shown that the problem of finding a suitable labeling of the vertices, the so-called *perfect matching*, is NP-complete.

For dimension $d > 2$ this is an open problem. In 3d Kossaczký has constructed a conforming refinement of any given coarse grid into an initial grid $\mathscr{T}_0$ that satisfies Assumption 1. This construction has been generalized by Stevenson to any space dimension. [70, Appendix A].

As mentioned above, the conditions of Bänsch [10], Liu and Joe [46], and Arnold et al. [2] on the initial tetrahedral mesh can be satisfied for any given conforming triangulation. But then it can only be shown that uniform refinements $\mathscr{T}_g$ with $g \bmod d = 0$ are conforming [2, 10, 46]. The property that any uniform refinement $\mathscr{T}_g$ for $g \in \mathbb{N}_0$ is conforming is the key tool for the complexity proof in Sect. 4.5.

We next define the class of *conforming* refinements of $\mathscr{T}_0$ to be

$$\mathbb{T} = \{\mathscr{T} = \mathscr{T}(\mathscr{F}) \mid \mathscr{F} \subset \mathbb{F} \text{ is finite and } \mathscr{T}(\mathscr{F}) \text{ is conforming}\}.$$

Then Theorem 9 has two direct consequences.

(a) The class $\mathbb{T}$ contains an infinite number of conforming refinements of $\mathscr{T}_0$.
(b) There exists a function $\mathsf{REFINE}(\mathscr{T}, \mathscr{M})$ that, given a conforming triangulation $\mathscr{T} \in \mathbb{T}$ and a subset $\mathscr{M} \subset \mathscr{T}$ of marked elements, bisects all simplices in $\mathscr{M}$ at least once, and outputs the smallest conforming refinement $\mathscr{T}_* \in \mathbb{T}$ of $\mathscr{T}$ with $\mathscr{T}_* \cap \mathscr{M} = \emptyset$.

Before constructing such function REFINE we analyze some basic properties of triangulations.

4.3 Basic Properties of Triangulations

In this section we analyze basic properties of refinement by bisection, namely uniform shape regularity, convergence of mesh-size functions, and the cardinality of an

overlay of two triangulations. The results can be easily derived using the structure of the master forest $\mathbb{F}$.

Uniform Shape Regularity. A direct consequence of Corollary 5 is that refinement by bisection only produces elements T with shape coefficient σ_T uniformly bounded by a constant solely depending on $\mathscr{T}_0$; recall Definition 6.

Lemma 4. *All elements in $\mathbb{F}$ are uniformly shape regular, i. e.,*

$$\sup_{T\in\mathbb{F}} \sigma_T = \sup_{T\in\mathbb{F}} \frac{\overline{h}_T}{\underline{h}_T} \leq C(\mathscr{T}_0) < \infty.$$

For any conforming mesh $\mathscr{T} \in \mathbb{T}$, the discrete neighborhood of $T \in \mathscr{T}$ is given by

$$N_{\mathscr{T}}(T) := \{T' \in \mathscr{T} |\ T' \cap T \neq \emptyset\}.$$

Lemma 4 implies that the cardinality of this patch is bounded uniformly and the measure of all its elements is comparable

$$\max_{T\in\mathscr{T}} \#N_{\mathscr{T}}(T) \leq C(\mathscr{T}_0), \qquad \max_{T'\in N_{\mathscr{T}}(T)} \frac{|T|}{|T'|} \leq C(\mathscr{T}_0), \tag{55}$$

with $C(\mathscr{T}_0)$ only depending on $\mathscr{T}_0$. This is usually called *local quasi-uniformity*.

Convergence of Mesh-Size Functions. Let $\{\mathscr{T}_k\}_{k\geq 0} \subset \mathbb{T}$ be any sequence of nested refinements, i. e., $\mathscr{T}_k \leq \mathscr{T}_{k+1}$ for $k \geq 0$. This sequence is accompanied by the sequence of mesh-size functions $\{h_k\}_{k\geq 0}$, defined as $h_k \in L^\infty(\Omega)$ with

$$h_{k|T} = h_T = |T|^{1/d} \qquad \text{for all } T \in \mathscr{T}_k.$$

If the sequence is produced by uniform refinement then we easily obtain from (54)

$$\lim_{k\to\infty} \|h_k\|_{L^\infty(\Omega)} = 0. \tag{56}$$

However, this may not hold when the sequence $\mathscr{T}_k$ is generated adaptively, i. e., we allow for local refinement. Therefore we have to generalize it appropriately. For a first generalization of (56), we observe that the skeleton $\Gamma_k := \bigcup\{\partial T \cap \Omega : T \in \mathscr{T}_k\}$ of $\mathscr{T}_k$ has d-dimensional Lebesgue measure zero. We may thus interpret h_k as a piecewise constant function in $L^\infty(\Omega)$. Moreover, the limiting skeleton $\Gamma_\infty := \cup_{k\geq 0}\Gamma_k$ has also d-dimensional Lebesgue measure zero. Since, for every $x \in \Omega \setminus \Gamma_\infty$, the sequence $h_k(x)$ is monotonically decreasing and bounded from below by 0,

$$h_\infty(x) := \lim_{k\to\infty} h_k(x) \tag{57}$$

is well-defined for $x \in \Omega \setminus \Gamma_\infty$ and defines a function in $L^\infty(\Omega)$. As the next lemma shows, the pointwise convergence in (57) holds actually in $L^\infty(\Omega)$. Another generalization of (56), where the limit function is 0, will be provided in Corollary 10 in Chap. 7.

Lemma 5 (Uniform Convergence of Mesh-Size Functions). *For any sequence $\{\mathscr{T}_k\}_{k\geq 0} \subset \mathbb{T}$ of nested refinements the corresponding sequence $\{h_k\}_{k\geq 0}$ of mesh-size functions converges uniformly in $\Omega \setminus \Gamma_\infty$ to h_∞, i. e.,*

$$\lim_{k\to\infty} \|h_k - h_\infty\|_{L^\infty(\Omega)} = 0.$$

Proof. [1] Denote by $\mathscr{F}_k = \mathscr{F}(\mathscr{T}_k)$ the corresponding forest of $\mathscr{T}_k$. From $\mathscr{T}_k \leq \mathscr{T}_{k+1}$ we conclude $\mathscr{F}_k \subset \mathscr{F}_{k+1}$ and thus the forest

$$\mathscr{F}_\infty := \bigcup_{k\geq 0} \mathscr{F}_k$$

is well defined. Note that in general $\mathscr{F}_\infty$ is infinite.
[2] For arbitrary $\varepsilon > 0$, let $g = g(\varepsilon) \in \mathbb{N}$ be the smallest number such that

$$g \geq \log(\varepsilon^d/M)/\log(\tfrac{1}{2})$$

with $M = \max\{|T| \mid T \in \mathscr{T}_0\}$. Obviously, $\hat{\mathscr{F}} := \{T \in \mathscr{F}_\infty \mid g(T) \leq g\}$ is a finite forest and $\mathscr{T}(\hat{\mathscr{F}})$ is a triangulation of Ω. Since $\hat{\mathscr{F}} \subset \mathscr{F}_\infty$ is finite there exists $k = k(\varepsilon) \geq 0$ with $\hat{\mathscr{F}} \subset \mathscr{F}_k$.
[3] Let $T \in \mathscr{T}_k$ be any leaf node of $\mathscr{F}_k$ and let $T \in \mathscr{F}(T_0)$ for some $T_0 \in \mathscr{T}_0$. To estimate $h_k - h_\infty$ on T, we distinguish the following two cases:

Case 1: $g(T) < g$. This implies that T is a leaf node of $\mathscr{F}_\infty$ and thus $h_{k|T} = h_{\infty|T}$ or, equivalently, $(h_k - h_\infty)_{|T} = 0$.

Case 2: $g(T) \geq g$. Hence, T is generated by at least g bisections of T_0. By (54), the monotonicity of the mesh-size functions, and the choice of g, we obtain

$$0 \leq (h_k - h_\infty)_{|T} \leq h_{k|T} = h_T \leq 2^{-g(T)/d} h_{T_0} \leq 2^{-g/d} M^{1/d} \leq \varepsilon.$$

Combining the two cases we end up with $0 \leq (h_k - h_\infty)_{|T} \leq \varepsilon$ for all $T \in \mathscr{T}_k$. Since ε is arbitrary and $0 \leq h_\ell - h_\infty \leq h_k - h_\infty$ in Ω for all $\ell \geq k$, this finishes the proof. □

Overlay of Triangulations. Let $\mathscr{T}_1, \mathscr{T}_2 \in \mathbb{T}$ be conforming triangulations with corresponding finite forests $\mathscr{F}_1$ and $\mathscr{F}_2$. Then $\mathscr{F}_1 \cup \mathscr{F}_2$ is also a finite forest and we call the unique triangulation

$$\mathscr{T}_1 \oplus \mathscr{T}_2 := \mathscr{T}(\mathscr{F}_1 \cup \mathscr{F}_2) \tag{58}$$

the *overlay of $\mathscr{T}_1$ and $\mathscr{T}_2$*. The name overlay is motivated by printing 2d triangulations $\mathscr{T}_1$ and $\mathscr{T}_2$ at the same position on two slides. The overlay is then the triangulation that can be seen when putting one slide on top of the other. It turns out that the overlay is the smallest conforming triangulation with $\mathscr{T}_1, \mathscr{T}_2 \leq \mathscr{T}_1 \oplus \mathscr{T}_2$ and its cardinality can be estimated by the ones of $\mathscr{T}_1$ and $\mathscr{T}_2$.

Lemma 6 (Overlay of Meshes). *For $\mathscr{T}_1, \mathscr{T}_2 \in \mathbb{T}$ the overlay $\mathscr{T} := \mathscr{T}_1 \oplus \mathscr{T}_2$ is the smallest common refinement of $\mathscr{T}_1$ and $\mathscr{T}_2$ and satisfies*

$$\#\mathscr{T} \leq \#\mathscr{T}_1 + \#\mathscr{T}_2 - \#\mathscr{T}_0.$$

Proof. Argue by contradiction and assume that $\mathscr{T}$ contains a non-conforming vertex z. That is, there exist $T_1, T_2 \in \mathscr{T}$ such that z is a vertex of T_1 and $z \in T_2$ is not a vertex of T_2. Without loss of generality let $T_1 \in \mathscr{T}_1$. Since $\mathscr{T}_1$ is conforming, there exists a $T' \in \mathscr{T}_1$, $T' \subset T_2$ such that z is a vertex of T'. Hence, T' is a descendant of T_2 in $\mathscr{F}_1$ and thus T_2 cannot be a leaf node of $\mathscr{F}(\mathscr{T})$, i.e., $T_2 \notin \mathscr{T}$, a contradiction. Since the overlay only contains elements from $\mathscr{T}_1$ or $\mathscr{T}_2$ and is conforming, it is the smallest conforming refinement.

For $T \in \mathscr{T}_0$ and $i = 1,2$ we denote by $\mathscr{F}_i(T) \subset \mathscr{F}(\mathscr{T})$ the binary trees with root T corresponding to $\mathscr{T}_i$ and let $\mathscr{T}_i(T)$ be the triangulation given by the leaf nodes of $\mathscr{F}_i(T)$. Since $\mathscr{T}(T) \subset \mathscr{T}_1(T) \cup \mathscr{T}_2(T)$, we infer that $\#\mathscr{T}(T) \leq \#\mathscr{T}_1(T) + \#\mathscr{T}_2(T)$. We now show that $\#\mathscr{T}(T) \leq \#\mathscr{T}_1(T) + \#\mathscr{T}_2(T) - 1$ by distinguishing two cases.

Case 1: $\mathscr{T}_1(T) \cap \mathscr{T}_2(T) \neq \emptyset$. Then there exists $T' \in \mathscr{T}_1(T) \cap \mathscr{T}_2(T)$, and so $T' \in \mathscr{T}(T)$. By counting T' only once in $\#(\mathscr{T}_1(T) \cup \mathscr{T}_2(T))$ we get $\#\mathscr{T}(T) \leq \#\mathscr{T}_1(T) + \#\mathscr{T}_2(T) - 1$.

Case 2: $\mathscr{T}_1(T) \cap \mathscr{T}_2(T) = \emptyset$. Then there exists $T' \in \mathscr{T}_1(T)$ (resp., $T' \in \mathscr{T}_2(T)$) so that $T' \notin \mathscr{T}(T)$, for otherwise $T' \in \mathscr{T}_2(T)$ (resp., $T' \in \mathscr{T}_1(T)$), thereby contradicting the assumption. We obtain again $\#\mathscr{T}(T) \leq \#\mathscr{T}_1(T) + \#\mathscr{T}_2(T) - 1$.

Finally, since $\mathscr{T}_i = \bigcup_{T \in \mathscr{T}_0} \mathscr{T}_i(T)$, the assertion follows by adding over the elements in $\mathscr{T}_0$. □

4.4 Refinement Algorithms

We discuss two refinement algorithms based on the bisection rule introduced in Sect. 4.1. Given a conforming triangulation $\mathscr{T}$ and a subset of marked elements $\mathscr{M}$ both variants output the smallest conforming refinement $\mathscr{T}_*$ of $\mathscr{T}$ such that all elements of $\mathscr{M}$ are bisected, i. e., $\mathscr{T}_* \cap \mathscr{M} = \emptyset$.

Iterative Refinement. The basic idea is to first bisect all marked elements in $\mathscr{T}$ leading to a non-conforming grid $\mathscr{T}_*$. In order to restore conformity, we identify all elements $T \in \mathscr{T}$ containing a so-called *irregular (or hanging) node* $z \in T$, namely a vertex $z \in \mathscr{V}_{\mathscr{T}_*}$ which is not a vertex of T. These elements are then also scheduled for refinement. This procedure has to be iterated until all irregular nodes are removed and this step is called *completion*. The core of iterative refinement is a routine that bisects all marked elements in a possibly non-conforming triangulation:

REFINE_MARKED($\mathscr{T}, \mathscr{M}$)
for all $T \in \mathscr{M}$ do
  $\{T_0, T_1\} =$ BISECT(T);
  $\mathscr{T} := \mathscr{T} \setminus \{T\} \cup \{T_0, T_1\}$;
end for
return($\mathscr{T}$)

The refinement of a given conforming grid $\mathscr{T}$ with a subset of marked elements $\mathscr{M}$ into a new conforming refinement is then executed by

```
REFINE(𝒯, ℳ)
while ℳ ≠ ∅ do
    𝒯 := REFINE_MARKED(𝒯, ℳ);
    ℳ := {T ∈ 𝒯 | T contains an irregular node};
end while
return(𝒯)
```

We let $\mathscr{T}_*$ be the output of REFINE_MARKED$(\mathscr{T},\mathscr{M})$ on its first call. Since non-conforming situations can only be cured by refining all elements containing an irregular node, the above algorithm outputs the smallest conforming refinement of $\mathscr{T}_*$ if the while-loop terminates. We let g be the maximal generation of any element in $\mathscr{T}_*$. By Theorem 9 the uniform refinement $\mathscr{T}_g$ is conforming, and by construction it satisfies $\mathscr{T}_* \leq \mathscr{T}_g$. Since REFINE$(\mathscr{T},\mathscr{M})$ only refines elements to remove non-conforming situations, any intermediate grid $\mathscr{T}$ produced by REFINE_MARKED$(\mathscr{T},\mathscr{M})$ satisfies $\mathscr{T} \leq \mathscr{T}_g$ and this implies that the while loop in the above algorithm terminates.

We point out that this algorithm works without any assumption on the ordering of vertices in $\mathscr{T}_0$ in 2d and with the less restrictive assumptions in [2, 10, 46] in 3d. This follows from the fact that $\mathscr{T}_g$ with $g \bmod d = 0$ is conforming and thus one can choose a suitable $\mathscr{T}_g \geq \mathscr{T}_*$; compare with Remark 11.

The above implementation of iterative refinement is not efficient since there are too many loops in the completion step. We observe that the bisection of a single element T enforces the bisection of all elements at its refinement edge. Some of these elements may also be marked for refinement and will directly be refined. Other elements have to be refined in the completion step. The algorithm can be speeded up by directly scheduling those elements for refinement.

This motivates the simultaneous bisection of all elements meeting at the refinement edge. This variant is discussed next.

Recursive Refinement. Let $\mathscr{T}$ be a given conforming grid and let $T \in \mathscr{T}$ be an element with refinement edge E. We define the *refinement patch of* T to be

$$R(\mathscr{T};T) := \{T' \in \mathscr{T} \mid T' \in \mathscr{T} \text{ with } E \subset T'\}.$$

As mentioned above, a bisection of T enforces a refinement of all elements in $R(\mathscr{T};T)$ for regaining conformity. We could avoid non-conforming situations by a simultaneous refinement of the whole refinement patch. This is only possible if all elements in $R(\mathscr{T};T)$ are compatibly divisible, i. e., E is the refinement edge of all $T' \in R(\mathscr{T};T)$ and $R(\mathscr{T};T)$ is a bisection patch. This is called the *atomic refinement operation* and is depicted in Fig. 7 for $d = 2$ (top) and $d = 3$ (bottom).

If there are elements in $R(\mathscr{T};T)$ that are not compatibly divisible with T, the basic idea is to recursively refine these elements first. This builds up the new refinement patch around E that in the end allows for the atomic refinement operation.

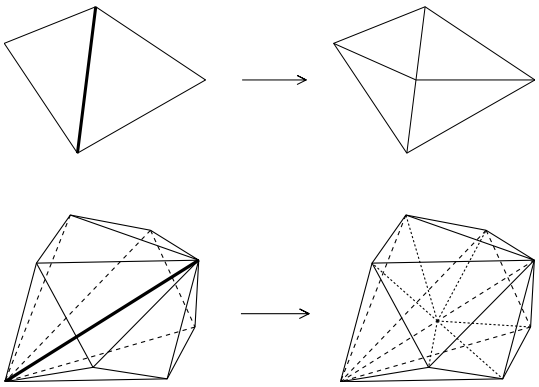

Fig. 7 Atomic refinement operation in 2d (top) and 3d (bottom): The common edge is the refinement edge for all elements.

In 2d, there is one neighbor sharing the refinement edge E in case E is interior. Either this neighbor is compatibly divisible, or the neighboring child is compatibly divisible after bisection of the neighbor. If E lies on the boundary, instead, bisection can be executed directly. Fig. 8 illustrates a situation that requires recursion.

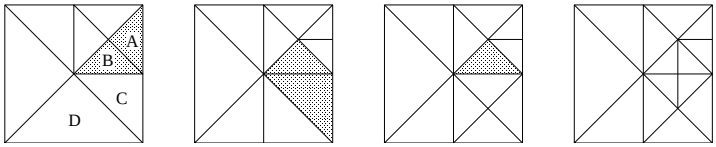

Fig. 8 Recursive refinement in 2d: Triangles A and B are initially marked for refinement.

In higher dimension there is in general a whole bunch of elements in $R(\mathscr{T};T)$. Since $R(\mathscr{T};T)\subset N_{\mathscr{T}}(T)$, (55) implies that the cardinality of $R(\mathscr{T};T)$ is uniformly bounded depending only on $\mathscr{T}_0$. Here it may happen that several elements have to be refined before we can perform the atomic refinement operation. It may also happen that an element inside the refinement patch has to be refined several times but the number of bisections is bounded by $d-1$; see Lemma 7 below. This lemma also allows for an elegant formulation of the recursive algorithm.

Lemma 7. *Let $\mathscr{T}_0$ be a conforming triangulation satisfying Assumption 1 and let $\mathscr{T}\in\mathbb{T}$ be a conforming refinement.*

Then any $T\in\mathscr{T}$ is of locally highest generation in $R(\mathscr{T};T)$, i. e.,

$$g(T)=\max\{g(T')\mid T'\in R(\mathscr{T};T)\}$$

and $T'\in R(\mathscr{T};T)$ is compatibly divisible with T if and only if $g(T')=g(T)$.

Furthermore, $\min\{g(T')\mid T'\in R(\mathscr{T};T)\}\geq g(T)-d+1$ *and a non-compatibly divisible neighboring element of T has generation $g(T)-1$.*

Proof. Denote by E the refinement edge of T and set $g := g(T)$. The uniform refinement $\mathscr{T}_{g+1}$ of $\mathscr{T}_0$ contains the midpoint $\bar{z}$ of E as a vertex and is a conforming refinement of $\mathscr{T}$. For any $T' \in R(\mathscr{T};T)$ the new vertex $\bar{z}$ is an irregular node on T', whence $T' \notin \mathscr{T}_{g+1}$. Since $\mathscr{T}_{g+1}$ is a conforming refinement of $\mathscr{T}$ we know that descendants of T' belong to $\mathscr{T}_{g+1}$ and thus $g(T') \le g$ for all $T' \in R(\mathscr{T};T)$.

If T' is compatibly divisible with T, then $\bar{z}$ is the new vertex of the two children, which belong to $\mathscr{T}_{g+1}$; hence, $g(T') = g$. If T' is not compatibly divisible with T, then $\bar{z}$ is the new vertex of descendants of one child of T', whence $g(T') < g$.

The refinement rule (53) implies that after d recurrent bisections all edges of the original simplex are bisected (see Problem 20). Consequently, any $T' \in R(\mathscr{T};T)$ has descendants of generation at most $g(T') + d$ that have $\bar{z}$ as a vertex and belong to $\mathscr{T}_{g+1}$. This yields $g(T') \ge g - d + 1$.

If T' is a non-compatibly divisible neighbor of T, then the refinement rule (53) implies that the refinement edge of one child T'' of T' is contained in the common side of T and T'. Since $\mathscr{T}_{g+1}$ is conforming this implies that T and T'' are compatibly divisible, and thus $g(T') = g - 1$. □

For $\mathscr{T} \in \mathbb{T}$ the recursive refinement of a single element $T \in \mathscr{T}$ now reads:

REFINE_RECURSIVE$(\mathscr{T}, T)$
do forever
  get refinement patch $R(\mathscr{T}, T)$;
  access $T' \in R(\mathscr{T}, T)$ with $g(T') = \min\{g(T'') \mid T'' \in R(\mathscr{T};T)\}$;
  if $g(T') < g(T)$ then
    $\mathscr{T} :=$ REFINE_RECURSIVE$(\mathscr{T}, T')$;
  else
    break;
  end if
end do
get refinement patch $R(\mathscr{T}, T)$;
for all $T' \in R(\mathscr{T}, T)$ do
  $\{T'_0, T'_1\} =$ BISECT(T');
  $\mathscr{T} := \mathscr{T} \setminus \{T'\} \cup \{T'_0, T'_1\}$;
end for
return($\mathscr{T}$)

Lemma 7 implies that only elements T' with $g(T') < g(T)$ are not compatibly divisible with T. Hence, recursion is only applied to elements with $g(T') < g(T)$ and thus the maximal depth of recursion is $g(T)$ and recursion terminates. Recursive refinement of an element T' may affect other elements of $R(\mathscr{T};T)$ with same generation $g(T')$. When the do-loop aborts, all elements in the refinement patch $R(\mathscr{T};T)$ are compatibly divisible, and the atomic refinement operation is executed in the for-loop: all elements $T' \in R(\mathscr{T};T)$ are refined, removed from $R(\mathscr{T};T)$, and replaced by the respective children sharing the refinement edge of T. Those children are all of generation $g(T') + 1 \le g(T)$. Since $\#R(\mathscr{T};T) \le C(\mathscr{T}_0)$, all elements in $R(\mathscr{T};T)$ are of the same generation $g(T)$ after a finite number of iterations. Observe that, except

for T, elements in $R(\mathscr{T};T)$ are only refined to avoid a non-conforming situation. This in summary yields the following result.

Lemma 8 (Recursive Refinement). *Let $\mathscr{T}_0$ be a conforming triangulation satisfying Assumption 1 and let $\mathscr{T} \in \mathbb{T}$ be any conforming refinement.*

Then, for any $T \in \mathscr{T}$ a call of REFINE_RECURSIVE($\mathscr{T}$,T) *terminates and outputs the smallest conforming refinement $\mathscr{T}_*$ of $\mathscr{T}$ where T is bisected. All newly created elements $T' \in \mathscr{T}_* \setminus \mathscr{T}$ satisfy $g(T') \leq g(T)+1$.*

Remark 12. Assumption 1 is a sufficient condition for recursion to terminate but it is not necessary. Such a characterization of recursive bisection is not known. Obviously, termination of the recursion for all elements of $\mathscr{T}_0$ is necessary. Practical experience shows that in 2d this is also sufficient, whereas this is not true in 3d.

We next formulate the algorithm for refining a given conforming grid $\mathscr{T}$ with marked elements $\mathscr{M}$ into a new conforming triangulation:

REFINE($\mathscr{T},\mathscr{M}$)
for all $T \in \mathscr{M} \cap \mathscr{T}$ do
  $\mathscr{T}$:= REFINE_RECURSIVE($\mathscr{T},T$);
end
return($\mathscr{T}$)

Let T be an element of the input set of marked elements $\mathscr{M}$. Then it may happen that there is an element $T_* \in \mathscr{M}$ scheduled prior to T for refinement and so that the refinement of T_* enforces the refinement of T, for instance $T \in R(\mathscr{T};T_*)$. In the bisection step T is replaced by its two children in $\mathscr{T}$ and thus $T \notin \mathscr{M} \cap \mathscr{T}$. This avoids to refine T twice. In addition, since REFINE_RECURSIVE($\mathscr{T},T$) outputs the smallest refinement such that T is bisected, REFINE($\mathscr{T},\mathscr{M}$) outputs the smallest conforming refinement $\mathscr{T}_*$ of $\mathscr{T}$ with $\mathscr{T}_* \cap \mathscr{M} = \emptyset$.

Remark 13 (Iterative vs Recursive Refinement). The iterative and recursive variant of REFINE produce the same output mesh whenever they both terminate. Proposition 2 in Sect. 4.5 makes use of the recursive refinement algorithm but is also valid for the iterative variant.

We concluded successful termination of both variants from the fact that the output grid $\mathscr{T}_*$ satisfies $\mathscr{T}_* \leq \mathscr{T}_g$ with g sufficiently large. Therefore, the used arguments do not imply that local refinement stays local. This property is an implication of Theorem 10 below.

On a first glance, the iterative variant seems to be easier to implement. But it turns out that handling non-conforming situations can become rather knotty, especially for $d \geq 3$. The implementation of the recursive variant avoids any non-conforming situation by performing the atomic refinement operation, which, as a consequence, simplifies the implementation drastically. The drawback of recursive refinement are stronger assumptions on the distribution of refinement edges on the initial grid.

4.5 Complexity of Refinement by Bisection

In this section we analyze the cardinality of conforming triangulations produced by adaptive iterations of the form (4). Assuming that a function $\mathsf{REFINE}(\mathscr{T},\mathscr{M})$ outputs the smallest conforming refinement of $\mathscr{T}$ with all elements in $\mathscr{M}$ bisected, we study a sequence of conforming refinements $\mathscr{T}_0 \leq \mathscr{T}_1 \leq \cdots \leq \mathscr{T}_k \leq \ldots$ generated by an iteration of the form

for $k \geq 0$ do
  determine a suitable subset $\mathscr{M}_k \subset \mathscr{T}_k$;
  $\mathscr{T}_{k+1} := \mathsf{REFINE}(\mathscr{T}_k, \mathscr{M}_k)$;
end

The main result is the following theorem.

Theorem 10 (Complexity of Refinement by Bisection). *Let $\mathscr{T}_0$ be a conforming triangulation satisfying Assumption 1.*

Then there exists a constant $\Lambda > 0$ solely depending on $\mathscr{T}_0$, such that for any $K \geq 0$ the conforming triangulation $\mathscr{T}_K$ produced by the above iteration verifies

$$\#\mathscr{T}_K - \#\mathscr{T}_0 \leq \Lambda \sum_{k=0}^{K-1} \#\mathscr{M}_k.$$

The proof of this theorem is split into several steps. Before embarking on it we want to remark that an estimate of the form

$$\#\mathscr{T}_{k+1} - \#\mathscr{T}_k \leq \Lambda \, \#\mathscr{M}_k \tag{59}$$

would imply Theorem 10 by summing up over $k = 0, \ldots, K-1$. But such a bound does not hold for refinement by bisection. To see this, consider the initial grid $\mathscr{T}_0$

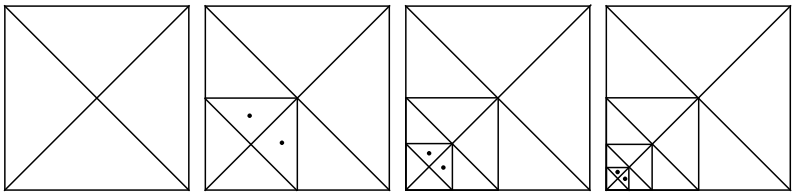

Fig. 9 An example showing that the depth of recursion in is only bounded by the generation of the selected element. Initial triangulation in the leftmost picture and grids $\mathscr{T}_K$ for $K = 2, 4, 6$. Recursion has depth K for the refinement of the elements marked with bullets.

depicted as the leftmost picture in Fig. 9. For all elements the boundary edge is selected as refinement edge and this choice satisfies Assumption 1. Pick up any even $K \in \mathbb{N}$ and let

$$\mathscr{M}_k := \{T \in \mathscr{T}_k \mid 0 \in T\} \quad \text{for } k = 0, \dots, K-1$$

and

$$\mathscr{M}_K := \{T \in \mathscr{T}_K \mid g(T) = K \text{ and } 0 \notin T\}.$$

In Fig. 9 we show the grids $\mathscr{T}_K$ for $K = 2,4,6$ and the two elements in $\mathscr{M}_K$ are indicated by a bullet. For $k \leq K$ we only refine marked elements, whence $\#\mathscr{T}_{k+1} - \#\mathscr{T}_k = \#\mathscr{M}_k = 2$ for $k = 0 \dots, K-1$. When refining $\mathscr{T}_K$ into $\mathscr{T}_{K+1}$ we have to recursively refine elements of generation $K-1, K-2, \dots, 0$ for both elements in $\mathscr{M}_K$. From this it is easy to verify that $\#\mathscr{T}_{K+1} - \#\mathscr{T}_K = 4K+2$. Since $\#\mathscr{M}_K = 2$ and K is an arbitrary even number it is obvious that (59) can not hold. On the other hand,

$$\sum_{k=0}^{K} \#\mathscr{T}_{k+1} - \#\mathscr{T}_k = (4K+2) + \sum_{k=0}^{K-1} 2 = 6K+2 \leq 3\,(2K+2) = 3 \sum_{k=0}^{K} \#\mathscr{M}_k.$$

This shows that Theorem 10 holds true for this example.

The proof of the theorem can be heuristically motivated as follows. Consider the set $\mathscr{M} := \bigcup_{k=0}^{K-1} \mathscr{M}_k$ used to generate the sequence $\mathscr{T}_0 \leq \mathscr{T}_1 \leq \cdots \leq \mathscr{T}_K =: \mathscr{T}$. Suppose that each element $T_* \in \mathscr{M}$ is assigned a fixed amount C_1 of money to spend on refined elements in $\mathscr{T}$, i. e., on $T \in \mathscr{T} \setminus \mathscr{T}_0$. Assume further that $\lambda(T, T_*)$ is the portion of money spent by T_* on T. Then it must hold

$$\sum_{T \in \mathscr{T} \setminus \mathscr{T}_0} \lambda(T, T_*) \leq C_1 \qquad \text{for all } T_* \in \mathscr{M}. \tag{60a}$$

In addition, we suppose that the investment of all elements in $\mathscr{M}$ is fair in the sense that each $T \in \mathscr{T} \setminus \mathscr{T}_0$ gets at least a fixed amount C_2, whence

$$\sum_{T_* \in \mathscr{M}} \lambda(T, T_*) \geq C_2 \qquad \text{for all } T \in \mathscr{T} \setminus \mathscr{T}_0. \tag{60b}$$

Therefore, summing up (60b) and using the upper bound (60a) we readily obtain

$$C_2(\#\mathscr{T} - \#\mathscr{T}_0) \leq \sum_{T \in \mathscr{T} \setminus \mathscr{T}_0} \sum_{T_* \in \mathscr{M}} \lambda(T, T_*) = \sum_{T_* \in \mathscr{M}} \sum_{T \in \mathscr{T} \setminus \mathscr{T}_0} \lambda(T, T_*) \leq C_1 \#\mathscr{M},$$

which proves the theorem for $\mathscr{T}$ and $\mathscr{M}$. In the remainder of this section we design such an allocation function $\lambda \colon \mathscr{T} \times \mathscr{M} \to \mathbb{R}^+$ in several steps and prove that recurrent refinement by bisection yields (60) provided $\mathscr{T}_0$ satisfies Assumption 1.

In view of (54), measure and diameter of an element are related to its generation:

$$D_1 2^{-g(T)} \leq |T| \quad \text{and} \quad \operatorname{diam}(T) \leq D_2 2^{-g(T)/d} \qquad \text{for all } T \in \mathbb{F}, \tag{61}$$

with $D_1 = \min\{|T_0| \mid T_0 \in \mathscr{T}_0\}$ and $D_2 \approx \max\{|T_0| \mid T_0 \in \mathscr{T}_0\}$. The constant hidden in $\approx$ solely depends on the shape regularity of $\mathbb{F}$ (and thus on $\mathscr{T}_0$).

Suppose now that T' is generated by REFINE_RECURSIVE($\mathscr{T}$,T). The constant D_2 enables us to relate the distance of T' to T with its generation $g(T')$, where

$$\operatorname{dist}(T,T') = \inf_{x\in T, x'\in T'} \left|x-x'\right|.$$

Proposition 2 (Distance and Generation). *Let $\mathscr{T}\in\mathbb{T}$, $T\in\mathscr{T}$ and assume that T' is created by* REFINE_RECURSIVE*($\mathscr{T}$,T). Then there holds*

$$\operatorname{dist}(T,T') \le D_2\, 2^{1/d} \sum_{g=g(T')}^{g(T)} 2^{-g/d} < D_2 \frac{2^{1/d}}{1-2^{-1/d}}\, 2^{-g(T')/d}.$$

Proof. We prove $\operatorname{dist}(T,T') \le D_2\, 2^{1/d} \sum_{g=g(T')}^{g(T)} 2^{-g/d}$ by induction over the generation of T. The rightmost inequality is a direct consequence of the geometric sum.
1 If $g(T)=0$, then the refinement patch $R(\mathscr{T};T)$ is compatibly divisible thanks to Lemma 7. Consequently REFINE_RECURSIVE($\mathscr{T}$,T) only creates elements T' with $\operatorname{dist}(T,T')=0$ and the assertion follows trivially.
2 Let now $g(T)>0$ and assume that the assertion holds for any $T''\in\mathscr{T}$ with $0\le g(T'')<g(T)$. We only need to consider $\operatorname{dist}(T,T')>0$, whence T' is created by a recursive call REFINE_RECURSIVE($\mathscr{T}$,T'') for an element $T''\in R(\mathscr{T};T)$ that is not compatibly divisible with T; thus $g(T'')<g(T)$ by Lemma 7. The induction hypothesis yields

$$\operatorname{dist}(T'',T') \le D_2\, 2^{1/d} \sum_{g=g(T')}^{g(T'')} 2^{-g/d}.$$

Since $T''\in R(\mathscr{T},T)$, and so T'' contains the refinement edge of T, we realize that $\operatorname{dist}(T'',T)=0$. Combining the last estimate with (61), we deduce

$$\begin{aligned}\operatorname{dist}(T,T') &\le \operatorname{dist}(T'',T') + \operatorname{diam}(T'') \le D_2\, 2^{1/d} \sum_{g=g(T')}^{g(T'')} 2^{-g/d} + D_2\, 2^{-g(T'')/d} \\ &= D_2\, 2^{1/d} \sum_{g=g(T')}^{g(T'')+1} 2^{-g/d} \le D_2\, 2^{1/d} \sum_{g=g(T')}^{g(T)} 2^{-g/d},\end{aligned}$$

where we have used $g(T'')<g(T)$ in the last step. This finishes the proof. □

We next construct the allocation function λ. The construction is based on two sequences $\{a(\ell)\}_{\ell=-1}^{\infty}, \{b(\ell)\}_{\ell=0}^{\infty} \subset \mathbb{R}^+$ of positive numbers satisfying

$$\sum_{\ell\ge -1} a(\ell) = A < \infty, \qquad \sum_{\ell\ge 0} 2^{-\ell/d}\, b(\ell) = B < \infty, \qquad \inf_{\ell\ge 1} b(\ell)\, a(\ell) = c_* > 0,$$

and $b(0)\ge 1$. Valid instances are $a(\ell)=(\ell+2)^{-2}$ and $b(\ell)=2^{\ell/(d+1)}$.

With these settings we are prepared to define $\lambda\colon \mathscr{T}\times\mathscr{M}\to\mathbb{R}^+$ by

$$\lambda(T,T_*) := \begin{cases} a(g(T_*)-g(T)), & \operatorname{dist}(T,T_*) < D_3 B 2^{-g(T)/d} \text{ and } g(T)\le g(T_*)+1 \\ 0, & \text{else,}\end{cases}$$

where $D_3 := D_2\big(1+2^{1/d}(1-2^{-1/d})^{-1}\big)$. Therefore, the investment of money by $T_* \in \mathscr{M}$ is restricted to cells T that are sufficiently close and are of generation $g(T) \le g(T_*)+1$. Only elements of such generation can be created during refinement of T_* according to Lemma 7.

The following lemma shows that the total amount of money spend by this allocation function per marked element is bounded.

Lemma 9 (Upper Bound). *There exists a constant $C_1 > 0$ only depending on $\mathscr{T}_0$ such that λ satisfies* (60a)*, i. e.,*

$$\sum_{T\in\mathscr{T}\setminus\mathscr{T}_0} \lambda(T,T_*) \le C_1 \qquad \text{for all } T_* \in \mathscr{M}.$$

Proof. [1] Given $T_* \in \mathscr{M}$ we set $g_* = g(T_*)$ and we let $0 \le g \le g_*+1$ be a generation of interest in the definition of λ. We claim that for such g the cardinality of the set

$$\mathscr{T}(T_*,g) = \{T \in \mathscr{T} \mid \operatorname{dist}(T,T_*) < D_3 B\, 2^{-g/d} \text{ and } g(T) = g\}$$

is uniformly bounded, i. e., $\#\mathscr{T}(T_*,g) \le C$ with C solely depending on D_1, D_2, D_3, B.

From (61) we learn that $\operatorname{diam}(T_*) \le D_2 2^{-g_*/d} \le 2D_2 2^{-(g_*+1)/d} \le 2D_2 2^{-g/d}$ as well as $\operatorname{diam}(T) \le D_2 2^{-g/d}$ for any $T \in \mathscr{T}(T_*,g)$. Hence, all elements of the set $\mathscr{T}(T_*,g)$ lie inside a ball centered at the barycenter of T_* with radius $(D_3B+3D_2)2^{-g/d}$. Again relying on (61) we thus conclude

$$\#\mathscr{T}(T_*,g) D_1 2^{-g} \le \sum_{T\in\mathscr{T}(T_*,g)} |T| \le c(d)(D_3B+3D_2)^d 2^{-g},$$

whence $\#\mathscr{T}(T_*,g) \le c(d)\, D_1^{-1}\,(D_3B+3D_2)^d =: C$.

[2] Accounting only for non-zero contributions $\lambda(T,T_*)$ we deduce

$$\sum_{T\in\mathscr{T}\setminus\mathscr{T}_0} \lambda(T,T_*) = \sum_{g=0}^{g_*+1} \sum_{T\in\mathscr{T}(T_*,g)} a(g_*-g) \le C \sum_{\ell=-1}^{\infty} a(\ell) = CA =: C_1,$$

which is the desired upper bound. □

The definition of λ also implies that each refined element receives a fixed amount of money.

Lemma 10 (Lower Bound). *There exists a constant $C_2 > 0$ only depending on $\mathscr{T}_0$ such that λ satisfies* (60b)*, i. e.,*

$$\sum_{T_*\in\mathscr{M}} \lambda(T,T_*) \ge C_2 \qquad \text{for all } T \in \mathscr{T}\setminus\mathscr{T}_0.$$

Proof. [1] Fix an arbitrary $T_0 \in \mathscr{T}\setminus\mathscr{T}_0$. Then there is an iteration count $1 \le k_0 \le K$ such that $T_0 \in \mathscr{T}_{k_0}$ and $T_0 \notin \mathscr{T}_{k_0-1}$. Therefore there exists an $T_1 \in \mathscr{M}_{k_0-1} \subset \mathscr{M}$ such that T_0 is generated during $\mathsf{REFINE\_RECURSIVE}(\mathscr{T}_{k_0-1}, T_1)$. Iterating this

process we construct a sequence $\{T_j\}_{j=1}^J \subset \mathscr{M}$ with corresponding iteration counts $\{k_j\}_{j=1}^J$ such that T_j is created by REFINE_RECURSIVE$(\mathscr{T}_{k_j-1}, T_{j+1})$. The sequence is finite since the iteration counts are strictly decreasing and thus $k_J = 0$ for some $J > 0$, or equivalently $T_J \in \mathscr{T}_0$.

Since T_j is created during refinement of T_{j+1} we infer from Lemma 8 that

$$g(T_{j+1}) \geq g(T_j) - 1.$$

Accordingly, $g(T_{j+1})$ can decrease the previous value of $g(T_j)$ at most by 1. Since $g(T_J) = 0$ there exists a smallest value s such that $g(T_s) = g(T_0) - 1$. Note that for $j = 1, \dots, s$ we have $\lambda(T_0, T_j) > 0$ if $\operatorname{dist}(T_0, T_j) \leq D_3 B g^{-g(T_0)/d}$.

[2] We next estimate the distance $\operatorname{dist}(T_0, T_j)$. For $1 \leq j \leq s$ and $\ell \geq 0$ we define the set

$$\mathscr{T}(T_0, \ell, j) := \{T \in \{T_0, \dots, T_{j-1}\} \mid g(T) = g(T_0) + \ell\}$$

and denote by $m(\ell, j)$ its cardinality. The triangle inequality combined with an induction argument yields

$$\begin{aligned} \operatorname{dist}(T_0, T_j) &\leq \operatorname{dist}(T_0, T_1) + \operatorname{diam}(T_1) + \operatorname{dist}(T_1, T_j) \\ &\leq \sum_{i=1}^{j} \operatorname{dist}(T_{i-1}, T_i) + \sum_{i=1}^{j-1} \operatorname{diam}(T_i). \end{aligned}$$

We apply Proposition 2 for the terms of the first sum and (61) for the terms of the second sum to obtain

$$\begin{aligned} \operatorname{dist}(T_0, T_j) &< D_2 \frac{2^{1/d}}{1 - 2^{-1/d}} \sum_{i=1}^{j} 2^{-g(T_{i-1})/d} + D_2 \sum_{i=1}^{j-1} 2^{-g(T_i)/d} \\ &= D_2 \left(1 + \frac{2^{1/d}}{1 - 2^{-1/d}} \right) \sum_{i=0}^{j-1} 2^{-g(T_i)/d} \\ &= D_3 \sum_{\ell=0}^{\infty} m(\ell, j)\, 2^{-(g(T_0)+\ell)/d} \\ &= D_3 2^{-g(T_0)/d} \sum_{\ell=0}^{\infty} m(\ell, j)\, 2^{-\ell/d}. \end{aligned}$$

For establishing the lower bound we distinguish two cases depending on the size of $m(\ell, s)$. This is done next.

[3] *Case 1:* $m(\ell, s) \leq b(\ell)$ for all $\ell \geq 0$. From this we conclude

$$\operatorname{dist}(T_0, T_s) < D_3 2^{-g(T_0)/d} \sum_{\ell=0}^{\infty} b(\ell)\, 2^{-\ell/d} = D_3 B\, 2^{-g(T_0)/d}$$

and the definition of λ then readily implies

$$\sum_{T_*\in\mathscr{M}} \lambda(T_0,T_*) \geq \lambda(T_0,T_s) = a(g(T_s)-g(T_0)) = a(-1) > 0.$$

4 *Case 2:* There exists $\ell \geq 0$ such that $m(\ell,s) > b(\ell)$. For each of these ℓ's there exists a smallest $j = j(\ell)$ such that $m(\ell, j(\ell)) > b(\ell)$. We let ℓ^* be the index ℓ that gives rise to the smallest $j(\ell)$, and set $j^* = j(\ell^*)$. Consequently

$$m(\ell, j^*-1) \leq b(\ell) \quad \text{for all } \ell \geq 0 \qquad \text{and} \qquad m(\ell^*, j^*) > b(\ell^*).$$

As in Case 1 we see $\operatorname{dist}(T_0,T_i) < D_3 B 2^{-g(T_0)/d}$ for all $i \leq j^*-1$, or equivalently

$$\operatorname{dist}(T_0,T_i) < D_3 B 2^{-g(T_0)/d} \qquad \text{for all } T_i \in \mathscr{T}(T_0,\ell,j^*).$$

We next show that the elements in $\mathscr{T}(T_0,\ell^*,j^*)$ spend enough money on T_0. We first consider $\ell^* = 0$ and note that $T_0 \in \mathscr{T}(T_0,0,j^*)$. Since $m(0,j^*) > b(0) \geq 1$ we discover $j^* \geq 2$. Hence, there is an $T_i \in \mathscr{T}(T_0,0,j^*) \cap \mathscr{M}$, which yields the estimate

$$\sum_{T_*\in\mathscr{M}} \lambda(T_0,T_*) \geq \lambda(T_0,T_i) = a(g(T_i)-g(T_0)) = a(0) > 0.$$

For $\ell^* > 0$ we see that $T_0 \notin \mathscr{T}(T_0,\ell^*,j^*)$, whence $\mathscr{T}(T_0,\ell^*,j^*) \subset \mathscr{M}$. In addition, $\lambda(T_0,T_i) = a(\ell^*)$ for all $T_i \in \mathscr{T}(T_0,\ell^*,j^*)$. From this we conclude

$$\begin{aligned}\sum_{T_*\in\mathscr{M}} \lambda(T_0,T_*) &\geq \sum_{T_*\in\mathscr{T}(T_0,\ell^*,j^*)} \lambda(T_0,T_*) = m(\ell^*,j^*)\, a(\ell^*) \\ &> b(\ell^*)\, a(\ell^*) \geq \inf_{\ell\geq 1} b(\ell)\, a(\ell) = c_* > 0.\end{aligned}$$

5 In summary we have proved the assertion since for any $T_0 \in \mathscr{T} \setminus \mathscr{T}_0$

$$\sum_{T_*\in\mathscr{M}} \lambda(T_0,T_*) \geq \min\{a(-1),a(0),c_*\} =: C_2 > 0. \qquad \square$$

Lemmas 9 and 10 show that the allocation function λ satisfies (60), which implies Theorem 10.

Remark 14 (Several Bisections). In practice, one often likes to bisect selected elements several times, for instance each marked element is scheduled for $b \geq 1$ bisections. This can be done by assigning the number $b(T) = b$ of bisections that have to be executed for each marked element T. If T is bisected then we assign $(b(T)-1)$ as the number of pending bisections to its children and the set of marked elements is $\mathscr{M} := \{T \in \mathscr{T} \mid b(T) > 0\}$.

To show the complexity estimate when REFINE performs $b > 1$ bisections, the set $\mathscr{M}_k$ is to be understood as a sequence of *single* bisections recorded in sets $\{\mathscr{M}_k(j)\}_{j=1}^b$, which belong to intermediate triangulations between $\mathscr{T}_k$ and $\mathscr{T}_{k+1}$ with $\#\mathscr{M}_k(j) \leq 2^{j-1}\#\mathscr{M}_k$, $j = 1,\ldots,b$. Then we also obtain Theorem 10 because

$$\sum_{j=1}^{b} \#\mathscr{M}_k(j) \leq \sum_{j=1}^{b} 2^{j-1} \#\mathscr{M}_k = (2^b - 1)\#\mathscr{M}_k.$$

Remark 15 (Optimal Constant). Trying to trace the value of the constant Λ one realizes that Λ becomes rather large since it depends via C_1 and C_2 on the constants A, B, c_*. Experiments suggest that $\Lambda \approx 14$ in 2d and $\Lambda \approx 180$ in 3d when $\mathscr{T}_0$ is the initial triangulation of the d-dimension cube $(0,1)^d$ build from the $d!$ Kuhn-simplices of type 0. According to Problem 22, $\mathscr{T}_0$ satisfies Assumption 1.

There is an interesting connection to a result by Atalay and Mount that can be formulated as follows [3]: there exists a constant $C_d \leq 3^d d!$ such that the smallest conforming refinement $\mathscr{T}_* \in \mathbb{T}$ of any non-conforming refinement $\mathscr{T}$ of $\mathscr{T}_0$ satisfies

$$\#\mathscr{T}_* \leq C_d \#\mathscr{T}.$$

For 2d the optimal constant is shown to be $C_2 = 14$ and the constant $C_3 = 162$ for $d = 3$ is quite close to the constant observed in experiments. The agreement between theory and experiments for 2d is quite exiting, but nevertheless the estimate by Atalay and Mount cannot be used to show Theorem 10.

4.6 Problems

Problem 20. Show that after d recurrent bisections of a simplex T all edges of T are bisected exactly once. To this end, let first $T = \{z_0, \ldots, z_d\}_0$ be of type 0 and show by induction that any sub-simplex T' of T with generation $t = g < d$ has the structure

$$T' = \left\{ z_{k_0}, \bar{z}_t, \bar{z}_{t-1}, \ldots, \bar{z}_1, z_{k_1}, z_{k_2}, \ldots, z_{k_{d-t}} \right\}_t,$$

where $\bar{z}_i$ are the new vertices of the bisection step i, $i = 1, \ldots, t$, and $k_0, \ldots, k_{d-t}$ are consecutive natural numbers, for instance $0, 1, 2, \ldots, d-1$ or $d, d-1, \ldots, 1$ for $t = 1$. Then generalize the claim to a simplex T of type $t \in \{0, \ldots, d-1\}$.

Problem 21. Show that the output of $\mathsf{BISECT}(T)$ and $\mathsf{BISECT}(T_R)$ is the same, i. e., the children of T and its reflected element T_R are identical.

Problem 22. Show that the set of the $d!$ Kuhn-simplices of type 0 is a conforming triangulation of the unit cube $(0,1)^d \subset \mathbb{R}^d$ satisfying Assumption 1.

Problem 23. Let $d = 2$ and $T = \{z_0, z_1, z_2\}_t, T' = \{z'_0, z'_1, z'_2\}_t$ be neighboring elements with common side $S = T \cap T'$. Show that

(a) T and T' are reflected neighbors if and only if $\overline{z_0 z_2} = \overline{z'_0 z'_2}$ or $z_1 = z'_1$.
(b) If T and T' are reflected neighbors, then so are their neighboring children.
(c) If $z_1 = z'_2$ and $z_2 = z'_1$, then T and T' are not reflected neighbors but their neighboring children are.
(d) If $S = \overline{z_0 z_2} = \overline{z'_0 z'_2}$ or $z_1, z'_1 \in S$, then T and T' are matching neighbors.

5 Piecewise Polynomial Approximation

The numerical solution of a boundary value problem may be seen as a special approximation problem where the target function is not given explicitly but implicitly. Theorem 5 shows that the error of a Petrov-Galerkin solution of a stable discretization is dictated by the best approximation from the discrete space. In this chapter we investigate approximation properties of continuous piecewise polynomials, the standard discretization for the model problem in Sect. 2.2.1. We do not strive for completeness but rather want to provide some background and motivation for the successive chapters. To this end, we depart from classical finite element approximation and end up with a result on nonlinear or adaptive approximation.

For more information about nonlinear and constructive approximation, we refer to the survey [28] and the book [29].

5.1 Quasi-Interpolation

We start with a brief discussion on piecewise polynomial interpolation of rough functions, namely those without point values as we expect H^1-functions to be. This leads to the concept of quasi-interpolation and to a priori error estimates for the standard discretization of our model problem in Sect. 2.2.1.

Using the Lagrange basis $\{\phi_z\}_{z\in\mathcal{N}_n(\mathscr{T})} \subset S^{n,0}(\mathscr{T})$ from Theorem 7 we have for any $v \in S^{n,0}(\mathscr{T})$ the representation $v = \sum_{z\in\mathcal{N}_n(\mathscr{T})} v(z)\phi_z$. This may suggest to use for given v the *Lagrange interpolant*

$$I_{\mathscr{T}} v(x) := \sum_{z\in\mathcal{N}_n(\mathscr{T})} v(z)\phi_z(x). \tag{62}$$

However, this operator requires that point values of v are well-defined. If $v \in W_p^s(\Omega)$, this entails the condition $\mathrm{sob}(W_p^s) > 0$, which in turn requires regularity beyond the trial space $H_0^1(\Omega)$ when $d \geq 2$.

Quasi-interpolants, like those in Clément [26] or Scott-Zhang [65], replace $v(z)$ in (62) by a suitable local average and so are well-defined also for rough functions, e.g. from $H_0^1(\Omega)$. For any conforming refinement $\mathscr{T} \geq \mathscr{T}_0$ of $\mathscr{T}_0$, the averaging process extends beyond nodes and so brings up the discrete neigborhood

$$N_{\mathscr{T}}(T) := \{T' \in \mathscr{T} \mid T' \cap T \neq \emptyset\}$$

for each element $T \in \mathscr{T}$ along with the uniform properties (55), namely,

$$\max_{T\in\mathscr{T}} \#N_{\mathscr{T}}(T) \leq C(\mathscr{T}_0), \qquad \max_{T'\in N_{\mathscr{T}}(T)} \frac{|T|}{|T'|} \leq C(\mathscr{T}_0),$$

where $C(\mathscr{T}_0)$ depends only on the shape coefficient of $\mathscr{T}_0$. We shall make use of the following estimate of the local interpolation error; see [16, 65].

Proposition 3 (Local Error Estimate for Quasi-Interpolant). *Let s be the regularity index with $0 \le s \le n+1$, and $1 \le p \le \infty$ be the integrability index.*
(a) There exists an operator $I_{\mathscr{T}} : L^1(\Omega) \to S^{n,0}(\mathscr{T})$ such that for all $T \in \mathscr{T}$ we have

$$\|D^t(v - I_{\mathscr{T}} v)\|_{L^q(T)} \lesssim h_T^{\operatorname{sob}(W_p^s) - \operatorname{sob}(W_q^t)} \|D^s v\|_{L^p(N_{\mathscr{T}}(T))} \tag{63}$$

where $0 \le t \le s$, $1 \le q \le \infty$ are such that $\operatorname{sob}(W_p^s) > \operatorname{sob}(W_q^t)$. The hidden constant depends on the shape coefficient of $\mathscr{T}_0$ and d.
(b) There exists an operator $I_{\mathscr{T}} : W_1^1(\Omega) \to S^{n,0}(\mathscr{T})$ satisfying (63) for $s \ge 1$ and, in addition, if $v \in W_1^1(\Omega)$ has a vanishing trace on $\partial\Omega$, then so does $I_{\mathscr{T}} v$.

Both operators are invariant in $S^{n,0}(\mathscr{T})$, namely $I_{\mathscr{T}} V = V$ for all $V \in S^{n,0}(\mathscr{T})$.

Proof. We sketch the proof; see [16, 65] for details. Recall that $\{\phi_z\}_{z \in \mathscr{N}_n(\mathscr{T})}$ is the global Lagrange basis of $S^{n,0}(\mathscr{T})$ and $\{\phi_z^*\}_{z \in \mathscr{N}_n(\mathscr{T})}$ is the global dual basis and, according to Remark 7, $\operatorname{supp} \phi_z^* = \operatorname{supp} \phi_z$ for all $z \in \mathscr{N}_n(\mathscr{T})$. We thus define $I_{\mathscr{T}} : L^1(\Omega) \to S^{n,0}(\mathscr{T})$ to be

$$I_{\mathscr{T}} v = \sum_{z \in \mathscr{N}_n(\mathscr{T})} \langle v, \phi_z^* \rangle \phi_z,$$

and observe that by construction this operator is invariant in $S^{n,0}(\mathscr{T})$, namely,

$$I_{\mathscr{T}} P = P \qquad \text{for all } P \in S^{n,0}(\mathscr{T}).$$

In particular, the averaging process giving rise to the values of $I_{\mathscr{T}} v$ for each element $T \in \mathscr{T}$ takes place in the neighborhood $N_{\mathscr{T}}(T)$, whence we also deduce the local invariance

$$I_{\mathscr{T}} P|_T = P \qquad \text{for all } P \in \mathbb{P}_n(N_{\mathscr{T}}(T))$$

as well as the local stability estimate

$$\|I_{\mathscr{T}} v\|_{L^q(T)} \lesssim \|v\|_{L^q(N_{\mathscr{T}}(T))}.$$

We thus may write

$$v - I_{\mathscr{T}} v|_T = (v - P) - I_{\mathscr{T}}(v - P)|_T \qquad \text{for all } T \in \mathscr{T},$$

where $P \in \mathbb{P}_{s-1}$ is arbitrary. It suffices now to prove (63) in the reference element $\hat{T}$ and scale back and forth via Lemma 1; the definition (5) of Sobolev number accounts precisely for this scaling. We keep the notation T for $\hat{T}$, apply the inverse estimate for $\mathbb{P}_n$-polynomials $\|D^t(I_{\mathscr{T}} v)\|_{L^q(T)} \lesssim \|I_{\mathscr{T}} v\|_{L^q(T)}$ to $v - P$ instead of v, and use the above local stability estimate, to infer that

$$\|D^t(v - I_{\mathscr{T}} v)\|_{L^q(T)} \lesssim \|v - P\|_{W_q^t(N_{\mathscr{T}}(T))} \lesssim \|v - P\|_{W_p^s(N_{\mathscr{T}}(T))}.$$

The last inequality is a consequence $W_p^s(N_{\mathscr{T}}(T)) \subset W_q^t(N_{\mathscr{T}}(T))$ because $\operatorname{sob}(W_p^s) > \operatorname{sob}(W_q^t)$. Estimate (63) now follows from the Bramble-Hilbert lemma [16, Lemma 4.3.8], [25, Theorem 3.1.1]

$$\inf_{P\in\mathbb{P}_{s-1}(N_{\mathscr{T}}(T))} \|v-P\|_{W^s_p(N_{\mathscr{T}}(T))} \lesssim \|D^s v\|_{L^p(N_{\mathscr{T}}(T))}. \tag{64}$$

This proves (a). To show (b) we modify the averaging process for boundary nodes and define a set of dual functions with respect to an L^2-scalar product over $(d-1)$-subsimplices contained on $\partial\Omega$; see again [16, 65] for details. This retains the invariance property of $I_{\mathscr{T}}$ on $S^{n,0}(\mathscr{T})$ and guarantees that $I_{\mathscr{T}}v$ has a zero trace if $v\in W^1_1(\Omega)$ does. Hence, the same argument as above applies and (63) follows. □

Remark 16 (Sobolev Numbers). We cannot expect (63) to be valid if $\mathrm{sob}(W^s_p) = \mathrm{sob}(W^t_q)$ since this may not imply $W^s_p(\Omega)\subset W^t_q(\Omega)$; recall the counterexample $W^s_p(\Omega)=W^1_d(\Omega)$ and $W^t_q(\Omega)=L^\infty(\Omega)$ of Sect. 2.1.1. However, equality of Sobolev numbers is allowed in (63) as long as the space embedding is valid.

Remark 17 (Fractional Regularity). We observe that (63) does not require the regularity indices t and s to be integer. The proof follows the same lines but replaces the polynomial degree $s-1$ by the greatest integer smaller that s; the generalization of (64) can be taken from [33].

Remark 18 (Local Error Estimate for Lagrange Interpolant). Let the regularity index s and integrability index $1\le p\le\infty$ satisfy $s-d/p>0$. This implies that $\mathrm{sob}(W^s_p)>\mathrm{sob}(L^\infty)$, whence $W^s_p(\Omega)\subset C(\overline{\Omega})$ and the Lagrange interpolation operator $I_{\mathscr{T}}: W^s_p(\Omega)\to S^{n,0}(\mathscr{T})$ is well defined and satisfies the fully local error estimate

$$\|D^t(v-I_{\mathscr{T}}v)\|_{L^q(T)} \lesssim h_T^{\mathrm{sob}(W^s_p)-\mathrm{sob}(W^t_q)}\|D^s v\|_{L^p(T)}, \tag{65}$$

provided $0\le t\le s$, $1\le q\le\infty$ are such that $\mathrm{sob}(W^s_p)>\mathrm{sob}(W^t_q)$. We point out that $N_{\mathscr{T}}(T)$ in (63) is now replaced by T in (65). We also remark that if v vanishes on $\partial\Omega$ so does $I_{\mathscr{T}}v$. The proof of (65) proceeds along the same lines as that of Proposition 3 except that the nodal evaluation does not extend beyond the element $T\in\mathscr{T}$ and the inverse and stability estimates over the reference element are replaced by

$$\|D^t I_{\mathscr{T}}v\|_{L^q(\hat{T})} \lesssim \|I_{\mathscr{T}}v\|_{L^q(\hat{T})} \lesssim \|v\|_{L^\infty(\hat{T})} \lesssim \|v\|_{W^s_p(\hat{T})}.$$

Remark 19 (Boundary values). The procedure described at the end of the proof of Proposition 3 can be used to interpolate functions with boundary values different from zero while retaining invariance over the finite element space. We refer to [16, 65] for details.

Remark 20 (Localized Estimate). Suppose that $v\in W^1_1(\Omega)$ happens to be a piecewise polynomial of degree $\le n$ on a subdomain Ω_* of Ω. Let ω be a connected component of $\Omega\backslash\Omega_*$ and let the quasi-interpolant $I_{\mathscr{T}}v$ preserve the boundary values of v on $\partial\omega$, as indicated in Remark 19. If we repeat this construction for each connected component ω of $\Omega\backslash\Omega_*$ and define $I_{\mathscr{T}}v=v$ in Ω_*, then $I_{\mathscr{T}}v\in S^{n,0}(\mathscr{T})$ and we deduce the localized estimate for all $1\le p\le\infty$

$$\sum_{T\subset\omega} h_T^{-2}\|v-I_{\mathscr{T}}v\|^p_{L^p(T)} + h_T^{-2+2/p}\|v-I_{\mathscr{T}}v\|^p_{L^p(\partial T)} \lesssim \|\nabla v\|^p_{L^p(\omega)}. \tag{66}$$

This property will be crucial in Chap. 9 to prove quasi-optimality of AFEM.

The local interpolation error estimate in Proposition 3 implies a global one. The latter will be discussed as an upper bound for the error of the finite element solution in the next section.

5.2 A Priori Error Analysis

Combining Theorem 5 with Proposition 3 we derive a so-called *a priori error estimate*, which bounds the error of the finite element solution in terms of the mesh-size function and regularity of the exact solution beyond $H^1(\Omega)$. We present a slightly more general variant than usual. This will help in the successive discussion on error reduction.

Theorem 11 (A Priori Error Estimate). *Let $1 \le s \le n+1, 1 \le p \le 2$, and let the solution u of the model problem* (13) *satisfy $u \in W_p^s(\Omega)$ with $r := \mathrm{sob}(W_p^s) - \mathrm{sob}(H^1) > 0$. Let $U \in \mathbb{V}(\mathscr{T}) = S^{n,0}(\mathscr{T}) \cap H_0^1(\Omega)$ be the corresponding discrete solution. If $h : \Omega \to \mathbb{R}$ denotes the piecewise constant mesh density function, then*

$$\|\nabla(u-U)\|_{L^2(\Omega)} \lesssim \frac{\alpha_2}{\alpha_1} \|h^r D^s u\|_{L^p(\Omega)}. \tag{67}$$

The hidden constant depends on shape coefficient of $\mathscr{T}_0$ and the dimension d.

Proof. Theorem 5 and Proposition 3 yield

$$\|\nabla(u-U)\|_{L^2(\Omega)}^2 \lesssim \frac{\alpha_2}{\alpha_1} \|\nabla(u-I_{\mathscr{T}}u)\|_{L^2(\Omega)}^2 \lesssim \frac{\alpha_2}{\alpha_1} \sum_{T\in\mathscr{T}} h_T^{2r} \|D^s u\|_{L^p(N_{\mathscr{T}}(T))}^2.$$

In order to sum up the right-hand side we need to accumulate in ℓ^p rather than ℓ^2. We recall the elementary property of series $\sum_n a_n \le (\sum_n a_n^q)^{1/q}$ for $0 < q \le 1$. We take $q = p/2$ and apply this property, in conjunction with (55), to arrive at

$$|u-I_{\mathscr{T}}u|_{H^1(\Omega)}^2 \lesssim \left(\sum_{T\in\mathscr{T}} h_T^{rp} \|D^s u\|_{L^p(N_{\mathscr{T}}(T))}^p \right)^{\frac{2}{p}} \lesssim \left(\int_\Omega h(x)^{rp} |D^s u(x)|^p \, dx \right)^{\frac{2}{p}}.$$

This is the asserted estimate (67). □

Notice that in Theorem 11 the exploitable number of derivatives of the exact solution is limited by the polynomial degree

$$1 \le s \le 1+n.$$

Moreover, decreasing the mesh-size function reduces the upper bound (67). The reduction rate is dictated by the difference of the Sobolev numbers

$$r = \mathrm{sob}(W_p^s) - \mathrm{sob}(H^1),$$

and is thus sensitive to the integrability of the relevant derivatives in both left and right-hand sides of (67). The best rate is obtained for integrability index $p = 2$, which coincides with the integrability of the error notion.

Relying solely on decreasing of the mesh-size function, and thus ignoring the local distribution of the derivative $D^s u$ of the exact solution u, leads to uniform refinement or quasi-uniform meshes. The specialization of Theorem 11 to this case reads as follows:

Corollary 6 (Quasi-Uniform Meshes). *Let $1 \le s \le n+1$, and let the solution u of the model problem* (13) *satisfy $u \in H^s(\Omega)$. Let $\mathscr{T}_N$ be a quasi-uniform partition of Ω with N interior nodes and let $U_N \in \mathbb{V}(\mathscr{T}_N)$ be the discrete solution corresponding to the model problem* (13)*. Then*

$$\|\nabla(u - U_N)\|_{L^2(\Omega)} \lesssim \frac{\alpha_2}{\alpha_1} |u|_{H^s(\Omega)} N^{-(s-1)/d}. \tag{68}$$

Proof. Quasi-uniformity of $\mathscr{T}_N$ implies

$$\max_{T\in\mathscr{T}_N} h_T^d \le \max_{T\in\mathscr{T}_N} \overline{h}_T^d \lesssim \min_{T\in\mathscr{T}_N} \underline{h}_T{}^d \le \frac{1}{N} \sum_{T\in\mathscr{T}_N} h_T^d = \frac{|\Omega|}{N}$$

Since $r = (s - d/2) - (1 - d/2) = s - 1$, the assertion follows (67). □

A simple consequence of (68), under full regularity $u \in H^{n+1}(\Omega)$ is the maximal decay rate in terms of degrees of freedom

$$\|\nabla(u - U_N)\|_{L^2(\Omega)} \lesssim \frac{\alpha_2}{\alpha_1} |u|_{H^{n+1}(\Omega)} N^{-n/d}. \tag{69}$$

One may wonder whether (68) is sharp whenever $s < n+1$. The following example addresses this question.

Example 1 (Corner Singularity). We consider the Dirichlet problem for $-\Delta u = f$, for which $\alpha_1 = \alpha_2 = 1$, with exact solution (in polar coordinates)

$$u(r,\theta) = r^{\frac{2}{3}} \sin(2\theta/3) - r^2/4,$$

on an L-shaped domain Ω; this function satisfies $u \in H^s(\Omega)$ for $s < 5/3$. Recall that even though s is fractional, the error estimates are still valid; see Remark 17. In particular, (68) can be derived by space interpolation between $H^1(\Omega)$ and $H^{n+1}(\Omega)$. In Figure 1 we depict the sequence of *uniform* meshes, for which $N \approx h^{-2}$, h being the mesh-size. In Table 1 we report the order of convergence for polynomial degrees $n = 1, 2, 3$. The asymptotic rate is about $h^{2/3}$, or equivalently $N^{-1/3}$, regardless of n and is consistent with the estimate (68). This indicates that (68) is sharp.

The question arises whether the rate $N^{-1/3}$ in Example 1 is just a consequence of uniform refinement or unavoidable. It is important to realize that $u \notin H^s(\Omega)$ for

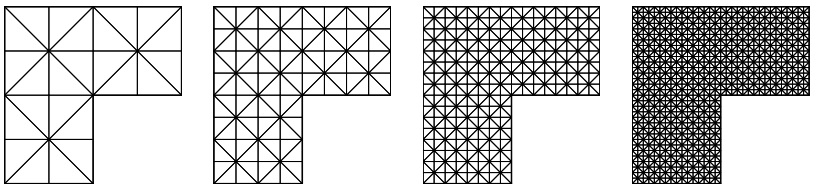

Fig. 10 Sequence of uniform meshes for L-shaped domain Ω

h	linear ($n=1$)	quadratic ($n=2$)	cubic ($n=3$)
1/4	1.14	9.64	9.89
1/8	0.74	0.67	0.67
1/16	0.68	0.67	0.67
1/32	0.66	0.67	0.67
1/64	0.66	0.67	0.67
1/128	0.66	0.67	0.67

Table 1 The asymptotic rate of convergence is about $h^{2/3}$, or equivalently $N^{-1/3}$, irrespective of the polynomial degree n as predicted by (68).

$s \geq 5/3$ and thus (68) is not applicable. However, the problem is not that second order derivatives of u do not exist but rather that they are not square-integrable. In particular, it is true that $u \in W^2_p(\Omega)$ if $1 \leq p < 3/2$. We therefore may apply Theorem 11 with, e.g., $n = 1$, $s = 2$, and $p \in [1, 3/2)$ and then ask whether the structure of (67) can be exploited, e.g., by compensating the local behavior of $D^s u$ with the local mesh-size h. If u is assumed to be known, this enterprise naturally leads to meshes *adapted to u* that may be *graded*. We discuss this possibility in Sect. 5.3 and propose a condition that should be satisfied by these meshes.

5.3 Principle of Error Equidistribution

For the model problem and its standard discretization, Theorem 5 and the considerations at the end of Sect. 5.2 suggest the optimization problem:

> *Given a function $u \in H^1(\Omega)$ and an integer $N > 0$ find conditions for a shape regular mesh $\mathscr{T}$ to minimize the error $|u - I_{\mathscr{T}} u|_{H^1(\Omega)}$ subject to the constraint that the number of degrees of freedom does not exceed N.*

In the framework of Chap. 4 this becomes a *discrete* optimization problem. Here we consider a simplified setting and, similar to Babuška and Rheinboldt [5], invoke a *continuous model*:

- The dimension is $d = 2$ and the regularity of $u \in C^2(\Omega) \cap W^2_p(\Omega)$ with $1 < p \leq 2$;
- There exists a C^1 function $h : \Omega \to \mathbb{R}$, a mesh density function, with the property that $h(x)$ is equivalent to h_T for all $T \in \mathscr{T}$ with equivalence constants only depending on shape regularity (thus on the shape coefficient of $\mathscr{T}_0$);

- The number of degrees of freedom and local mesh-size are related through the relation
$$N = \int_\Omega \frac{dx}{h(x)^2}.$$
- The mesh $\mathscr{T}$ is sufficiently fine so that D^2u is essentially constant within each element $T \in \mathscr{T}$;
- The error is given by the formula
$$\left(\int_\Omega h(x)^{2(p-1)} |D^2u(x)|^p \, dx \right)^{\frac{2}{p}}.$$

A few comments about this model are in order. The first condition is motivated by the subsequent discussion and avoids dealing with Besov spaces with integrability index $p < 1$; in particular, all corner singularities for $d = 2$ are of the form $u(x) \approx |x|^\gamma$ and satisfy $u \in C^2(\Omega) \cap W^2_p(\Omega)$ for some $p > 1$. The second assumption is quite realistic since shape regularity is sufficient for the existence of a C^∞ mesh density with the property $D^l h \approx h^{1-t}$; see Nochetto et al. [57]. The third condition is based on the heuristics that the number of elements per unit of area is about $h(x)^{-2}$. The fourth assumption can be rephrased as follows: $\int_T |D^2u|^p \approx h_T^2 |D^2u(x_T)|^p$ where x_T is the barycenter of $T \in \mathscr{T}$. Finally, the fifth assumption replaces the error by an upper bound. In fact, if $I_\mathscr{T}$ is the Lagrange interpolation operator, we can use the local interpolation estimates (65) to write

$$|u - I_\mathscr{T} u|_{H^1(T)} \lesssim h_T^{\mathrm{sob}(W^2_p) - \mathrm{sob}(H^1)} |u|_{W^2_p(T)} \lesssim h_T^{2-\frac{2}{p}} |u|_{W^2_p(T)} \qquad \text{for all } T \in \mathscr{T}$$

and then argue as in the proof of Theorem 11 to derive the upper bound

$$\|\nabla(u-U)|^2_{L^2(\Omega)} \lesssim \left(\int_\Omega h(x)^{2(p-1)} |D^2u(x)|^p \, dx \right)^{\frac{2}{p}}.$$

Since we would like to minimize the error for a given number of degrees of freedom N, we propose the Lagrangian

$$\mathscr{L}[h, \lambda] = \int_\Omega \left(h(x)^{2(p-1)} |D^2u(x)|^p - \frac{\lambda}{h(x)^2} \right) dx,$$

with Lagrange multiplier $\lambda \in \mathbb{R}$. A stationary point of $\mathscr{L}$ satisfies (see Problem 25)

$$h(x)^{2(p-1)+2} |D^2u(x)|^p = \text{constant},$$

and thus requires a variable mesh-size $h(x)$ that compensates the local behavior of $D^2u(x)$. This relation can be interpreted as follows: since the error E_T associated with element $T \in \mathscr{T}$ satisfies

$$E_T = h_T^{2(p-1)} \int_T |D^2u|^p \approx h_T^{2(p-1)+2} |D^2u(x_T)|^p,$$

we infer that the *element error* is equidistributed.

Summarizing (and ignoring the asymptotic aspects of the above continuous model), a candidate for the sought condition is

$$E_T \approx \Lambda \quad \text{(constant)} \quad \text{for all } T \in \mathscr{T}.$$

Meshes satisfying this property have been constructed by Babuška et al [4] for corner singularities and $d = 2$; see also [39]. Problem 27 explores this matter and proposes a specific mesh grading towards the origin. However, what the above argument does not address is whether such meshes exist in general and whether they can be actually constructed upon bisecting the initial mesh $\mathscr{T}_0$, namely that $\mathscr{T} \in \mathbb{T}$.

5.4 Adaptive Approximation

The purpose of this concluding section is to show that the maximum decay rate $N^{-n/d}$ in (69) can be reached under weaker regularity assumption when using suitably adapted meshes. Following the work of Binev et al. [14], we use an adaptive algorithm that is based on the knowledge of the element errors and on bisection.

The algorithm can be motivated with the above equidistribution principle in the following manner. Let $\delta > 0$ be a given tolerance and the polynomial degree $n = 1$. If the element error is equidistributed, that is $E_T \approx \delta^2$, and the global error decays with maximum rate $N^{-1/2}$, then

$$\delta^4 N \approx \sum_{T \in \mathscr{T}_N} E_T^2 = |u - I_{\mathscr{T}} u|^2_{H^1(\Omega)} \lesssim N^{-1}$$

that is $N \lesssim \delta^{-2}$. With this in mind, we impose $E_T \leq \delta^2$ as a common threshold to stop refining and expect $N \lesssim \delta^{-2}$.

The following algorithm implements this idea.

Algorithm (Thresholding). Given a tolerance $\delta > 0$ and a conforming mesh $\mathscr{T}_0$, THRESHOLD finds a conforming refinement $\mathscr{T} \geq \mathscr{T}_0$ of $\mathscr{T}_0$ by bisection such that $E_T \leq \delta^2$ for all $T \in \mathscr{T}$: let $\mathscr{T} = \mathscr{T}_0$ and

THRESHOLD$(\mathscr{T}, \delta)$
while $\mathscr{M} := \{T \in \mathscr{T} \,|\, E_T > \delta^2\} \neq \emptyset$
  $\mathscr{T} :=$ REFINE$(\mathscr{T}, \mathscr{M})$
end while
return$(\mathscr{T})$

We now discuss the situation mentioned above. Assume

$$u \in W^2_p(\Omega), \quad p > 1, d = 2, \tag{70}$$

which implies that u is uniformly continuous in Ω and we can take $I_{\mathscr{T}}$ to be the Lagrange interpolation operator. Since $p > 1$ we have $r = 2(1 - 1/p) > 0$, according

to (65), and

$$E_T \lesssim h_T^r \, \|D^2 u\|_{L^p(T)}. \tag{71}$$

Therefore, THRESHOLD *terminates* because h_T decreases monotonically to 0 with bisection. The quality of the resulting mesh is assessed next.

Theorem 12 (Thresholding). *If $u \in H_0^1(\Omega)$ verifies* (70)*, then the output $\mathscr{T} \in \mathbb{T}$ of* THRESHOLD *satisfies*

$$|u - I_{\mathscr{T}} u|_{H^1(\Omega)} \leq \delta^2 (\#\mathscr{T})^{1/2}, \qquad \#\mathscr{T} - \#\mathscr{T}_0 \lesssim \delta^{-2} \, |\Omega|^{1-1/p} \|D^2 u\|_{L^p(\Omega)}.$$

Proof. Let $k \geq 1$ be the number of iterations of THRESHOLD before termination. Let $\mathscr{M} = \mathscr{M}_0 \cup \cdots \cup \mathscr{M}_{k-1}$ be the set of marked elements. We organize the elements in $\mathscr{M}$ by size in such a way that allows for a counting argument. Let $\mathscr{P}_j$ be the set of elements T of $\mathscr{M}$ with size

$$2^{-(j+1)} \leq |T| < 2^{-j} \quad \Rightarrow \quad 2^{-(j+1)/2} \leq h_T < h_T^{-j/2}.$$

We proceed in several steps.

1 We first observe that all T's in $\mathscr{P}_j$ are *disjoint*. This is because if $T_1, T_2 \in \mathscr{P}_j$ and $\mathring{T}_1 \cap \mathring{T}_2 \neq \emptyset$, then one of them is contained in the other, say $T_1 \subset T_2$, due to the bisection procedure. Thus

$$|T_1| \leq \frac{1}{2} \, |T_2|$$

contradicting the definition of $\mathscr{P}_j$. This implies

$$2^{-(j+1)} \, \#\mathscr{P}_j \leq |\Omega| \quad \Rightarrow \quad \#\mathscr{P}_j \leq |\Omega| \, 2^{j+1}. \tag{72}$$

2 In light of (71), we have for $T \in \mathscr{P}_j$

$$\delta^2 \leq E_T \lesssim 2^{-(j/2)r} \|D^2 u\|_{L^p(T)}.$$

Therefore

$$\delta^{2p} \, \#\mathscr{P}_j \lesssim 2^{-(j/2)rp} \sum_{T \in \mathscr{P}_j} \|D^2 u\|_{L^p(T)}^p \leq 2^{-(j/2)rp} \, \|D^2 u\|_{L^p(\Omega)}^p$$

whence

$$\#\mathscr{P}_j \lesssim \delta^{-2p} \, 2^{-(j/2)rp} \, \|D^2 u\|_{L^p(\Omega)}^p. \tag{73}$$

3 The two bounds for $\#\mathscr{P}$ in (72) and (73) are complementary. The first is good for j small whereas the second is suitable for j large (think of $\delta \ll 1$). The crossover takes place for j_0 such that

$$2^{j_0+1} |\Omega| = \delta^{-2p} \, 2^{-j_0(rp/2)} \|D^2 u\|_{L^p(\Omega)}^p \quad \Rightarrow \quad 2^{j_0} \approx \delta^{-2} \, \frac{\|D^2 u\|_{L^p(\Omega)}}{|\Omega|^{1/p}}.$$

4 We now compute

$$\#\mathscr{M} = \sum_j \#\mathscr{P}_j \lesssim \sum_{j\le j_0} 2^j|\Omega| + \delta^{-2p}\,\|D^2u\|^p_{L^p(\Omega)} \sum_{j>j_0} (2^{-rp/2})^j.$$

Since

$$\sum_{j\le j_0} 2^j \approx 2^{j_0}, \qquad \sum_{j>j_0} (2^{-rp/2})^j \lesssim 2^{-(rp/2)j_0} = 2^{-(p-1)j_0}$$

we can write

$$\#\mathscr{M} \lesssim \left(\delta^{-2} + \delta^{-2p}\delta^{2(p-1)}\right)|\Omega|^{1-1/p}\,\|D^2u\|_{L^p(\Omega)} \approx \delta^{-2}\,|\Omega|^{1-1/p}\,\|D^2u\|_{L^p(\Omega)}.$$

We finally apply Theorem 10 to arrive at

$$\#\mathscr{T} - \#\mathscr{T}_0 \lesssim \#\mathscr{M} \lesssim \delta^{-2}\,|\Omega|^{1-1/p}\,\|D^2u\|_{L^p(\Omega)}.$$

5 It remains to estimate the energy error. We have, upon termination of THRESHOLD, that $E_T \le \delta^2$ for all $T \in \mathscr{T}$. Then

$$|u - I_{\mathscr{T}}u|^2_{H^1(\Omega)} = \sum_{T\in\mathscr{T}} E_T^2 \le \delta^4 \#\mathscr{T}.$$

This concludes the Theorem. □

By relating the threshold value δ and the number of refinements N, we obtain a result about the convergence rate.

Corollary 7 (Convergence Rate). *Let $u \in H^1_0(\Omega)$ satisfy* (70)*. Then for $N > \#\mathscr{T}_0$ integer there exists $\mathscr{T} \in \mathbb{T}$ such that*

$$|u - I_{\mathscr{T}}u|_{H^1(\Omega)} \lesssim |\Omega|^{1-1/p}\,\|D^2u\|_{L^p(\Omega)} N^{-1/2}, \qquad \#\mathscr{T} - \#\mathscr{T}_0 \lesssim N.$$

Proof. Choose $\delta^2 = |\Omega|^{1-1/p}\,\|D^2u\|_{L^p(\Omega)} N^{-1}$ in Theorem 12. Then, there exists $\mathscr{T} \in \mathbb{T}$ such that $\#\mathscr{T} - \#\mathscr{T}_0 \lesssim N$ and

$$\begin{aligned} |u - I_{\mathscr{T}}u|_{H^1(\Omega)} &\lesssim |\Omega|^{1-1/p}\,\|D^2u\|_{L^p(\Omega)} N^{-1}\left(N + \#\mathscr{T}_0\right)^{1/2} \\ &\lesssim |\Omega|^{1-1/p}\,\|D^2u\|_{L^p(\Omega)} N^{-1/2} \end{aligned}$$

because $N > \#\mathscr{T}_0$. This finishes the Corollary. □

Remark 21 (Piecewise smoothness). The global regularity (70) can be weakened to *piecewise W^2_p regularity* over the initial mesh $\mathscr{T}_0$, namely $W^2_p(\Omega;\mathscr{T}_0)$, and global $H^1_0(\Omega)$. This is because $W^2_p(T) \hookrightarrow C^0(\overline{T})$ for all $T \in \mathscr{T}_0$, whence $I_{\mathscr{T}}$ can be taken to be the Lagrange interpolation operator.

Remark 22 (Case $p < 1$). Consider either polynomial degree $n > 1$ and $d = 2$ or $n \ge 1$ for $d > 2$. The Sobolev number corresponding to a space with regularity of order $n+1$ is

$$n+1-\frac{d}{p}=\mathrm{sob}(H^1)=1-\frac{d}{2} \quad \Rightarrow \quad p=\frac{d}{n+d/2}.$$

For $d=2$ this implies $p<1$. Spaces based on $L^p(\Omega)$, $p<1$, are unusual in finite element theory but not in approximation theory [71, 30, 28]. The argument of Theorem 12 works provided we replace (71) by a modulus of regularity; in fact, $D^{n+1}u$ would not be locally integrable and so would fail to be a distribution. This requires two ingredients:

- The construction of a quasi-interpolation operator $I_{\mathscr{T}} : L^p(\Omega) \to S^{n,0}(\mathscr{T})$ for $p<1$ with optimal approximation properties; such operator $I_{\mathscr{T}}$ is inevitably nonlinear. We refer to [30, 28, 58], as well as [37] where the following key property is proven: $I_{\mathscr{T}}(v+P)=I_{\mathscr{T}}(v)+P$ for all $P \in S^{n,0}(\mathscr{T})$ and $v \in L^p(\Omega)$.
- Besov regularity properties of the solution u of an elliptic boundary value problem; we refer to [27] for such an endeavor for 2d Lipschitz domains and the Laplace operator. For the model problem with discontinuous coefficients as well as for $d>2$ this issue seems to be open in general.

Applying Corollary 7 to Example 1, we see that the maximum decay rate $N^{-1/2}$ for polynomial degree $n=1$ and dimension $d=2$, as well as $N^{-n/d}$ for $n \geq 1, d \geq 2$ when taking Remark 22 into account, can be reestablished by judicious mesh grading. Of course the thresholding algorithm cannot be applied directly within the finite element method because the exact solution u is typically unknown. In fact, we are only able to replace the element energy error by computable element error indicators, and thus gain access to u indirectly. This is the topic of a posteriori error analysis and is addressed in Chap. 6.

5.5 Problems

Problem 24. Let $\mathscr{T}$ be a shape regular and quasi-uniform triangulation of $\Omega \subset \mathbb{R}^d$. Let $\mathbb{V}_{\mathscr{T}}$ be the space of (possibly discontinuous) finite elements of degree $\leq n$. Given $u \in L^2(\Omega)$, the L^2-projection $U_{\mathscr{T}} \in \mathbb{V}_{\mathscr{T}}$ is defined by

$$\int_\Omega (u-U_{\mathscr{T}})V=0 \quad \text{for all } V \in \mathbb{V}_{\mathscr{T}}.$$

Show

(a) $\|u-U_{\mathscr{T}}\|_{L^2(\Omega)} \lesssim h^{n+1}\,|u|_{H^{n+1}(\Omega)}$
(b) $\|u-U_{\mathscr{T}}\|_{H^{-m}(\Omega)} \lesssim h^{n+1+m}\,|u|_{H^{n+1}(\Omega)}$

for $0 \leq m \leq n+1$ and h being the maximal mesh size of $\mathscr{T}$. The estimate in (b) ensures *superconvergence*.

Problem 25. Let $h(x)$ a smooth function locally equivalent to the mesh-size. Prove that a stationary point of the Lagrangian

$$\mathscr{L}[h,\lambda] = \int_\Omega \left(h(x)^{2(p-1)} \, |D^2 u(x)|^p - \frac{\lambda}{h(x)^2} \right) dx$$

satisfies the optimality condition

$$h^{2(p-1)+2} \, |D^2 u|^p = \text{constant}.$$

Problem 26. Consider the solution u of the model problem in Sect. 2.2.1 with corner singularity:

$$u(r,\theta) = r^\gamma \phi(\theta) \qquad 0 < \gamma < 1$$

in polar coordinates (r,θ). Show that $u \in W_p^2(\Omega) \setminus H^2(\Omega)$ for $1 \le p < 2/(2-\gamma)$.

Problem 27. Use the Principle of Equidistribution to determine the grading of an mesh for a corner singularity

$$u(r,\theta) = r^\gamma \phi(\theta) \qquad (0 < \gamma < 1).$$

In fact, show that

$$h_T = \Lambda \, \text{dist}(T,0)^{1-\gamma/2} \qquad (\Lambda = \text{ constant}).$$

Count the number of elements using the expression $N \approx \int_\Omega \frac{dx}{h(x)^2}$ and derive an optimal bound $|u - I_{\mathscr{T}_N} u|_{H^1(\Omega)} \lesssim N^{-1/2}$ for polynomial degree $n = 1$.

Problem 28. Consider the function u of Problem 26.
(a) Examine the construction of an graded mesh via the Thresholding Algorithm.
(b) Repeat the proof of Theorem 12 replacing the W_p^2 regularity by the corresponding local H^2 regularity of u depending on the distance to the origin.

6 A Posteriori Error Analysis

Suppose, as it is generically the case, that the solution of a boundary value problem is unknown. Then we may use a numerical method to compute an approximate solution. Of course, it is useful to have information about the error of such an approximation. Moreover, if the error is still to big, one would like to know how to modify the discretization so as to reduce the error effectively.

The results of the preceding chapters provide little such information, because they involve the exact solution and/or are of asymptotic nature. However, so-called *a posteriori error estimators* extract such information from the given problem and the approximate solution, without invoking the exact solution. Starting with the pioneering work [5] of Babuška and Rheinboldt , a great deal of work has been devoted to their derivation. We refer to [1, 6, 76] for an overview of the state-of-the-art.

This chapter is an introduction to a posteriori error estimators, providing the essentials for the following chapters about adaptive algorithms. To this end, we shall mainly restrict ourselves to the model problem of Sect. 2.2.1 and we will drop the index N or $\mathscr{T}$, since it will be kept fixed during the whole chapter.

6.1 Error and Residual

Let u be the exact solution of (10) and U be a corresponding Petrov-Galerkin solution as in (37). We want to obtain information about the error function $u-U$, which is typically unknown. The so-called *residual* $\mathscr{R} = \mathscr{R}(U,f) \in W^*$ given by

$$\langle \mathscr{R}, w\rangle := \langle f, w\rangle - \mathscr{B}[U,w] \quad \text{for all } w \in \mathbb{W}$$

depends only on data and the approximate solution U and is related to the error function by

$$\langle \mathscr{R}, w\rangle = \mathscr{B}[u-U,w] \quad \text{for all } w \in \mathbb{W}. \tag{74}$$

If the error notion of interest is $\|u-U\|_{\mathbb{V}}$, the following lemma determines a dual-norm of $\mathscr{R}$ that is equivalent to the error.

Lemma 11 (Abstract A posteriori Error Estimate). *There holds*

$$\alpha\|u-U\|_{\mathbb{V}} \le \|\mathscr{R}\|_{\mathbb{W}^*} \le \|\mathscr{B}\|\|u-U\|_{\mathbb{V}}, \tag{75}$$

where $0<\alpha\le\|\mathscr{B}\|$ *are the inf-sup and continuity constants of* $\mathscr{B}$ *from* (21a) *and* (16).

Proof. The inf-sup condition (21a) and (74) imply

$$\alpha\|u-U\|_{\mathbb{V}} \le \sup_{\|w\|_{\mathbb{W}}=1} \mathscr{B}[u-U,w] = \|\mathscr{R}\|_{\mathbb{W}^*},$$

while (74) and (16) imply

$$\|\mathscr{R}\|_{\mathbb{W}^*} = \sup_{\|w\|_{\mathbb{W}}=1} \mathscr{B}[u-U,w] \le \|\mathscr{B}\|\|u-U\|_{\mathbb{V}}. \qquad \square$$

In view of the result, we are left with (approximately) evaluting $\|\mathscr{R}\|_{\mathbb{W}^*}$ at an acceptable cost. Notice that, while the quasi-best approximation property (5) relies on the stability of the discretization, Lemma 11 relies on the well-posedness of the continuous problem (10). It is thus a first, discretization-independent step. Here it was rather straight-forward, it can get more involved depending on problem and error notion.

There are various techniques for evaluting $\|\mathscr{R}\|_{\mathbb{W}^*}$. This second step depends on the discretization. In what follows, we present the most basic and common approach, *standard residual estimation*, in the case of our model problem of Sect. 2.2.1 and its standard discretization of Sect. 3.2.2.

Before embarking on it, it is instructive to analyze the structure of the residual for the model problem, where $\mathbb{W}^* = H^{-1}(\Omega)$, U is piecewise polynomial function over a triangulation $\mathscr{T}$, and the residual is the distribution

$$\mathscr{R} = f + \operatorname{div}(\boldsymbol{A}\nabla U) \in H^{-1}(\Omega).$$

To this end, we suppose $f \in L^2(\Omega)$. This allows us to write $\langle \mathscr{R}, w\rangle$ as integrals over each $T \in \mathscr{T}$ and integration by parts yields the representation:

$$\begin{aligned} \langle \mathscr{R}, w\rangle &= \int_\Omega fw - \nabla U \cdot \boldsymbol{A}\nabla w = \sum_{T\in\mathscr{T}} \int_T fw - \nabla U \cdot \boldsymbol{A}\nabla w \\ &= \sum_{T\in\mathscr{T}} \int_T rw + \sum_{S\in\mathring{\mathscr{S}}} \int_S jw, \end{aligned} \tag{76}$$

with

$$\begin{aligned} &r = f + \operatorname{div}(\boldsymbol{A}\nabla U) \quad \text{in any simplex } T \in \mathscr{T}, \\ &j = [\![\boldsymbol{A}\nabla U]\!] = \boldsymbol{n}^+ \cdot \boldsymbol{A}\nabla U_{|T^+} + \boldsymbol{n}^- \cdot \boldsymbol{A}\nabla U_{|T^-} \quad \text{on any internal side } S \in \mathring{\mathscr{S}} \end{aligned}$$

and $\boldsymbol{n}^+$, $\boldsymbol{n}^-$ are unit normals pointing towards T^+, $T^- \in \mathscr{T}$. We see that the distribution $\mathscr{R}$ consists of a regular part r, called *interior or element residual*, and a singular part j, called *jump or interelement residual*. The regular part is absolutely continuous w.r.t. the d-dimensional Lebesgue measure and is related to the strong form of the PDE. The singular part is supported on the skeleton $\Gamma = \bigcup_{S\in\mathring{\mathscr{S}}} S$ of $\mathscr{T}$ and is absolutely continuous w.r.t. the $(d-1)$-dimensional Hausdorff measure.

We point out that this structure is not special to the model problem and its discretization but rather arises from the weak formulation of the PDE and the piecewise construction of finite element spaces.

6.2 Global Upper Bound

As already mentioned, we provide an a posteriori analysis for the model problem in Sect. 2.2.1 using standard residual estimation. This approach provides an upper bound $\|\mathscr{R}\|_{\mathbb{W}^*}$ with the help of suitably weighted Lebesgue norms (which are considered to be computable). We will see below that the weights are crucial for the sharpness of the derived bound.

In what follows, we shall write '$\lesssim$' instead of $\leq C$, where the constant C is bounded in terms of the shape coefficent $\sigma_{\mathscr{T}}$ of the triangulation $\mathscr{T}$ and the dimension d. The presentation here is a simplified version of [74], which has been influenced by [5, 20, 54] and provides in particular constants that are explicit in terms of local Poincaré constants.

6.2.1 Tools

For bounding $\|\mathscr{R}\|_{\mathbb{W}^*}$ we need two tools: a trace inequality that will help to bound the singular part with the jump residual and a Poincaré-type inequality that will take care of the lower order norms arising in the trace inequality and from the regular part with the element residual. We start by deriving the trace inequality.

Lemma 12 (Trace Identity). *Let T be a d-simplex, S a side of T, and z the vertex opposite to S. Defining the vector field $\boldsymbol{q}_S$ by*

$$\boldsymbol{q}_S(x) := x - z$$

the following equality holds

$$\frac{1}{|S|}\int_S v = \frac{1}{|T|}\int_T v + \frac{1}{d|T|}\int_T \boldsymbol{q}_S \cdot \nabla v \quad \textit{for all } v \in W_1^1(T).$$

Proof. We start with properties of the vector field $\boldsymbol{q}_S$. Let S' be an arbitrary side of T and fix some $y \in S'$. We then see $\boldsymbol{q}_S(x)\cdot\boldsymbol{n}_T = \boldsymbol{q}_S(y)\cdot\boldsymbol{n}_T + (x-y)\cdot\boldsymbol{n}_T = \boldsymbol{q}_S(y)\cdot\boldsymbol{n}_T$ for any $x \in S'$ since $x-y$ is a tangent vector to S'. Therefore, on each side of T, the associated normal flux $\boldsymbol{q}_S\cdot\boldsymbol{n}_T$ is constant. In particular, we see $\boldsymbol{q}_S\cdot\boldsymbol{n}_T$ vanishes on $\partial T \setminus S$ by choosing $y = z$ for sides emanating from z. Moreover, $\operatorname{div}\boldsymbol{q}_S = d$. Thus, if $v \in C^1(\overline{T})$, the Divergence Theorem yields

$$\int_T \boldsymbol{q}_S\cdot\nabla v = -d\int_T v + (\boldsymbol{q}_S\cdot\boldsymbol{n}_T)_{|S}\int_S v.$$

Take $v = 1$ to show $(\boldsymbol{q}_S\cdot\boldsymbol{n}_T)_{|S} = d|T|/|S|$ and extend the result to $v \in W_1^1(T)$ by density. □

The following corollary is a ready-to-use form for our purposes.

Corollary 8 (Scaled Trace Inequality). *For any side $S \subset T$ the following inequality holds*

$$\|v\|_{L^2(S)} \lesssim h_S^{-1/2}\|v\|_{L^2(T)} + h_S^{1/2}\|\nabla v\|_{L^2(T)} \quad \textit{for all } v \in H^1(T) \tag{77}$$

where $h_S =: |S|^{1/(d-1)}$.

Proof. Problem 30. □

We next present the Poincaré-type inequality. Let

$$\omega_z = \cup_{T \ni z} T$$

be the star (or patch) around a vertex $z \in \mathscr{V}$ of $\mathscr{T}$. We define

$$h_z := |\omega_z|^{1/d}$$

and notice that this quantity is, up to the shape coefficient of $\mathscr{T}$, equivalent to the diameter of ω_z, to h_T if $T \subset \omega_z$ and to h_S if $S \subset \omega_z$.

Lemma 13 (Local Poincaré-Type Inequality). *For any $v \in H^1_0(\Omega)$ and $z \in \mathscr{V}$ there exists $c_z \in \mathbb{R}$ such that*

$$\|v - c_z\|_{L^2(\omega_z)} \lesssim h_z \|\nabla v\|_{L^2(\omega_z)}. \tag{78}$$

If $z \in \partial\Omega$ is a boundary vertex, then we can take $c_z = 0$.

Proof. [1] In fact, for any $z \in \mathscr{V}$ the value

$$\bar{c}_z = \frac{1}{|\omega_z|}\int_{\omega_z} v$$

is an optimal choice and (78) can be shown with $c_z = \bar{c}_z$ as (64).
[2] If $z \in \partial\Omega$, then we observe that there exists a side $S \subset \partial\omega_z \cap \partial\Omega$ such that $v = 0$ on S. We therefore can write

$$v = v - \frac{1}{|S|}\int_S v = (v - \bar{c}_z) - \frac{1}{|S|}\int_S (v - \bar{c}_z)$$

and thus, using Corollary 8 and Step 1 for the second term,

$$\|v\|_{L^2(\omega_z)} \lesssim \|v - \bar{c}_z\|_{L^2(\omega_z)} + h_z\|\nabla v\|_{L^2(\omega_z)} \lesssim h_z\|\nabla v\|_{L^2(\omega_z)},$$

which establishes the supplement for boundary vertices. □

6.2.2 Derivation of the Upper Bound

We now pass to the proper derivation of the upper bound. The following properties of the Courant basis $\{\phi_z\}_{z\in\mathscr{V}}$ from Theorem 6 are instrumental:

- It provides a discrete partition of unity:

$$\sum_{z\in\mathscr{N}} \phi_z = 1 \quad \text{in } \Omega. \tag{79}$$

- Each function ϕ_z is contained in $S^{n,0}(\mathscr{T})$ and so the residual is orthogonal to the interior contributions of the partition of unity:

$$\langle \mathscr{R}, \phi_z \rangle = 0 \quad \text{for all } z \in \mathring{\mathscr{V}} := \mathscr{V} \cap \Omega. \tag{80}$$

The second property corresponds to the Galerkin orthogonality. Notice that the first property involve all vertices, while in the second one the boundary vertices are excluded. For this reason, the supplement on boundary vertices in Lemma 78 is important.

For any $w \in H_0^1(\Omega)$ we start by applying (79) and then (80) with c_z from Lemma 13 for w to write

$$\langle \mathscr{R}, w \rangle = \sum_{z\in\mathscr{V}} \langle \mathscr{R}, w\,\phi_z \rangle = \sum_{z\in\mathscr{V}} \langle \mathscr{R}, (w-c_z)\phi_z \rangle,$$

where $c_z = 0$ whenever $z \in \partial\Omega$. In view of representation (76), we can write

$$|\langle \mathscr{R}, (w-c_z)\phi_z \rangle| \le \int_{\omega_z} |r||w-c_z|\phi_z + \int_{\gamma_z} |j||w-c_z|\phi_z$$

where γ_z is the skeleton of ω_z, i.e. the union of all sides emanating from z. We examine each term on the right hand side seperately. Invoking $\|\phi_z\|_{L^\infty(\omega_z)} \le 1$ and (78), we obtain

$$\int_{\omega_z} |r||w-c_z|\phi_z \le \|r\,\phi_z^{1/2}\|_{L^2(\omega_z)} \|w-c_z\|_{L^2(\omega_z)} \lesssim h_z \|r\,\phi_z^{1/2}\|_{L^2(\omega_z)} \|\nabla w\|_{L^2(\omega_z)}.$$

Likewise, employing (77) and (78), we get

$$\int_{\gamma_z} |j||w-c_z|\phi_z \le \|j\,\phi_z^{1/2}\|_{L^2(\gamma_z)} \|w-c_z\|_{L^2(\gamma_z)} \lesssim h_z^{1/2} \|j\,\phi_z^{1/2}\|_{L^2(\gamma_z)} \|\nabla w\|_{L^2(\omega_z)}.$$

Therefore,

$$|\langle \mathscr{R}, w\,\phi_z \rangle| \lesssim \left(h_z \|r\,\phi_z^{1/2}\|_{L^2(\omega_z)} + h_z^{1/2} \|j\,\phi_z^{1/2}\|_{L^2(\gamma_z)} \right) \|\nabla w\|_{L^2(\omega_z)}.$$

Summing over $z \in \mathscr{V}$ and using Cauchy-Schwarz in $\mathbb{R}^{\#\mathscr{T}}$ gives

$$|\langle \mathscr{R}, w \rangle| \lesssim \left(\sum_{z\in\mathscr{V}} h_z^2 \|r\,\phi_z^{1/2}\|_{L^2(\omega_z)}^2 + h_z \|j\,\phi_z^{1/2}\|_{L^2(\gamma_z)}^2 \right)^{1/2} \left(\sum_{z\in\mathscr{V}} \|\nabla w\|_{L^2(\omega_z)}^2 \right)^{1/2}.$$

Denote by $h\colon \Omega \to \mathbb{R}^+$ the mesh-size function given by $h(x) := |S|^{1/k}$ if x belongs to the interior of the k-subsimplex S of $\mathscr{T}$ with $k \in \{1,\dots,d\}$. Then for all $x \in \omega_z$ we have $h_z \lesssim h(x)$. Therefore employing (79) once more and recalling that Γ is the union of all interior sides of $\mathscr{T}$, we proceed by

$$\sum_{z\in\mathscr{V}} h_z^2\|r\,\phi_z^{1/2}\|^2_{L^2(\omega_z)} + h_z\|j\,\phi_z^{1/2}\|^2_{L^2(\gamma_z)} \lesssim \sum_{z\in\mathscr{V}} \|h\,r\,\phi_z^{1/2}\|^2_{L^2(\Omega)} + \|h^{1/2}\,j\,\phi_z^{1/2}\|^2_{L^2(\Gamma)}$$
$$= \|h\,r\|^2_{L^2(\Omega)} + \|h^{1/2}\,j\|^2_{L^2(\Gamma)}.$$

We next resort to the finite overlapping property of stars, namely

$$\sum_{z\in\mathscr{V}} \chi_{\omega_z}(x) \le d+1$$

to deduce

$$\sum_{z\in\mathscr{V}} \|\nabla w\|^2_{L^2(\omega_z)} \lesssim \|\nabla w\|^2_{L^2(\Omega)}.$$

Thus, introducing the *element indicators*

$$\mathscr{E}^2(U,T) := h_T^2\|r\|^2_{L^2(T)} + h_T\|j\|^2_{L^2(\partial T\setminus\partial\Omega)} \tag{81}$$

and the *error estimator*

$$\mathscr{E}^2(U,\mathscr{T}) = \sum_{T\in\mathscr{T}} \mathscr{E}^2(U,T), \tag{82}$$

we have derived

$$\|\mathscr{R}\|_{\mathbb{W}^*} \lesssim \mathscr{E}(U,\mathscr{T}).$$

Combing this with the abstract a posteriori bound in Lemma 11, we obtain the main result of this section.

Theorem 13 (Upper Bound). *Let u and U be exact and Galerkin solution of the model problem and its standard discretization. Then there holds the following global upper bound:*

$$\|\nabla(u-U)\|_{L^2(\Omega)} \le \frac{C}{\alpha_1}\mathscr{E}(U,\mathscr{T}) \tag{83}$$

where α_1 is the global smallest eigenvalue of $\boldsymbol{A}(x)$ and C depends only on the shape coefficient $\sigma_{\mathscr{T}}$ and on the dimension d.

6.2.3 Sharpness of Weighted Lebesgue Norms

The indicators $\mathscr{E}(U,T)$, $T\in\mathscr{T}$, consists of weighted L^2-norms. The weights h_T and $h_T^{1/2}$ arise from the local Poincaré inequalities (78), which in turn rely on the orthogonality (80) of the residual. If we do not exploit orthogonality and use a global Poincaré-type inequality instead of the local ones, the resulting weights are 1 and $h_T^{-1/2}$ and the corresponding upper bound has a lower asymptotic decay rate. We wonder whether the ensuing weights h_T and $h_T^{1/2}$ are accurate and explore this issue for the first weight h_T of the element residual. The following discussion is a elaborated version of [62, Remark 3.1].

First we notice that the local counterpart of $\|\mathscr{R}\|_{H^{-1}(\Omega)}$ is $\|\mathscr{R}\|_{H^{-1}(T)}$ and observe

$$\|\mathscr{R}\|_{H^{-1}(T)} = \sup_{\|\nabla w\|_{L^2(T)}\le 1} \langle \mathscr{R}, w\rangle = \sup_{\|\nabla w\|_{L^2(T)}\le 1} \int_T rw = \|r\|_{H^{-1}(T)} \tag{84}$$

thanks to the representation (76). This suggests to compare the weighted norm $h_T\|r\|_{L^2(T)}$ in the indicator with the local negative norm $\|r\|_{H^{-1}(T)}$. Mimicking the local part in the argument of Sect. 6.2.2, we derive

$$\int_T rw \le \|r\|_{L^2(T)}\|w\|_{L^2(T)} \lesssim h_T\|r\|_{L^2(T)}\|\nabla w\|_{L^2(T)}$$

with the help of the Poincaré-Friedrichs inequality (7). Consequently there holds

$$\|r\|_{H^{-1}(T)} \lesssim h_T\|r\|_{L^2(T)}. \tag{85}$$

Since $L^2(\Omega)$ is a proper subspace of $H^{-1}(\Omega)$ the inverse inequality cannot hold for arbitrary r. Consequently, $h_T\|r\|_{L^2(T)}$ may overestimates $\|r\|_{H^{-1}(T)}$. On the other hand, if $r \in \mathbb{R}$ is *constant* and η denotes a non-negative function with properties

$$|T| \lesssim \int_T \eta, \quad \operatorname{supp}\eta = T, \quad \|\nabla\eta\|_{L^\infty(T)} \lesssim h_T^{-1} \tag{86}$$

(postpone the question of existence until (90) below), we deduce

$$\begin{aligned}\|r\|^2_{L^2(T)} &\lesssim \int_T r(r\eta) \le \|r\|_{H^{-1}(T)}\|\nabla(r\eta)\|_{L^2(T)}\\ &\le \|r\|_{H^{-1}(T)}\|r\|_{L^2(T)}\|\nabla\eta\|_{L^\infty(T)} \lesssim h_T^{-1}\|r\|_{H^{-1}(T)}\|r\|_{L^2(T)}.\end{aligned}$$

whence

$$h_T\|r\|_{L^2(T)} \lesssim \|r\|_{H^{-1}(T)}. \tag{87}$$

This shows that overestimation in (85) is caused by *oscillation* of r at a scale finer than the mesh-size. The estimate (87) is also valid for $r \in \mathbb{P}_l(T)$, but the constant deteriorates with the degree l; see Problem 34.

To conclude this discussion, we observe that $h_T\|r\|_{L^2(T)}$ can be easily approximated with the help of numerical integration, while this is not true for $\|r\|_{H^{-1}(T)}$. We therefore may say the weights are asymptotically acurate and that the possible overestimation of the weighted Lebesgue norms in (81) is the price for (almost) computability. This view is consistent with the fact that the indicators associated with the approximation of the Dirichlet boundary values in [62], which do not to invoke weighted Lebesgue norms, are overestimation-free.

6.3 Lower Bounds

The discussion in Sect. 6.2.3 suggests that $h_T\|r\|_{L^2(T)}$ bounds 'asymptotically' $\|\mathscr{R}\|_{H^{-1}(T)}$ from below. This is the main step towards a *local* lower bound for the

error. Such local lower bounds are the subject of this section. They do not contradict the global nature of the boundary value problem and their significance goes beyond a verification of the sharpness of the global upper bound (83).

For the sake of presentation, we present the case with polynomial degree $n = 1$ and leave the general case as problems to the reader.

6.3.1 Interior Residual

Let us start with a lower bound in terms of the interior residual and first check that $h_T \|r\|_{L^2(T)}$ bounds asymptotically $\|\mathscr{R}\|_{H^{-1}(T)}$ from below. To this end, we introduce the *oscillation of the interior residual* in T by

$$h_T \|r - \bar{r}_T\|_{L^2(T)},$$

where $\bar{r}_T$ denotes the mean value of r in T. Replacing r in (85) by $r - \bar{r}_T$ and in (87) by $\bar{r}_T$ as well as recalling (84), we derive

$$\begin{aligned} h_T \|r\|_{L^2(T)} &\leq h_T \|\bar{r}_T\|_{L^2(T)} + h_T \|r - \bar{r}_T\|_{L^2(T)} \\ &\lesssim \|\bar{r}_T\|_{H^{-1}(T)} + h_T \|r - \bar{r}_T\|_{L^2(T)} \\ &\lesssim \|r\|_{H^{-1}(T)} + \|r - \bar{r}_T\|_{H^{-1}(T)} + h_T \|r - \bar{r}_T\|_{L^2(T)} \\ &\lesssim \|\mathscr{R}\|_{H^{-1}(T)} + h_T \|r - \bar{r}_T\|_{L^2(T)}. \end{aligned} \tag{88}$$

This is the desired statement because the oscillation $h_T \|r - \bar{r}_T\|_{L^2(T)}$ is expected to convergence faster than $h_T \|r\|_{L^2(T)}$ under refinement. In the case $n = 1$ at hand there holds $r = f$ and, for example, there is one additional order if $f \in H^1(\Omega)$.

Since

$$\|\mathscr{R}\|_{H^{-1}(T)} = \sup_{w \in H^1_0(T)} \frac{\langle \mathscr{R}, w \rangle}{\|\nabla w\|_{L^2(T)}} = \sup_{w \in H^1_0(T)} \frac{\mathscr{B}[u - U, w]}{\|\nabla w\|_{L^2(T)}} \leq \alpha_2 \|\nabla(u - U)\|_{L^2(T)},$$

we have derived the following local lower bound

$$h_T \|r\|_{L^2(T)} \lesssim \alpha_2 \|\nabla(u - U)\|_{L^2(T)} + h_T \|r - \bar{r}_T\|_{L^2(T)}, \tag{89}$$

which also holds with $\bar{r}_T$ chosen from $\mathbb{P}_l(T)$ at the price of a larger constant hidden in $\lesssim$.

Finally we comment on the choice of the cut-off function $\eta_T \in W^1_\infty(T)$ with (86). For example, we may take

$$\eta_T = (d+1)^{d+1} \prod_{z \in \mathscr{V} \cap T} \lambda_z, \tag{90}$$

where λ_z, $z \in \mathscr{V} \cap T$, are the barycentric coordinates of T; see Lemma 3. This choice is due to Verfürth [75, 76]. Another choice, due to Dörfler [32], can be defined as

follows: refine T such that there appears an interior node and take the corresponding Courant basis function on the virtual triangulation of T; see Fig. 11 for the 2-dimensional case.

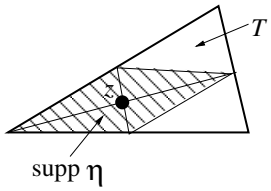

Fig. 11 Virtual refinement of a triangle for the Dörfler cut-off function.

The Dörfler cut-off function has the additional property that it is an element of a refined finite element space. This is not important here but useful when proving lower bounds for the differences of two discrete solutions. Such estimates are therefore called *discrete lower bound* whereas the bound for the true error is called *continuous lower bound*.

6.3.2 Jump Residual

We next strive for a local lower bound for the error in terms of the jump residual $h_S^{1/2}\|j\|_{L^2(S)}$, $S \in \mathring{\mathscr{S}}$, and use Sect. 6.3.1 on the interior residual as guideline.

We first notice that $j = [\![\boldsymbol{A}\nabla U]\!]$ is not necessary constant on an interior side $S \in \mathring{\mathscr{S}}$ due to the presence of A. We therefore introduce the oscillation of the jump residual in S:

$$h_S^{1/2}\|j - \bar{j}_S\|_{L^2(S)},$$

where $\bar{j}_S$ stands for the mean value of j on S. Notice that the important question about the order of this oscillation is not obvious because, in contrast to the oscillation of the element residual, the approximate solution U is involved. We postpone a corresponding discussion to Remark 23.

To choose a counterpart of η_T, let ω_S denote the patch composed of the two elements of $\mathscr{T}$ sharing S; see Fig. 12 for the 2-dimensional case. Obviously ω_S has a nonempty interior. Let $\eta_S \in W_\infty^1(\omega_S)$ be a cut-off function with the properties

$$|S| \lesssim \int_S \eta_S, \quad \operatorname{supp}\eta_S = \omega_S, \quad \|\eta_S\|_{L^\infty(\omega_S)} = 1, \quad \|\nabla\eta_S\|_{L^\infty(\omega_S)} \lesssim h_S^{-1}. \tag{91}$$

Following Verfürth [75, 76] we may take η_S given by

$$\eta_{S|T} = d^d \prod_{z \in \mathscr{V} \cap S} \lambda_z^T, \tag{92}$$

where $T \subset \omega_S$ and λ_z^T, $z \in \mathscr{V} \cap T$, are the barycentric coordinates of T. Also here

Fig. 12 Patch ω_S of triangles associated to interior side (left) and its refinement for Dörfler cut-off function (right).

Dörfler [32] proposed an alternative which is obtained as follows: refine ω_S such that there appears an interior node of S and take the corresponding Courant basis function on the virtual triangualtion of ω_S; see Fig. 12 for the 2-dimensional case.

After these preparations we are ready to derive a counterpart of (88). In view of the properties of η_S, we have

$$\|\bar{j}_S\|^2_{L^2(S)} \lesssim \int_S \bar{j}_S(\bar{j}_S\eta_S) = \int_S j\,\psi_S + \int_S (\bar{j}_S - j)\,\psi_S \tag{93}$$

with $\psi_S = \bar{j}_S\eta_S$. We rewrite the first term on the right hand side with the representation formula (76) as follows:

$$\int_S j\,\psi_S = -\int_{\omega_S} r\,\psi_S + \langle \mathscr{R},\, \psi_S\rangle,$$

where, in contrast to Sect. 6.3.1, the jump residual couples with the element residual. Hence

$$\left|\int_{\omega_S} j\,\psi_S\right| \le \|r\|_{L^2(\omega_S)}\|\psi_S\|_{L^2(\omega_S)} + \|\mathscr{R}\|_{H^{-1}(\omega_S)}\|\nabla\psi_S\|_{L^2(\omega_S)}.$$

In view of $|\omega_S| \lesssim h_S|S|$ and (91), we have

$$\|\psi_S\|_{L^2(\omega_S)} \le \|\bar{j}_S\|_{L^2(\omega_S)}\|\eta_S\|_{L^\infty(\omega_S)} \lesssim h_S^{1/2}\|\bar{j}_S\|_{L^2(S)}$$

and

$$\|\nabla\psi_S\|_{L^2(\omega_S)} \le \|\bar{j}_S\|_{L^2(\omega_S)}\|\nabla\eta_S\|_{L^\infty(\omega_S)} \lesssim h_S^{-1/2}\|\bar{j}_S\|_{L^2(S)}.$$

We infer that

$$\left|\int_{\omega_S} j\,\psi_S\right| \lesssim \left(h_S^{1/2}\|r\|_{L^2(\omega_S)} + h_S^{-1/2}\|\mathscr{R}\|_{H^{-1}(\omega_S)}\right)\|\bar{j}_S\|_{L^2(S)}.$$

In addition

$$\left|\int_S (\bar{j}_S - j)\,\psi_S\right| \le \|\bar{j}_S - j\|_{L^2(S)}\|\psi_S\|_{L^2(S)} \lesssim \|\bar{j}_S - j\|_{L^2(S)}\|\bar{j}_S\|_{L^2(S)}.$$

Inserting these estimates into (93) yields

$$\|\overline{j}_S\|^2_{L^2(S)} \lesssim \left(h_S^{1/2}\|r\|_{L^2(\omega_S)} + h_S^{-1/2}\|\mathscr{R}\|_{H^{-1}(\omega_S)} + \|\overline{j}_S - j\|_{L^2(S)}\right)\|\overline{j}_S\|_{L^2(S)}$$

whence, using (89) and $\|\mathscr{R}\|_{H^{-1}(T)} \le \|\mathscr{R}\|_{H^{-1}(\omega_S)}$ for $T \subset \omega_S$,

$$h_S^{1/2}\|j\|_{L^2(S)} \lesssim \|\mathscr{R}\|_{H^{-1}(\omega_S)} + \|h(r-\overline{r})\|_{L^2(\omega_S)} + \|h^{1/2}(\overline{j}-j)\|_{L^2(S)}, \tag{94}$$

where h denotes the mesh-size function from Sect. 6.2.2 and $\overline{r}$ and $\overline{j}$ are given by

$$\overline{r}_{|T} = \overline{r}_T \quad \text{for all } T \in \mathscr{T} \qquad \text{and} \qquad \overline{j}_{|S} = \overline{j}_S \quad \text{for all } S \in \mathring{\mathscr{S}}.$$

Since

$$\|\mathscr{R}\|_{H^{-1}(\omega_S)} \le \alpha_2 \|\nabla(u-U)\|_{L^2(\omega_S)},$$

we obtain the local lower bound in terms of the jump residual:

$$h_S^{1/2}\|j\|_{L^2(S)} \lesssim \alpha_2\|\nabla(u-U)\|_{L^2(\omega_S)} + \|h(r-\overline{r})\|_{L^2(\omega_S)} + \|h^{1/2}(\overline{j}-j)\|_{L^2(S)}. \tag{95}$$

Also this estimate holds with $\overline{j}$ piecewise polynomial of degree l; see Problem 39.

6.3.3 Local Lower Bound

We combine the two results on interior and jump residual and discuss its significance. To this end, we associate with each simplex $T \in \mathscr{T}$ the patch

$$\omega_T := \bigcup_{S \subset \partial T \setminus \partial\Omega} \omega_S,$$

see Fig. 13 for the 2-dimensional case, and define the oscillation in ω_T by

$$\operatorname{osc}(U, \omega_T) = \|h(r-\overline{r})\|_{L^2(\omega_T)} + \|h^{1/2}(j-\overline{j})\|_{L^2(\partial T\setminus\partial\Omega)}. \tag{96}$$

Recall that the higher order nature of $h_T\|r-\overline{r}_T\|_{L^2(T)}$ in (88) is crucial. We therefore compare the convergence order of (96) with that of the local error.

Remark 23 (On Asymptotics of Oscillation). For simplicity, we consider only polynomial degree $n = 1$, maximum convergence rates and suppose that $\boldsymbol{A}$ and f are smooth. One then expects that the local error vanishes like

$$\|\nabla(u-U)\|^2_{L^2(T)} = \mathscr{O}(h_T^{d+2})$$

and interior and jump residual oscillations like

$$\|h(r-\overline{r})\|^2_{L^2(\omega_T)} + \|h^{1/2}(j-\overline{j})\|^2_{L^2(\partial T\setminus\partial\Omega)} = \mathscr{O}(h_T^{d+4}).$$

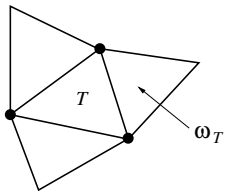

Fig. 13 Patch associated to a triangle in the local lower bound.

We already argued about the higher order of the interior residual after (88). Regarding the jump residual, the fact that ∇U is piecewise constant entails the identity

$$j - \overline{j}_S = [\![(\boldsymbol{A} - \overline{\boldsymbol{A}}_S)\nabla U]\!] = (\boldsymbol{A} - \overline{\boldsymbol{A}}_S)\boldsymbol{A}^{-1} j \quad \text{on an interior } S \in \mathring{\mathscr{S}},$$

which reveals the additional order for sufficiently smooth $\boldsymbol{A}$.

The oscillation $\operatorname{osc}(U, \omega_T)$ is therefore expected to be a higher order term for $h_T \downarrow 0$. However, as we shall see from the example in Remark 26 below, it may dominate on relatively coarse triangulations.

Similar arguments may be used to determine an appropriate polynomial degree of $\overline{j}_S$ and $\overline{r}_T$ in the case of general n. We do not insist on this and anticipate that in Chaps. 8 and 9 $\overline{r}_T$ will be the $L^2(T)$-best approximation in $P_{2n-2}(T)$ and $\overline{j}_S$ the $L^2(S)$-best approximation in $P_{2n-1}(S)$. This choices ensure, also for piecewise smooth A and f, that the oscillation is of higher order.

Since (89) and (95) hold also for piecewise polynomial $\overline{r}$ and $\overline{j}$, we have the following result for single indicators.

Theorem 14 (Local Lower Bound). *Let u and U be exact and Galerkin solution of the model problem and its standard discretization. Then, up to oscillation, each indicator is bounded by the local error:*

$$\mathscr{E}(U,T) \lesssim \alpha_2 \, \|\nabla(u-U)\|_{L^2(\omega_T)} + \operatorname{osc}(U, \omega_T) \qquad \textit{for all } T \in \mathscr{T}, \tag{97}$$

where α_2 is the largest eigenvalue of $\boldsymbol{A}(x)$ in ω_T and the hidden constant depends only on the shape coefficients of the simplices in ω_T, the dimension d and the polynomial degrees for $\overline{r}$ and $\overline{j}$.

Proof. Simply add the generalizations of (89) and (95) for all interior sides $S \in \mathring{\mathscr{S}}$ with $S \subset \partial T$. □

It is worthwhile to observe that in proving the local lower bound we have used the following the abstract notion of *local continuity* of the bilinear form $\mathscr{B}$. Let $\mathbb{V}$, $\mathbb{W}$ be normed spaces over Ω that are equipped with integral norms. If ω is a subdomain of Ω, then

$$\mathscr{B}[v, w] \le C_{\mathscr{B}} \|v\|_{\mathbb{V}(\omega)} \|w\|_{\mathbb{W}} \quad \text{for all } w \text{ with } w = 0 \text{ in } \Omega \setminus \overline{\omega}, \tag{98}$$

where $\|\cdot\|_{\mathbb{V}(\omega)}$ stands for the restriction of $\|\cdot\|_{\mathbb{V}}$ to ω. Obviously, the continuity constant $\|\mathscr{B}\|$ satisfies $\|\mathscr{B}\| \leq C_{\mathscr{B}}$ and therefore local continuity is stronger than global continuity. Property (98) readily implies an abstract local counterpart of the lower bound in Lemma 11.

We conclude this section with a remark about the importance of the fact that the lower bound in Theorem 14 is local and a remark about a simplifying setting in following chapters.

Remark 24 (Local Lower Bound and Marking). If $\operatorname{osc}(U,\omega_T) \ll \|\nabla(u-U)\|_{L^2(\omega_T)}$, as we expect asymptotically, then (97) translates into

$$\mathscr{E}(U,T) \lesssim \alpha_2 \|\nabla(u-U)\|_{L^2(\omega_T)}.$$

This means that an element T with relatively large error indicator contains a large portion of the error. To improve the solution U effectively, such T must be split giving rise to a procedure that tries to equidistribute errors. This is consistent with the discussion of adaptive approximation in 1d of Sect. 1.1 and constructive approximation of Chap. 5.

Remark 25 (Oscillation vs Data Oscillation). The quantity (96) measures oscillations of both interior residual r and jump residual j beyond the local mesh scale. Note that if U is piecewise affine and $\boldsymbol{A}(x)$ is piecewise constant, then

$$r = f + \operatorname{div}(\boldsymbol{A}\nabla U) = f \quad \text{and} \quad j = [\![\boldsymbol{A}\nabla U]\!]_S = \overline{j}_S.$$

Consequently

$$\operatorname{osc}(U,\omega_T) = \|h(f-\overline{f})\|_{L^2(\omega_T)}$$

becomes *data oscillation*, which is independent of the discrete solution U. Otherwise, for variable $\boldsymbol{A}$, osc depends on the discrete solution U. This additional dependence creates a nonlinear interaction in the adaptive algorithm and so leads to difficulties in characterizing an appropriate approximation class for adaptive methods, see Chap. 9.

6.3.4 Global Lower Bound and Equivalence

We derive a global lower bound from Theorem 14 and summarize the achievements of global nature in this chapter.

To formulate the global lower bound, we introduce the global oscillation

$$\operatorname{osc}(U,\mathscr{T}) = \|h(r-\overline{r})\|_{L^2(\Omega)} + \|h^{1/2}(\overline{j}-j)\|_{L^2(\Gamma)}, \tag{99}$$

recalling that Γ is the interior skeleton of $\mathscr{T}$. By summing (97) over all $T \in \mathscr{T}$ and taking into account (55), which entails a finite overlapping of the patches ω_T, we obtain the following global result.

Corollary 9 (Global Lower Bound). *Let u and U be exact and Ritz-Galerkin solutions of the model problem and its standard discretization. Then there holds the following global lower bound:*

$$\mathscr{E}(U,\mathscr{T}) \lesssim \alpha_2 \|\nabla(u-U)\|_{L^2(\Omega)} + \operatorname{osc}(U,\mathscr{T}) \tag{100}$$

where α_2 is the largest global eigenvalues of $\boldsymbol{A}$ and the hidden constant depends on the shape coefficient of $\mathscr{T}$, the dimension d, and the polynomial degrees for $\bar{r}$ and $\bar{j}$.

As already alluded to in Sect. 6.2.3, the presence of $\operatorname{osc}(U,\mathscr{T})$ in the lower bound is the price to pay for having a simple and computable estimator $\mathscr{E}(U,\mathscr{T})$. In the following remark, we present an example that shows that $\operatorname{osc}(U,\mathscr{T})$ cannot be removed from (100).

Remark 26 (Necessity of oscillation). Let $\varepsilon = 2^{-K}$ for K integer and extend the function $\frac{1}{2}x(\varepsilon - |x|)$ defined on $(-\varepsilon,\varepsilon)$ to a 2ε-periodic C^1 function u_ε on $\Omega = (-1,1)$. Moreover, let the forcing function be $f_\varepsilon = -u''$, which is 2ε-periodic and piecewise constant with values ± 1 that change at multiples of ε; see Fig. 14. Let $\mathscr{T}$ be a uni-

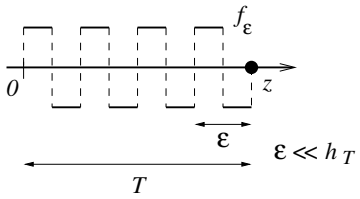

Fig. 14 An strongly oscillating forcing function.

form mesh with mesh-size $h = 2^{-k}$, with $k \ll K$. We consider piecewise linear finite elements $\mathbb{V}(\mathscr{T}_\varepsilon)$ and corresponding Galerkin solution $U_\varepsilon \in \mathbb{V}(\mathscr{T}_\varepsilon)$. It is easy to verify that f_ε is L^2-orthogonal to both the space of piecewise constants and linears over $\mathscr{T}_\varepsilon$, whence $U_\varepsilon = \bar{f}_\varepsilon = 0$ and

$$\begin{aligned}\|u_\varepsilon' - U_\varepsilon'\|_{L^2(\Omega)} = \|u_\varepsilon'\|_{L^2(\Omega)} &= \frac{\varepsilon}{\sqrt{6}} = \frac{2^{-K}}{\sqrt{6}} \\ &\ll 2^{-k} = h = \|hf_\varepsilon\|_{L^2(\Omega)} = \operatorname{osc}(U_\varepsilon,\mathscr{T}) = \mathscr{E}(U_\varepsilon,\mathscr{T}).\end{aligned}$$

Therefore, the ratio $\|u_\varepsilon' - U_\varepsilon'\|_{L^2(\Omega)}/\mathscr{E}(U_\varepsilon,\mathscr{T})$ can be made arbitrarily small by increasing K/k, and $\operatorname{osc}(U_\varepsilon,\mathscr{T})$ accounts for the discrepancy. On the other hand, measuring the oscillation in $H^{-1}(\Omega)$, as suggested in [13, 69],

$$\|f_\varepsilon - \bar{f}_\varepsilon\|_{H^{-1}(\Omega)} = \|f_\varepsilon\|_{H^{-1}(\Omega)} = \|u_\varepsilon'\|_{L^2(\Omega)} \approx \varepsilon,$$

would avoid overestimation but brings us back to the question how to (approximately) evaluate the $H^{-1}(\Omega)$-norm at acceptable cost.

This 1d example can be extended via a checkerboard pattern to any dimension.

We see that $\operatorname{osc}(U,\mathscr{T})$ may be dominant in early stages of the adaptive iteration (4). Therefore, it cannot be ignored in an optimality analysis without fineness assumptions on the initial mesh $\mathscr{T}_0$; compare with Example 3.

We conclude by combing the two global bounds in Theorem 13 and Corollary 9.

Theorem 15 (Asymptotic Equivalence). *Let u and U be exact and Galerkin solutions of the model problem and its standard discretization. Then the error estimator* (82) *is asymptotically equivalent to the error:*

$$\frac{1}{\alpha_2}\Big(\mathscr{E}(U,\mathscr{T})-\operatorname{osc}(U,\mathscr{T})\Big)\lesssim\|\nabla(u-U)\|_{L^2(\Omega)}\lesssim\frac{1}{\alpha_1}\mathscr{E}(U,\mathscr{T}) \tag{101}$$

where $0<\alpha_1\leq\alpha_2$ are the smallest and largest global eigenvalues of $\boldsymbol{A}$ and the hidden constants depend only on the shape coefficient of $\mathscr{T}$, the dimension d and the polynomial degrees for $\overline{r}$ and $\overline{j}$.

We thus have derived a computable quantity that may be used to stop the adaptive iteration (4) and, in view of the local lower bound in Sect. 6.3.3, the indicators may be used to provide the problem-specific information for local refinement.

6.4 Problems

Problem 29. The gap in (75) is dictated by $\|\mathscr{B}\|/\alpha$. Determine this quantity for the model problem in Sect. 2.2.1 and

(a) $\|v\|_{\mathbb{V}}=|v|_{1,\Omega}$,
(b) $\|v\|_{\mathbb{V}}=\big(\int_\Omega\nabla v\cdot A\nabla v\big)^{1/2}$.

Problem 30. Prove the scaled trace inequality (Corollary 8)

$$\|v\|_{L^2(S)}\lesssim h_S^{-1/2}\|v\|_{L^2(T)}+h_S^{1/2}\|\nabla v\|_{L^2(T)}\quad\text{for all } v\in H^1(T).$$

Problem 31. Show that, up to oscillation terms, the jump residual

$$\eta_{\mathscr{J}}(U,\mathscr{T})=\Big(\sum_{S\in\mathscr{S}}\|h^{1/2}j\|^2_{L^2(S)}\Big)^{1/2}$$

bounds $\|\mathscr{R}\|_{H^{-1}(\Omega)}$, which entails that the estimator $\mathscr{E}(U,\mathscr{T})$ is dominated by $\eta_{\mathscr{J}}(U,\mathscr{T})$. To this end, revise the proof of the upper bound for $\|\mathscr{R}\|_{H^{-1}(\Omega)}$, use

$$c_z=\frac{1}{\int_{\omega_z}\phi_z}\int_{\omega_z}r\,\phi_z.$$

and rewrite $\int_{\omega_z}r\,(w-c_z)\,\phi_z$ by exploiting this weighted L^2-orthogonality.

Problem 32. Considering the model problem with its standard discretization, derive the upper a posteriori error bound without using the discrete partition of unity. To this end use (76) and combine the scaled trace inequality (77) with the local interpolation error estimate (63). Derive as an intermediate step the upper bound:

$$|\langle \mathscr{R}, w\rangle| \leq \sum_{T\in\mathscr{T}} \mathscr{E}(U,T)\|\nabla w\|_{L^2(N(T))},$$

with $N(T)$ from (55). Discuss the differences of the two derivations.

This form of the upper bound is useful in Chap. 7.

Problem 33. Verify that a suitable multiple of the Verfürth cut-off function (90) satisfies the properties (86). To this end, recall Lemma 1. Repeat for the Dörfler cut-off function.

Problem 34. (Try this problem after Problem 33.) Show that the choice (90) for η_T verifies, for all $p \in \mathbb{P}_l(T)$,

$$\int_T p^2 \lesssim \int_T p^2\eta_T, \qquad \|\nabla(p\eta_T)\|_{L^2(T)} \lesssim h_T^{-1}\|p\|_{L^2(T)}$$

with constants depending on l and the shape coefficient of T. To this end, recall the equivalence of norms in finite-dimensional spaces. Derive the estimate

$$h_T\,\|r\|_{L^2(T)} \lesssim \|r\|_{H^{-1}(T)}$$

for $r \in \mathbb{P}_l(T)$.

Problem 35. Consider the model problem and its discretization for $d = 2$ and $n = 1$. Let U_1 be the solution over a triangulation $\mathscr{T}_1$ and U_2 the solution over $\mathscr{T}_2$, where $\mathscr{T}_2$ has been obtained by applying at least 3 bisections to every triangle of $\mathscr{T}_1$. Moreover, suppose that f is piecewise constant over $\mathscr{T}_1$. Show

$$\|\nabla(U_2 - U_1)\|_{L^2(\Omega)} \geq \|h_1 f\|_{L^2(\Omega)},$$

where h_1 is the mesh-size function of $\mathscr{T}_1$.

Problem 36. Verify that a suitable multiple of the Verfürth cut-off function (92) satisfies the properties (91). Repeat for the Dörfler cut-off function.

Problem 37. Let S be a side of a simplex T. Show that for each $q \in \mathbb{P}_l(S)$ there exists a $p \in \mathbb{P}_l(T)$ such that

$$p = q \text{ on } S \qquad \text{and} \qquad \|p\|_{L^2(T)} \lesssim h_T^{1/2}\|q\|_{L^2(S)}.$$

Problem 38. Let S be a side of a simplex T. Show that the choice (92) for η_S verifies, for all $q \in \mathbb{P}_k(\mathscr{S})$ and all $p \in \mathbb{P}_l(T)$,

$$\int_S q^2 \lesssim \int_S q^2\eta_S, \qquad \|\nabla(p\eta_S)\|_{L^2(T)} \lesssim h_T^{-1}\|p\|_{L^2(T)}$$

with constants depending on l and the shape coefficient of T.

Problem 39. Derive the estimate (95), where $\bar{r}$ and $\bar{j}$ are piecewise polynomials of degree l_1 and l_2.

Problem 40. Generalize Remark 23 to polynomial degree $n \geq 2$.

Problem 41. Supposing (98), formulate and prove an abstract local lower bound in the spirit of Lemma 11.

Problem 42. Derive a posteriori error bounds for the energy norm

$$|||v|||_{\Omega} = \left(\int_{\Omega} \nabla v \cdot \boldsymbol{A} \nabla v \right)^{1/2}$$

and compare with Theorem 15.

7 Adaptivity: Convergence

The purpose of this chapter is to prove that the standard adaptive finite element method characterized by the iteration

$$\mathsf{SOLVE} \longrightarrow \mathsf{ESTIMATE} \longrightarrow \mathsf{MARK} \longrightarrow \mathsf{REFINE} \tag{102}$$

generates a sequence of discrete solutions converging to the exact one. This will be established under assumptions that are quite weak or even minimal. In particular, we will not suppose any regularity of the exact solution that goes beyond the natural one in the variational formulation. We therefore can expect only a plain convergence result that does not give any convergence rate in terms of degrees of freedom. The assumptions on the general variational problem allow for various examples that are of quite different from the model problem in Sect. 2.2.1. Examples are left as problems to the reader.

The presentation is based on the basic convergence result by Morin et al. [55] and the modifications by Siebert [67].

7.1 The Adaptive Algorithm

Given a continuous bilinear form $\mathscr{B}\colon \mathbb{V}\times\mathbb{W}\to\mathbb{R}$ and an element $f\in\mathbb{W}^*$ we consider the variational problem

$$u\in\mathbb{V}: \quad \mathscr{B}[u,w]=\langle f,w\rangle \qquad \text{for all } w\in\mathbb{W} \tag{103}$$

introduced in Chap. 2. We assume that $\mathscr{B}$ satisfies the inf-sup condition (21).

For the adaptive approximation of the solution u we consider a loop of the form (102). To be more precise, starting with an initial conforming triangulation $\mathscr{T}_0$ of the underlying domain Ω and a refinement procedure REFINE as described in Sect. 4.4 we execute an iteration of the following main steps:

$$\begin{aligned}
&(1)\quad U_k := \mathsf{SOLVE}\big(\mathbb{V}(\mathscr{T}_k),\mathbb{W}(\mathscr{T}_k)\big);\\
&(2)\quad \{\mathscr{E}_k(U_k,T)\}_{T\in\mathscr{T}_k} := \mathsf{ESTIMATE}\big(U_k,\mathscr{T}_k\big);\\
&(3)\quad \mathscr{M}_k := \mathsf{MARK}\big(\{\mathscr{E}_k(U_k,T)\}_{T\in\mathscr{T}_k},\mathscr{T}_k\big);\\
&(4)\quad \mathscr{T}_{k+1} := \mathsf{REFINE}\big(\mathscr{T}_k,\mathscr{M}_k\big), \text{ increment } k \text{ and go to Step (1).}
\end{aligned} \tag{104}$$

In practice, a stopping test is used after Step (2) for terminating the iteration; here we shall ignore it for notational convenience. Besides the initial grid $\mathscr{T}_0$ and the module REFINE from Sect. 4.4, the realization of these steps requires the following objects and modules:

- For any grid $\mathscr{T} \in \mathbb{T}$, there are finite element spaces $\mathbb{V}(\mathscr{T})$ and $\mathbb{W}(\mathscr{T})$ and the module SOLVE outputs the corresponding Petrov-Galerkin approximation $U_{\mathscr{T}}$ to u.
- A module ESTIMATE that, given a grid $\mathscr{T} \in \mathbb{T}$ and the corresponding discrete solution $U_{\mathscr{T}}$, outputs the a posteriori error estimator $\{\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T)\}_{T\in\mathscr{T}}$, where the so-called indicator $\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T) \geq 0$ is associated with the element $T \in \mathscr{T}$.
- A strategy in the module MARK that, based upon the a posteriori error indicators $\{\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T)\}_{T\in\mathscr{T}}$, collects elements of $\mathscr{T}$ in $\mathscr{M}$, which serves as input for REFINE.

Obviously, the modules SOLVE and ESTIMATE do strongly depend on the variational problem, i. e., on data $\mathscr{B}$ and f; compare with Sects. 3.1.3 and 6. For convenience of notation we have suppressed this dependence. The refinement module REFINE is problem independent and the same applies in general to the module MARK. We list the most popular marking strategies for (104):

(a) Maximum Strategy: For given parameter $\theta \in [0,1]$ we let

$$\mathscr{M} = \big\{T \in \mathscr{T} \mid \mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T) \geq \theta \mathscr{E}_{\mathscr{T},\max}\big\} \quad \text{with} \quad \mathscr{E}_{\mathscr{T},\max} = \max_{T\in\mathscr{T}} \mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T).$$

(b) Equidistribution Strategy: For given parameter $\theta \in [0,1]$ we let

$$\mathscr{M} = \Big\{T \in \mathscr{T} \mid \mathscr{E}_{\mathscr{T}}(U_k,T) \geq \theta \mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},\mathscr{T})/\sqrt{\#\mathscr{T}}\Big\}.$$

(c) Dörfler's Strategy: For given parameter $\theta \in (0,1]$ we let $\mathscr{M} \subset \mathscr{T}$ such that

$$\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},\mathscr{M}) \geq \theta \mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},\mathscr{T}).$$

For efficiency reasons one wants to mark as few elements as possible. This can be achieved by selecting the elements holding the largest indicators, whence

$$\min_{T\in\mathscr{M}} \mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T) \geq \max_{T\in\mathscr{T}\setminus\mathscr{M}} \mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T).$$

The objective of this chapter is to prove that, under quite weak assumptions on the modules SOLVE, ESTIMATE, and MARK, the sequence $\{U_k\}_{k\geq 0}$ of discrete solutions converges to u, i. e.,

$$\lim_{k\to\infty} \|U_k - u\|_{\mathbb{V}} = 0. \tag{105}$$

This is a priori not clear, since the estimator only provides a global upper bound for the error. All the techniques used in Chap. 5 are based on completely local interpolation estimates and therefore cannot be used when working with an estimator. Then again, as long as $U_k \neq u$ the estimator is non-zero. This should lead to convergence provided that the indicators $\{\mathscr{E}_k(U_k,T)\}_{T\in\mathscr{T}_k}$ pick up some local error information and the selection of elements in MARK accounts for that.

For convenience of notation we replace in what follows the argument $\mathscr{T}_k$ by a subscript k, for instance we set $\mathbb{V}_k := \mathbb{V}(\mathscr{T}_k)$.

7.2 Density and Convergence

Plain convergence for a sequence of uniformly refined grids is a simple consequence of density. To see this, we set $\mathscr{M}_k = \mathscr{T}_k$ in each iteration of (104). Then (61) implies

$$h_{\max}(\mathscr{T}_k) := \max\{h_T \mid T \in \mathscr{T}_k\} \leq D_2\, 2^{-kb/d} \to 0 \qquad \text{as } k \to \infty,$$

if elements in $\mathscr{M}_k$ are scheduled for $b \geq 1$ bisections. Furthermore, let $\mathbb{V}^s \subset \mathbb{V}$ be a dense subspace and $I_k \colon \mathbb{V}^s \to \mathbb{V}_k$ be an interpolation operator with

$$\|I_k v - v\|_{\mathbb{V}} \leq C\, h_{\max}^s(\mathscr{T}_k)\|v\|_{\mathbb{V}^s} \qquad \text{for all } v \in \mathbb{V}^s \tag{106}$$

for $s > 0$. In case of the model problem with $\mathbb{V} = H_0^1(\Omega)$ we could take for instance $\mathbb{V}_k$ to be conforming Lagrange finite elements over $\mathscr{T}_k$, and I_k the Lagrange interpolant, which satisfies (106) with $s = 1$ on $\mathbb{V}^2 = H^2(\Omega) \cap H_0^1(\Omega)$; compare with Remark 18. For any $v \in \mathbb{V}$ and $\bar{v} \in \mathbb{V}^s$ we then derive

$$\|I_k\bar{v} - v\|_{\mathbb{V}} \leq \|I_k\bar{v} - \bar{v}\|_{\mathbb{V}} + \|\bar{v} - v\|_{\mathbb{V}} \leq C\, h_{\max}^s(\mathscr{T}_k)\|\bar{v}\|_{\mathbb{V}^s} + \|\bar{v} - v\|_{\mathbb{V}}.$$

For given v and ε we first can choose $\bar{v} \in \mathbb{V}^s$ such that $\|v - \bar{v}\|_{\mathbb{V}} \leq \varepsilon/2$ by density of $\mathbb{V}^s$ in $\mathbb{V}$. Then (106) implies $C\, h_{\max}^s(\mathscr{T}_k)\|\bar{v}\|_{\mathbb{V}^s} \leq \varepsilon/2$ provided k is sufficiently large, whence $\|I_k\bar{v} - v\|_{\mathbb{V}} \leq \varepsilon$. Therefore,

$$\lim_{k\to\infty} \min_{V_k \in \mathbb{V}_k} \|V_k - v\|_{\mathbb{V}} = 0 \qquad \text{for all } v \in \mathbb{V}$$

or, equivalently,

$$\mathbb{V} = \overline{\bigcup_{k\geq 0} \mathbb{V}_k}. \tag{107}$$

This density property already implies convergence if the sequence $\{\mathbb{V}_k, \mathbb{W}_k\}_{k\geq 0}$ is stable, i. e., it satisfies a uniform inf-sup condition. Recalling the quasi-best approximation property of the Petrov-Galerkin solution U_k established in Theorem 5, stability of the discretization yields

$$\|U_k - u\|_{\mathbb{V}} \leq \frac{\|\mathscr{B}\|}{\beta} \min_{V_k \in \mathbb{V}_k} \|V_k - u\|_{\mathbb{V}} \to 0 \qquad \text{as } k \to \infty, \tag{108}$$

thanks to density (107). Note, that this convergence result holds true irrespective of any regularity property of u beyond $\mathbb{V}$.

Assume now that the sequence $\{\mathscr{T}_k\}_{k\geq 0}$ is adaptively generated. We observe that (107) still holds whenever

$$\lim_{k\to\infty} h_{\max}(\mathscr{T}_k) = 0, \tag{109}$$

whence (108) is also true. But (109) does not hold in general for an adaptively generated sequence of meshes, as was already observed by Babuška and Vogelius [7]. Recalling the definition of the mesh-size function

$$h_k \in L^\infty(\Omega): \qquad h_{k|T} = |T|^{1/d}, \quad T \in \mathscr{T}_k$$

in Sect. 4.3 and its L^∞ limit $h_\infty \in L^\infty(\Omega)$ of Lemma 5, Eq. (109) is equivalent to $h_\infty \equiv 0$ in Ω. If $h_\infty \not\equiv 0$, then there exists an $x \in \Omega \setminus \Gamma_\infty$ with $h_\infty(x) > 0$. This implies that there is an element $T \ni x$ and an iteration counter $K = K(x)$ such that $T \in \mathscr{T}_k$ for all $k \geq K$.

This motivates to split the triangulations $\mathscr{T}_k$ into two classes of elements

$$\mathscr{T}_k^+ := \bigcap_{\ell \geq k} \mathscr{T}_k = \{T \in \mathscr{T}_k \mid T \in \mathscr{T}_\ell \; \forall \ell \geq k\}, \quad \text{and} \quad \mathscr{T}_k^0 := \mathscr{T}_k \setminus \mathscr{T}_k^+. \tag{110}$$

The set $\mathscr{T}_k^+$ contains all elements that are not refined after iteration k and we observe that the sequence $\left\{\mathscr{T}_k^+\right\}_{k\geq 0}$ is nested, i. e., $\mathscr{T}_\ell^+ \subset \mathscr{T}_k^+$ for all $k \geq \ell$. The set $\mathscr{T}_k^0$ contains all elements that are refined at least once more in a forthcoming step of the adaptive procedure. Since the sequence $\{\mathscr{T}_k^+\}_{k\geq 0}$ is nested the set

$$\mathscr{T}^+ := \bigcup_{k\geq 0} \mathscr{T}_k^+$$

is well-defined and we conclude

$$h_\infty \equiv 0 \qquad \text{if and only if} \qquad \mathscr{T}^+ = \emptyset.$$

If $\mathscr{T}^+ \neq \emptyset$ then the finite element spaces cannot be dense in $\mathbb{V}$ since inside $T \in \mathscr{T}^+$ we can only approximate discrete functions. Therefore, taking into accout the arguments at the beginning of the section, we have that (107) is equivalent to $h_\infty \equiv 0$.

On the other hand, when using adaptivity we do not aim at approximating *all* functions in $\mathbb{V}$ but rather one single function, namely the solution u to (103). A necessary condition for being able to approximate u is

$$\lim_{k\to\infty} \min_{V_k \in \mathbb{V}_k} \|u - V_k\|_{\mathbb{V}} = 0.$$

Assuming that the finite element spaces are nested, the space

$$\mathbb{V}_\infty := \overline{\bigcup_{k\geq 0} \mathbb{V}_k}$$

is well-defined and we can approximate u by discrete functions if and only if $u \in \mathbb{V}_\infty$. We realize that $\mathbb{V}_\infty$ is defined via the adaptively generated spaces $\mathbb{V}_k$. Therefore, $u \in \mathbb{V}_\infty$ hinges on properties of the modules SOLVE, ESTIMATE, MARK, and REFINE. In addition, if $\mathbb{V}_\infty$ is a proper subspace of $\mathbb{V}$ and $u \in \mathbb{V}_\infty$ then u is locally a discrete function. This implies, that the adaptive method must only decide not

to refine an element any more if u locally belongs to the finite element space, for instance u is affine in some part of the domain in case of Courant elements.

But this is not the generic case. If u is not locally discrete, then the decisions of the adaptive method have to yield $\mathscr{T}^+ = \emptyset$, and if so, convergence is a direct consequence of density as for uniform refinement. We aim at a convergence result for adaptive finite elements that just relies on this density argument in this case. In doing this we shall use a *local density* property of the finite element spaces in the region $\{h_\infty \equiv 0\}$ and properties of the adaptive method in its complement.

7.3 Properties of the Problem and the Modules

In this section we state structural assumptions on the Hilbert spaces $\mathbb{V}$ and $\mathbb{W}$ and the modules SOLVE, ESTIMATE, and MARK. For notational convenience we use '$a \lesssim b$' for '$a \le Cb$' whenever the constant C only depends on $\mathscr{T}_0$ and data of (103) like $\mathscr{B}$ and f.

7.3.1 Properties of Hilbert Spaces

We assume that $\mathbb{V}$ is a subspace of $L^2(\Omega;\mathbb{R}^m)$ with some $m \in \mathbb{N}$ and that $\|\cdot\|_{\mathbb{V}}$ is an L^2-type integral norm implying the following properties: The square of the norm $\|\cdot\|_{\mathbb{V}(\Omega)}$ is set-additive, i. e., for any subset $\omega \subset \Omega$ that is decomposed into $\omega = \omega_1 \cup \omega_2$ with $|\omega_1 \cap \omega_2| = 0$ there holds

$$\|v\|^2_{\mathbb{V}(\omega)} = \|v\|^2_{\mathbb{V}(\omega_1)} + \|v\|^2_{\mathbb{V}(\omega_2)} \qquad \text{for all } v \in \mathbb{V}(\omega). \tag{111}$$

In addition, we ask $\|\cdot\|_{\mathbb{V}}$ to be absolutely continuous with respect to the Lebesgue measure, this is, for any $v \in \mathbb{V}$ holds

$$\|v\|_{\mathbb{V}(\omega)} \to 0 \quad \text{as} \quad |\omega| \to 0.$$

Finally we require $\mathbb{W}$ to have the same properties.

7.3.2 Properties of SOLVE

For any grid $\mathscr{T} \in \mathbb{T}$ we assume the existence of a pair of finite element spaces $\{\mathbb{V}(\mathscr{T}), \mathbb{W}(\mathscr{T})\}$ and suppose the following properties:

(1) They are conforming

$$\mathbb{V}(\mathscr{T}) \subset \mathbb{V}, \quad \mathbb{W}(\mathscr{T}) \subset \mathbb{W} \qquad \text{for all } \mathscr{T} \in \mathbb{T} \tag{112a}$$

and nested

$$\mathbb{V}(\mathscr{T}) \subset \mathbb{V}(\mathscr{T}_*), \quad \mathbb{W}(\mathscr{T}) \subset \mathbb{W}(\mathscr{T}^*) \qquad \text{for all } \mathscr{T} \leq \mathscr{T}_* \in \mathbb{T}. \tag{112b}$$

(2) The finite element spaces are a stable discretization, i. e., there exists $\beta > 0$ such that for all $\mathscr{T} \in \mathbb{T}$

$$\dim \mathbb{V}(\mathscr{T}) = \dim \mathbb{W}(\mathscr{T}) \quad \text{and} \quad \inf_{\substack{V \in \mathbb{V}(\mathscr{T}) \\ \|V\|_{\mathbb{V}}=1}} \sup_{\substack{W \in \mathbb{W}(\mathscr{T}) \\ \|W\|_{\mathbb{W}}=1}} \mathscr{B}[V, W] \geq \beta. \tag{112c}$$

(3) Let $\mathbb{W}^s \subset \mathbb{W}$ be a dense sub-space with norm $\|\cdot\|_{\mathbb{W}^s}$ such that $\|\cdot\|^2_{\mathbb{W}^s}$ is set-additive and let $I_{\mathscr{T}} \in L(\mathbb{W}^s, \mathbb{W}(\mathscr{T}))$ be a continuous, linear interpolation operator such that

$$\|w - I_{\mathscr{T}} w\|_{\mathbb{W}(T)} \lesssim h_T^s \|w\|_{\mathbb{W}^s(T)} \qquad \text{for all } T \in \mathscr{T} \text{ and } w \in \mathbb{W}^s \tag{112d}$$

with $s > 0$.

(4) We suppose that $\mathsf{SOLVE}\big(\mathbb{V}(\mathscr{T}), \mathbb{W}(\mathscr{T})\big)$ outputs the *exact* Petrov-Galerkin approximation of u, i. e.,

$$U_{\mathscr{T}} \in \mathbb{V}(\mathscr{T}): \qquad \mathscr{B}[U_{\mathscr{T}}, W] = \langle f, W \rangle \qquad \text{for all } w \in \mathbb{W}(\mathscr{T}). \tag{112e}$$

This entails exact integration and linear algebra; see Remarks 9 and 10.

Note, that for non-adaptive realizations of (104), condition (112c) is necessary for the well-posedness of (112e) and convergence irrespective of $f \in \mathbb{W}^*$; compare with Problem 12. Although phrasing the interpolation estimate (112d) as a condition on the choice of the finite element space, the construction of any finite element space is based on such a local approximation property.

7.3.3 Properties of ESTIMATE

Given a grid $\mathscr{T} \in \mathbb{T}$ and the Petrov-Galerkin approximation $U_{\mathscr{T}} \in \mathbb{V}_{\mathscr{T}}$ of (112e) we suppose that we can *compute* a posteriori error indicators $\{\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}}, T)\}_{T \in \mathscr{T}}$ by

$$\{\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}}, T)\}_{T \in \mathscr{T}} = \mathsf{ESTIMATE}(U_{\mathscr{T}}, \mathscr{T})$$

with the following properties:

(1) The estimator provides the following upper bound for the residual $\mathscr{R}_{\mathscr{T}} \in \mathbb{W}^*$ of $U_{\mathscr{T}}$:

$$|\langle \mathscr{R}_{\mathscr{T}}, w \rangle| \lesssim \sum_{T \in \mathscr{T}} \mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}}, T) \|w\|_{\mathbb{W}(N_{\mathscr{T}}(T))} \qquad \text{for all } w \in \mathbb{W}. \tag{113a}$$

(2) The estimator is efficient in that it satisfies the continuous local lower bound

$$\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}}, T) \lesssim \|U_{\mathscr{T}} - u\|_{\mathbb{V}(N_{\mathscr{T}}(T))} + \operatorname{osc}_{\mathscr{T}}(U_{\mathscr{T}}, T) \qquad \text{for all } T \in \mathscr{T}, \tag{113b}$$

where the oscillation indicator $\operatorname{osc}_{\mathscr{T}}(U_{\mathscr{T}}, T)$ satisfies

$$\operatorname{osc}_{\mathscr{T}}(U_{\mathscr{T}},T) \lesssim h_T^q \Big(\|U_{\mathscr{T}}\|_{\mathbb{V}(N_{\mathscr{T}}(T))} + \|D\|_{L^2(N_{\mathscr{T}}(T))} \Big). \tag{113c}$$

Hereafter, $q > 0$ and $D \in L^2(\Omega)$ is given by data of (103).

The upper bound as stated in (113a) is usually an intermediate step when deriving a posteriori error estimates; compare with Problem 32. It allows us to extract *local* information about the residual. This is not possible when directly using the global upper bound $\|U_{\mathscr{T}} - u\|_{\mathbb{V}(\Omega)} \lesssim \mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}}, \mathscr{T})$.

7.3.4 Properties of MARK

The last module for the adaptive algorithm is a function

$$\mathscr{M} = \mathsf{MARK}\big(\{\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T)\}_{T\in\mathscr{T}}, \mathscr{T}\big)$$

that, given a mesh $\mathscr{T} \in \mathbb{T}$ and indicators $\{\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T)\}_{T\in\mathscr{T}}$, selects elements subject to refinement. Given a fixed function $g : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ that is continuous at 0 with $g(0) = 0$, we ask that the set $\mathscr{M}$ of marked elements has the property

$$\max\{\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T) \mid T \in \mathscr{T} \setminus \mathscr{M}\} \le g\big(\max\{\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T) \mid T \in \mathscr{M}\}\big). \tag{114}$$

Marking criterion (114) implies that all indicators in $\mathscr{T}$ are controlled by the maximal indicator in $\mathscr{M}$. Marking strategies that pick up the elements holding the largest indicator, as those from Sect. 7.1, satisfy (114) with $g(s) = s$.

7.4 Convergence

In this section we show that the realization of (104) generates a sequence of Petrov-Galerkin solutions that converges to the true solution in $\mathbb{V}$ under the above assumptions.

Theorem 16 (Convergence). *Let u be the exact solution of* (103) *and suppose that* (21) *holds. Let the finite element spaces and the functions* SOLVE, ESTIMATE, *and* MARK *satisfy* (112), (113), *and* (114), *respectively.*

Then the sequence of Galerkin approximations $\{U_k\}_{k\ge0}$ generated by iteration (104) *satisfies*

$$\lim_{k\to\infty} \|U_k - u\|_{\mathbb{V}} = 0 \qquad \textit{and} \qquad \lim_{k\to\infty} \mathscr{E}_k(U_k, \mathscr{T}_k) = 0.$$

In particular, any prescribed tolerance $\mathsf{TOL} > 0$ *for the estimator is reached in a finite number of steps. In other words: there is an iteration k^* with*

$$\|U_{k^*} - u\|_{\mathbb{V}} \lesssim \mathscr{E}_{k^*}(U_{k^*}, \mathscr{T}_{k^*}) \le \mathsf{TOL}.$$

We split the proof in several steps.

7.4.1 Two limits

In this paragraph we give another generalization of (56) for a sequence of adaptively generated triangulations. In combination with the interpolation estimate (112d) this result yields a local density property of adaptively generated finite element spaces. Additionally we show that for any realization of (104) the Petrov-Galerkin solutions are a Cauchy-sequence in $\mathbb{V}$.

The uniform convergence $h_k \to h_\infty$ shown in Lemma 5 helps to locate the set $\{h_\infty \equiv 0\}$ in terms of the splitting $\mathscr{T}_k = \mathscr{T}_k^0 \cup \mathscr{T}_k^+$ introduced in (110). According to $\mathscr{T}_k^0$ and $\mathscr{T}_k^+$ we decompose the domain Ω into

$$\bar{\Omega} = \Omega(\mathscr{T}_k^+) \cup \Omega(\mathscr{T}_k^0) =: \Omega_k^+ \cup \Omega_k^0,$$

where for any sub-triangulation $\mathscr{T}_k' \subset \mathscr{T}_k$ we let

$$\Omega(\mathscr{T}_k') := \bigcup \{T : T \in \mathscr{T}_k'\}$$

be the part of Ω covered by $\mathscr{T}_k'$. A direct consequence of Lemma 5 is the following result.

Corollary 10 ($\{h_\infty \equiv 0\}$)**.** *Denote by χ_k^0 the characteristic function of Ω_k^0.*
Then the definition of $\mathscr{T}_k^0$ implies

$$\lim_{k\to\infty} \left\| h_k \chi_k^0 \right\|_{L^\infty(\Omega)} = \lim_{k\to\infty} \|h_k\|_{L^\infty(\Omega_k^0)} = 0.$$

Proof. The definition of $\mathscr{T}_k^0$ implies that all elements in $\mathscr{T}_k^0$ are refined at least once. Hence, $h_\infty \le 2^{-\frac{1}{d}} h_k$ in Ω_k^0, yielding $\left(1 - 2^{-1/d}\right) h_k \le h_k - h_\infty$ in Ω_k^0. This in turn implies with $\gamma = 1 - 2^{1/d} > 0$

$$\left\| h_k \chi_k^0 \right\|_{L^\infty(\Omega)} \le \gamma^{-1} \left\| (h_k - h_\infty) \chi_k^0 \right\|_{L^\infty(\Omega)} \le \gamma^{-1} \|h_k - h_\infty\|_{L^\infty(\Omega)} \to 0$$

for $k \to \infty$ thanks to Lemma 5. □

Remark 27 (Local Density). We employ set-additivity of $\|\cdot\|_{\mathbb{W}^s}^2$ combined with the local approximation property (112d) to deduce for any sub-triangulation $\mathscr{T}_k' \subset \mathscr{T}_k$ and any $\bar{w} \in \mathbb{W}^s$ the local interpolation estimate

$$\|\bar{w} - I_k \bar{w}\|_{\mathbb{W}(\Omega(\mathscr{T}_k'))} \lesssim \|h_k^s\|_{L^\infty(\Omega(\mathscr{T}_k'))} \|\bar{w}\|_{\mathbb{W}^s(\Omega(\mathscr{T}_k'))}. \tag{115}$$

Using this estimate for $\mathscr{T}_k' = \mathscr{T}_k^0$ the above corollary implies

$$\|I_k \bar{w} - \bar{w}\|_{\mathbb{V}(\Omega_k^0)} \lesssim \|h_k^s\|_{L^\infty(\Omega_k^0)} \|\bar{w}\|_{\mathbb{W}^s(\Omega)} \qquad \text{for all } \bar{w} \in \mathbb{W}^s.$$

For any pair $w \in \mathbb{W}$ and $\bar{w} \in \mathbb{W}^s$ we then argue as in Sect. 7.2 for uniform refinement but restricted to subdomain $\Omega(\mathscr{T}_k^0)$ to conclude the 'local density'

$$\lim_{k\to\infty} \min_{W_k \in \mathbb{W}_k} \|w - W_k\|_{\mathbb{V}(\Omega_k^0)} = 0 \qquad \text{for all } w \in \mathbb{W}. \tag{116}$$

We use the interpolation estimate (115) in Proposition 4 below.

We next turn to the sequence $\{U_k\}_{k\geq 0}$ of approximate solutions. For characterizing the limit of this sequence we need the spaces

$$\mathbb{V}_\infty := \overline{\bigcup_{k\geq 0} \mathbb{V}_k} \qquad \text{and} \qquad \mathbb{W}_\infty := \overline{\bigcup_{k\geq 0} \mathbb{W}_k}.$$

Lemma 14 (Convergence of Petrov-Galerkin Approximations). *Assume that the sequence* $\{(\mathbb{V}_k, \mathbb{W}_k)\}_{k\geq 0}$ *satisfies* (112c) *and* (112b).

Then the sequence $\{U_k\}_{k\geq 0}$ *of approximate solutions converges in* $\mathbb{V}$ *to the solution* u_∞ *with respect to the pair* $(\mathbb{V}_\infty, \mathbb{W}_\infty)$*, which is characterized by*

$$u_\infty \in \mathbb{V}_\infty : \qquad \mathscr{B}[u_\infty, w] = f(w) \qquad \textit{for all } w \in \mathbb{W}_\infty. \tag{117}$$

Proof. 1 Let us first prove that the pair $(\mathbb{V}_\infty, \mathbb{W}_\infty)$ satisfies the inf-sup condition

$$\inf_{\substack{v\in\mathbb{V}_\infty \\ \|v\|_{\mathbb{V}}=1}} \sup_{\substack{w\in\mathbb{W}_\infty \\ \|w\|_{\mathbb{W}}=1}} \mathscr{B}[v, w] \geq \beta, \qquad \inf_{\substack{w\in\mathbb{W}_\infty \\ \|w\|_{\mathbb{W}}=1}} \sup_{\substack{v\in\mathbb{V}_\infty \\ \|v\|_{\mathbb{V}}=1}} \mathscr{B}[v, w] \geq \beta \tag{118}$$

with $\beta > 0$ from (112c).

To this end, fix first any $v \in \mathbb{V}_\infty \setminus \{0\}$ and choose a sequence $\{V_k\}_{k\geq 0}$ of functions $V_k \in \mathbb{V}_k$ such that $V_k \to v$ in $\mathbb{V}$ as $k \to \infty$. Moreover, with the help of (112c) choose a sequence $\{W_k\}_{k\geq 0}$ of functions $W_k \in \mathbb{W}_k$ such that

$$\|W_k\|_{\mathbb{W}} = 1 \qquad \text{and} \qquad \mathscr{B}[V_k, W_k] \geq \beta \|V_k\|_{\mathbb{V}}. \tag{119}$$

Thanks to (112a), the sequence $\{W_k\}_{k\geq 0}$ is in $\mathbb{W}$. Since the latter is reflexive, there exists a subsequence $\{W_{k_j}\}_{j\geq 0}$ and a function $w \in \mathbb{W}$ such that $W_{k_j} \rightharpoonup w$ weakly in $\mathbb{W}$ as $j \to \infty$. Since $\mathbb{W}_\infty$ is closed and convex as well as $\|\cdot\|_{\mathbb{W}}$ weakly lower semicontinuous, we have $w \in \mathbb{W}_\infty$ and $\|w\|_{\mathbb{W}} \leq \lim_{j\to\infty} \|W_{k_j}\|_{\mathbb{W}} = 1$. Combing this with (112c) yields

$$\mathscr{B}[v, w] \geq \beta \|v\|_{\mathbb{V}} \geq \beta \|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}.$$

In view of the first inequality, $w \neq 0$ and the first part of (118) is proved.

Proposition 1 states that (112c) is equivalent to

$$\inf_{\substack{W\in\mathbb{W}(\mathscr{T}) \\ \|W\|_{\mathbb{W}}=1}} \sup_{\substack{V\in\mathbb{V}(\mathscr{T}) \\ \|V\|_{\mathbb{V}}=1}} \mathscr{B}[V, W] \geq \beta. \tag{120}$$

In the same way, but using (120) instead of (112c), we show that for any $w \in \mathbb{W}_\infty$ there exists $v \in \mathbb{V}_\infty \setminus \{0\}$ such that $\mathscr{B}[v, w] \geq \beta \|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}$. This shows (118).

[2] The spaces $\mathbb{V}_\infty \subset \mathbb{V}$ and $\mathbb{W}_\infty \subset \mathbb{W}$ are closed and thus Hilbert spaces. The bilinear form $\mathscr{B}$ is continuous on $\mathbb{V}_\infty \times \mathbb{W}_\infty$ and satisfies the inf-sup condition (118). Therefore, by Theorem 2 there exits a unique $u_\infty \in \mathbb{V}_\infty$ with (117).
[3] By construction, $\mathbb{V}_k \subset \mathbb{V}_\infty$, which implies that the Petrov-Galerkin solution U_k is a $\|\cdot\|_{\mathbb{V}}$-quasi-optimal choice in $\mathbb{V}_k$ with respect to u_∞, i. e., there holds

$$\|u_\infty - U_k\|_{\mathbb{V}} \leq \frac{\|B\|}{\beta} \min_{V \in \mathbb{V}_k} \|u_\infty - V\|_{\mathbb{V}};$$

compare with Theorem 5. Besides that, $\bigcup_{k\geq 0} \mathbb{V}_k$ is dense in $\mathbb{V}_\infty$ and therefore

$$\lim_{k\to\infty} \|U_k - u_\infty\|_{\mathbb{V}} = 0. \qquad \square$$

In case of coercive $\mathscr{B}$ the proof is much simpler since coercivity is inherited from $\mathbb{V}$ to $\mathbb{V}_\infty$ and Step 1 of the proof is trivial. Existence of u_∞ is then a direct consequence of Corollary 2 (Lax-Milgram theorem). For symmetric and coercive $\mathscr{B}$ the above result has already been shown by Babuška and Vogelius [7].

Lemma 14 yields convergence of $U_k \to u_\infty$ in $\mathbb{V}$ as $k \to \infty$ irrespective of the decisions in the module MARK. We are going to prove below that the residual $\mathscr{R}_\infty$ of U_∞ satisfies $\mathscr{R}_\infty = 0$ in $\mathbb{W}^*$. The latter is equivalent to $u_\infty = u$ and thus shows Theorem 16. This, of course, hinges on the properties of ESTIMATE and MARK.

7.4.2 Auxiliary Results

Next we prove two auxiliary results, namely boundedness of the estimator and convergence of the indicators. Before embarking on this, we observe that the set-additivity of $\|\cdot\|^2_{\mathbb{V}}$ allows us to sum over overlapping patches, if the overlap is finite; compare also with the proof of Theorem 11. To be more precise: Local quasi-uniformity of $\mathscr{T}_k$ (55) implies $\#N_k(T) \lesssim 1$ for all $T \in \mathscr{T}_k$. Thus set-additivity (111) of $\|\cdot\|^2_{\mathbb{V}}$ gives for any subset $\mathscr{T}'_k \subset \mathscr{T}_k$ and any $v \in \mathbb{V}$

$$\sum_{T\in\mathscr{T}'_k} \|v\|^2_{\mathbb{V}(N_k(T))} = \sum_{T\in\mathscr{T}'_k} \sum_{T'\in N_k(T)} \|v\|^2_{\mathbb{V}(T')} \lesssim \sum_{T\in\mathscr{T}^*_k} \|v\|^2_{\mathbb{V}(T)} = \|v\|^2_{\mathbb{V}(\Omega^*_k)} \tag{121}$$

with $\mathscr{T}^*_k = \{T' \in \mathscr{T}_k \mid T' \in N_k(T), T \subset \mathscr{T}'_k\}$ and $\Omega^*_k := \Omega(\mathscr{T}^*_k)$. The same argument applies to $\|\cdot\|^2_{\mathbb{W}}$, $\|\cdot\|^2_{\mathbb{W}^s}$, and $\|\cdot\|^2_{L^2(\Omega)}$.

In the next results we use the stability estimate

$$\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}}, T) \lesssim \|U_{\mathscr{T}}\|_{\mathbb{V}(N_{\mathscr{T}}(T))} + \|\tilde{D}\|_{L^2(N_{\mathscr{T}}(T))} \qquad \text{for all } T \in \mathscr{T}, \tag{122}$$

where $\tilde{D} = \tilde{D}(u, D) \in L^2(\Omega)$ with D from (113b). This bound can be derived as follows. Combining the lower bound (113b) and the triangle inequality we infer

$$\begin{aligned}\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T) &\lesssim \|U_{\mathscr{T}}-u\|_{\mathbb{V}(N_{\mathscr{T}}(T))}+\operatorname{osc}_{\mathscr{T}}(U_{\mathscr{T}},T)\\ &\lesssim \|U_{\mathscr{T}}\|_{\mathbb{V}(N_{\mathscr{T}}(T))}+\|u\|_{\mathbb{V}(N_{\mathscr{T}}(T))}+\|D\|_{L^2(N_{\mathscr{T}}(T))},\end{aligned}$$

where the constant in $\lesssim$ also depends on $\|h_0^q\|_{L^\infty(\Omega)}$ via (113c). Since $\|\cdot\|_{\mathbb{V}(\Omega)}$ is an L^2-type norm, the stability of the indicators (122) is a direct consequence of (113b) with $\tilde{D}=\tilde{D}(u,D)\in L^2(\Omega)$.

Lemma 15 (Stability). *Let the finite element spaces and the error the indicators satisfy* (112c) *respectively* (122)*.*

Then the estimators $\mathscr{E}_k(U_k,\mathscr{T}_k)$ are uniformly bounded, i. e.,

$$\mathscr{E}_k(U_k,\mathscr{T}_k)\lesssim 1 \qquad \text{for all } k\geq 0.$$

Proof. Using (121) and the stability of the indicators (122) we derive for all $k\geq 0$

$$\mathscr{E}_k^2(U_k,\mathscr{T}_k)\lesssim \sum_{T\in\mathscr{T}_k}\|U_k\|^2_{\mathbb{V}(N_k(T))}+\|\tilde{D}\|^2_{L^2(N_k(T))}\lesssim \|U_k\|^2_{\mathbb{V}(\Omega)}+\|\tilde{D}\|^2_{L^2(\Omega)}.$$

The uniform estimate $\|U_k\|_{\mathbb{V}(\Omega)}\leq\beta^{-1}\|f\|_{\mathbb{V}^*}$ implies the claim. □

We next investigate the maximal indicator in the set of marked elements. In addition to convergence of the discrete solutions and mesh-size functions we exploit stability of the indicators, and properties of REFINE.

Lemma 16 (Marking). *Suppose that the finite element spaces fulfill* (112) *and the estimator* (113b) *and* (113c)*.*

Then the maximal indicator of the marked elements vanishes in the limit:

$$\lim_{k\to\infty}\max\{\mathscr{E}_k(U_k,T)\mid T\in\mathscr{M}_k\}=0.$$

Proof. Let $T_k\in\mathscr{M}_k$ such that $\mathscr{E}_k(U_k,T_k)=\max\{\mathscr{E}_k(U_k,T)\mid T\in\mathscr{M}\}$. All elements in $\mathscr{M}_k$ are refined and therefore $T_k\in\mathscr{T}_k^0$. Local quasi-uniformity (55) of $\mathscr{T}_k$ implies

$$|N_k(T_k)|\lesssim|T_k|\leq\|h_k^d\|_{L^\infty(T_k)}\leq\|h_k^d\|_{L^\infty(\Omega_k^0)}\to 0 \tag{123}$$

as $k\to\infty$ by Corollary 10.

As shown above (113b) and (113c) imply the stability (122), whence we can proceed by the triangle inequality to estimate the maximal indicator by

$$\mathscr{E}_k(U_k,T_k)\lesssim\|U_k-u_\infty\|_{\mathbb{V}(\Omega)}+\|u_\infty\|_{\mathbb{V}(N_k(T_k))}+\|\tilde{D}\|_{L^2(N_k(T_k))}.$$

The first term on the right hand side converges to 0 as $k\to\infty$, thanks to Lemma 14 and the other terms vanish in the limit too, by continuity of $\|\cdot\|_{\mathbb{V}(\Omega)}$ and $\|\cdot\|_{L^2(\Omega)}$ with respect to the Lebesgue measure $|\cdot|$ and (123).

7.4.3 Convergence of the Residuals

In this section we establish the weak convergence $\mathscr{R}_k \rightharpoonup 0$ in $\mathbb{W}^*$. In doing this, we distinguish two regions in Ω: in Ω_k^0 we use local density of the finite element spaces $\mathbb{W}_k$ in $\mathbb{W}$, and in Ω_k^+ we rely on properties of estimator and marking.

Proposition 4 (Weak Convergence of the Residuals). *Assume that* (112)*,* (113)*, and* (114) *are satisfied.*

Then the sequence of discrete solutions $\{U_k\}_{k\geq 0}$ *generated by iteration* (104) *verifies*

$$\lim_{k\to\infty} \langle \mathscr{R}_k, w\rangle = 0 \qquad \text{for all } w \in \mathbb{W}^s.$$

Proof. [1] For $k \geq \ell$ the inclusion $\mathscr{T}_\ell^+ \subset \mathscr{T}_k^+ \subset \mathscr{T}_k$ holds. Therefore, the sub-triangulation $\mathscr{T}_k \setminus \mathscr{T}_\ell^+$ of $\mathscr{T}_k$ covers the sub-domain Ω_ℓ^0, i. e., $\Omega_\ell^0 = \Omega(\mathscr{T}_k \setminus \mathscr{T}_\ell^+)$. We notice that any refinement of $\mathscr{T}_k$ does not affect any element in $\mathscr{T}_\ell^+$. Therefore, defining

$$\mathscr{T}_k^* = \{T' \mid T' \in N_k(T), T \in \mathscr{T}_k \setminus \mathscr{T}_\ell^+\}.$$

we also see that for $k \geq \ell$

$$\Omega_k^* = \Omega(\mathscr{T}_k^*) = \bigcup \{T' \colon T' \in N_\ell(T), T \in \mathscr{T}_\ell^0\}. \tag{124}$$

[2] Let $w \in \mathbb{W}^s$ with $\|w\|_{\mathbb{W}^s(\Omega)} = 1$ be arbitrarily chosen. Since U_k is the Petrov-Galerkin solution we can employ Galerkin orthogonality (41) in combination with the upper bound (113a) to split for $k \geq \ell$

$$\begin{aligned}
|\langle \mathscr{R}_k, w\rangle| &= |\langle \mathscr{R}_k, w - I_k w\rangle| \\
&\lesssim \sum_{T\in\mathscr{T}_k\setminus\mathscr{T}_\ell^+} \mathscr{E}_k(U_k,T)\|w - I_k w\|_{\mathbb{V}(N_k(T))} + \sum_{T\in\mathscr{T}_\ell^+} \mathscr{E}_k(U_k,T)\|w - I_k w\|_{\mathbb{V}(N_k(T))} \\
&\lesssim \mathscr{E}_k(U_k, \mathscr{T}_k \setminus \mathscr{T}_\ell^+)\|w - I_k w\|_{\mathbb{V}(\Omega_k^*)} + \mathscr{E}_k(U_k, \mathscr{T}_\ell^+)\|w - I_k w\|_{\mathbb{V}(\Omega)},
\end{aligned}$$

by the Cauchy-Schwarz inequality and (121) for $\|\cdot\|_{\mathbb{V}}^2$. In view of Lemma 15 we bound $\mathscr{E}_k(U_k, \mathscr{T}_k \setminus \mathscr{T}_\ell^+) \leq \mathscr{E}_k(U_k, \mathscr{T}_k) \lesssim 1$. We next use (115) with $\mathscr{T}_k' = \mathscr{T}_k^*$ to obtain $\|w - I_k w\|_{\mathbb{W}(\Omega_k^*)} \lesssim \|h_k^s\|_{L^\infty(\Omega_k^*)}$, recalling $\|w\|_{\mathbb{W}^s(\Omega)} = 1$. From (124) we see that for any $T' \in \mathscr{T}_k^*$ we find $T \in \mathscr{T}_\ell^0$ with $T' \subset N_\ell(T)$. Local quasi-uniformity (55) of $\mathscr{T}_\ell$ and monotonicity of the mesh-size functions therefore imply

$$\|h_k\|_{L^\infty(\Omega_k^*)} \lesssim \|h_k\|_{L^\infty(\Omega_\ell^0)} \leq \|h_\ell\|_{L^\infty(\Omega_\ell^0)}.$$

In summary this yields

$$\|w - I_k w\|_{\mathbb{V}(\Omega_\ell^*)} \lesssim \|h_\ell^s\|_{L^\infty(\Omega_\ell^0)} \qquad \text{and} \qquad \|w - I_k w\|_{\mathbb{V}(\Omega)} \lesssim 1,$$

which entails the existence of constants $0 \leq C_1, C_2 < \infty$, such that

$$|\langle \mathscr{R}_k, w\rangle| \leq C_1 \|h_\ell^s\|_{L^\infty(\Omega_\ell^0)} + C_2 \mathscr{E}_k(U_k, \mathscr{T}_\ell^+) \qquad \text{for all } k \geq \ell. \tag{125}$$

[3] For any given $\varepsilon > 0$, convergence of the mesh-size function $\|h_\ell\|_{L^\infty(\Omega_\ell^0)} \to 0$ for $\ell \to \infty$, proven in Corollary 10, and $s > 0$ allows us to first choose $\ell \geq 0$ such that

$$\|h_\ell^s\|_{L^\infty(\Omega_\ell^0)} \leq \frac{\varepsilon}{2C_1}.$$

Employing the marking rule (114), we conclude

$$\lim_{k\to\infty} \max\{\mathscr{E}_k(U_k,T) \mid T \in \mathscr{T}_k \setminus \mathscr{M}_k\} \leq \lim_{k\to\infty} g\big(\max\{\mathscr{E}_k(U_k,T) \mid T \in \mathscr{M}_k\}\big) = 0$$

by Lemma 16 and continuity of g in 0 with $g(0) = 0$. Since $\mathscr{T}_\ell^+ \cap \mathscr{M}_k = \emptyset$, this especially implies $\max\{\mathscr{E}_k(U_k,T) \mid T \in \mathscr{T}_\ell^+\} \to 0$, whence we can next choose $K \geq \ell$ such that

$$\mathscr{E}_k(U_k,T) \leq \frac{\varepsilon}{2C_2}(\#\mathscr{T}_\ell^+)^{-1/2} \qquad \text{for all } T \in \mathscr{T}_\ell^+ \text{ and all } k \geq K,$$

yielding $C_2\mathscr{E}_k(U_k,\mathscr{T}_\ell^+) \leq \varepsilon/2$ for those k. In summary, estimate (125) then implies $|\langle \mathscr{R}_k, w\rangle| \leq \varepsilon$ for $k \geq K$. Since ε is arbitrary this finishes the proof. □

7.4.4 Proof of Convergence

Collecting the auxillary results, we are in the position to prove the main result.

Proof of Theorem 16. [1] We first show convergence $U_k \to u$ in $\mathbb{V}$. For any $w \in \mathbb{W}^s$ we deduce

$$\begin{aligned}\langle \mathscr{R}_\infty, w\rangle &= \langle \mathscr{R}_\infty - \mathscr{R}_k, w\rangle + \langle \mathscr{R}_k, w\rangle = \mathscr{B}[u_\infty - U_k, w] + \langle \mathscr{R}_k, w\rangle \\ &\leq \|\mathscr{B}\|\|u_\infty - U_k\|_{\mathbb{V}(\Omega)}\|w\|_{\mathbb{V}(\Omega)} + \langle \mathscr{R}_k, w\rangle \to 0\end{aligned}$$

as $k \to \infty$ by Lemma 14 and Proposition 4, whence $\langle \mathscr{R}_\infty, w\rangle = 0$ for all $w \in \mathbb{W}^s$. This implies $\mathscr{R}_\infty = 0$ in $\mathbb{W}^*$ since $\mathbb{W}^s$ is dense in $\mathbb{W}$. The continuous inf-sup condition (21) yields

$$\alpha\|u_\infty - u\|_{\mathbb{V}} \leq \sup_{\|w\|_{\mathbb{W}}=1} \mathscr{B}[u_\infty - u, w] = \sup_{\|w\|_{\mathbb{W}}=1} \langle \mathscr{R}_\infty, w\rangle = 0,$$

which shows $u = u_\infty$. Convergence of the Galerkin approximations finally implies

$$\lim_{k\to\infty} U_k = u_\infty = u \qquad \text{in } \mathbb{V}.$$

[2] After proving $U_k \to u$ we next turn to the convergence of the estimators. Just like in the proof of Proposition 4 we split for $k \geq \ell$

$$\mathscr{E}_k^2(U_k,\mathscr{T}_k) = \mathscr{E}_k^2(U_k,\mathscr{T}_k \setminus \mathscr{T}_\ell^+) + \mathscr{E}_k^2(U_k,\mathscr{T}_\ell^+)$$

and we estimate the first term with the help of the local lower bound (113b), (113c) by

$$\begin{aligned}\mathscr{E}_k^2(U_k,\mathscr{T}_k\setminus\mathscr{T}_\ell^+) &\lesssim \sum_{T\in\mathscr{T}_k\setminus\mathscr{T}_\ell^+}\|U_k-u\|^2_{\mathbb{V}(N_k(T))}+h_T^{2q}\Big(\|U_k\|^2_{\mathbb{V}(N_k(T))}+\|D\|^2_{L^2(N_k(T))}\Big)\\ &\lesssim \|U_k-u\|^2_{\mathbb{V}}+\big\|h_k^{2q}\big\|_{L^\infty(\Omega_\ell^0)}\big(\beta^{-2}\|f\|^2_{\mathbb{V}^*}+\|D\|^2_{L^2(\Omega)}\big),\end{aligned}$$

where we have used (121) for $\|\cdot\|_{\mathbb{V}(\Omega)}$ and $\|\cdot\|_{L^2(\Omega)}$ as well as $\|U_k\|_{\mathbb{V}}\le\beta^{-1}\|f\|_{\mathbb{W}^*}$ in the second step. Using once again monotonicity of the mesh-size functions we deduce for some constants C_1, C_2

$$\mathscr{E}_k^2(U_k,\mathscr{T}_k)\le C_1\big\|h_\ell^{2q}\big\|_{L^\infty(\Omega_\ell^0)}+C_2\|U_k-u\|^2_{\mathbb{V}(\Omega)}+\mathscr{E}_k^2(U_k,\mathscr{T}_\ell^+).$$

By Corollary 10 we can make the first term small by choosing ℓ sufficiently large. In the proof of Proposition 4 we already have shown $\mathscr{E}_k(U_k,\mathscr{T}_\ell^+)\to 0$ for fixed ℓ and $k\to\infty$. Step 1 implies $\|U_k-u\|_{\mathbb{V}(\Omega)}\to 0$ as $k\to\infty$ which allows to make the last two terms small by choosing k large after fixing ℓ. This proves $\mathscr{E}_k(U_k,\mathscr{T}_k)\to 0$ as $k\to\infty$ and finishes the proof. □

Remark 28 (Lower Bound). For convergence $U_k\to u$ we have only utilized the stability (122) of the indicators, which is much weaker than efficiency (113b) because it allows for overestimation. Since most of the estimators for linear problems are shown to be reliable and efficient, we directly asked for efficiency of the estimator. For nonlinear problems this might be different and just asking for (122) may provide access for proving convergence for a larger problem class.

All convergence results but [21, 67] rely on a *discrete* local lower bound. For the model problem there is no difference in deriving the continuous or the discrete lower bound; compare with Sect. 6.3. In general, the derivation of a discrete lower bound is much more involved than its continuous counterpart. For instance, in Problem 44 below the discrete lower bound is not known and in Problem 45 it is only known for the lowest order elements. In respect thereof a convergence proof without lower bound enlarges the problem class where it applies to.

Yet, only asking for (122) yields convergence $U_k\to u$ but the progress without convergence $\mathscr{E}_k(U_k,\mathscr{T}_k)\to 0$ is not observable in the adaptive iteration. Therefore, a convergence result for non-efficient estimators is of little practical use.

Remark 29 (Characterization of Convergent Marking). The results in [55] and [67] also give a characterization of convergent marking. In our setting

$$\lim_{k\to\infty}\max\{\mathscr{E}_k(U_k,T)\mid T\in\mathscr{M}_k\}=0\quad\Longrightarrow\quad\lim_{k\to\infty}\mathscr{E}_k(U_k,T)=0\quad\text{for all } T\in\mathscr{T}^+ \tag{126}$$

is necessary and sufficient for convergence of (104). To see this, the hypothesis of (126) we have shown in Lemma 16 and the conclusion of (126) is obviously necessary for $\mathscr{E}_k(U_k,\mathscr{T}_k)\to 0$. If $\lim_{k\to\infty}\operatorname{osc}_k(U_k,T)=0$ for all $T\in\mathscr{T}^+$ then it is also necessary for $\|U_k-u\|_{\mathbb{V}}\to 0$ by the lower bound (113b), for instance in the

model problem when $\boldsymbol{A}$ and f are piecewise constant over $\mathscr{T}_0$. Condition (114) on marking we only have used in Step 3 of the proof to Proposition 4 and there it can be replaced by (126), whence (126) is also sufficient.

On the one hand, this assumption is not 'a posteriori' in that it can not be checked at iteration k of the adaptive loop and thus seems of little practical use. On the other hand, being a characterization of convergent marking it may be used to treat marking strategies that are based on extrapolation techniques involving indicators from previous iterations [5], or that are based on some optimization procedure [41].

Similarly, the condition on marking can be generalized to marking procedures where a given tolerance of the adaptive method enters the selection of elements, for instance the original equidistribution strategy for parabolic problems in [34]. Such methods then in turn only aim at convergence into tolerance. For details we refer to [67, Sect. 5].

7.5 Problems

Problem 43. Consider the general 2^{nd} order elliptic problem from Sect. 2.2.2, where $\boldsymbol{A}$ piecewise Lipschitz over $\mathscr{T}_0$ with smallest eigenvalue strictly bounded away from 0 and $c - 1/2 \operatorname{div} \boldsymbol{b} \geq 0$. Therefore, the corresponding bilinear form $\mathscr{B}$ is coercive on $\mathbb{V} = H_0^1(\Omega)$; compare with Sect.2.5.2.

Show that a discretization with H_0^1 conforming Lagrange elements of order $n \geq 1$ introduced in Sect. 3.2.2 and the residual estimator from Sect. 6.2 satisfy the assumptions (112) and (113). This implies convergence of the adaptive iteration (104) for the general 2^{nd} order elliptic equation with any of the marking strategies from Sect. 7.3.4.

Problem 44. Consider the biharmonic equation in 2d from Sect. 2.2.2 which leads to a variational problem in $\mathbb{V} = H_0^2(\Omega)$ with a continuous and coercive bilinear form.

Show that the discretization with the Argyris triangle defined in [25, Theorems 2.2.11 and 2.2.13] of Ciarlet's book satisfies (112). In addition verify that the estimator derived by Verfürth in [76, Section 3.7] fulfills (113). This implies convergence of the adaptive iteration (104) for the biharmonic equation with any of the marking strategies from Sect. 7.3.4.

Problem 45. Consider the 3d Eddy Current Equations from Sect. 2.2.2 which leads to a variational problem in $\mathbb{V} = H_0(\operatorname{curl}; \Omega)$ with a continuous and coercive bilinear form.

Show that the discretization with Nédélec finite elements of order $n \in \mathbb{N}$ comply with (112); compare with [51, Sect. 5.5]. Consider the estimator derived by Schöberl [64, Corollary 2] that has been shown to be efficient by Beck et al. [12, Theorem 3.3]. Show that it fulfills (113). This implies convergence of the adaptive iteration (104) for the 3d Eddy Current Equations with any of the marking strategies from Sect. 7.3.4.

Problem 46. Consider the Stokes problem from Sect. 2.2.2 that leads to a variational problem in $\mathbb{V} = H^1_0(\Omega;\mathbb{R}^d) \times L^2_0(\Omega)$ with a non-coercive bilinear form $\mathscr{B}$ that satisfies the inf-sup condition (23).

For the discretization with the Taylor-Hood element of order $n \geq 2$, this means we approximate the velocity with continuous piecewise polynomials of degree n and the pressure with continuous piecewise polynomials of degree $n-1$, Otto has shown (112c) in [59]. Prove that the Taylor-Hood element satisfies the other requirements of (112). Finally show that the estimator by Verfürth for the Stokes system [75] complies with (113). This implies convergence of the adaptive iteration (104) for the Stokes problem with any of the marking strategies from Sect. 7.3.4.

8 Adaptivity: Contraction Property

This chapter discusses the contraction property of AFEM for the *model problem* of Sect. 2.2.1, namely

$$-\operatorname{div}(\boldsymbol{A}(x)\nabla u) = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega. \tag{127}$$

The variational formulation of (127) from Sect. 2.5.1 reads with $\mathbb{V} = \mathbb{W} = H^1(\Omega)$

$$u \in \mathbb{V}: \quad \mathscr{B}[u,v] := \int_\Omega \nabla v \cdot \boldsymbol{A}(x)\nabla u = \int_\Omega f\, v =: \langle f, v\rangle \qquad \text{for all } v \in \mathbb{V}.$$

We revisit the modules of the basic adaptive loop (4), i. e.,

$$\mathsf{SOLVE} \longrightarrow \mathsf{ESTIMATE} \longrightarrow \mathsf{MARK} \longrightarrow \mathsf{REFINE}.$$

Similar to Chap. 7, the outcome of each iteration with counter $k \geq 1$ is a sequence $\{\mathscr{T}_k, \mathbb{V}_k, U_k\}_{k=0}^\infty$ of conforming bisection refinements $\mathscr{T}_k$ of $\mathscr{T}_0$, spaces of conforming finite element spaces $\mathbb{V}_k = \mathbb{W}_k = S^{n,0}(\mathscr{T}_k) \cap H_0^1(\Omega)$, i. e., C^0 continuous piecewise polynomials of degree $\leq n$ for both ansatz and test spaces, and Ritz-Galerkin solutions $U_k \in \mathbb{V}_k$.

Since error monotonicity is closely related to a minimization principle, we cannot in general expect a contraction property for problems governed by an inf-sup condition. We thus restrict ourselves to the special class of coercive and symmetric problems of the form (127). The first contribution in dimension $d > 1$ is due to Dörfler [32], who introduced a crucial marking, the so-called Dörfler marking of Sect. 7.1, and proved strict energy error reduction for the Laplacian provided the initial mesh $\mathscr{T}_0$ satisfies a fineness assumption. The Dörfler marking will play an essential role in the present discussion, which does not seem to extend to other marking strategies such as those in Sect. 7.1. Morin, Nochetto, and Siebert [52, 53] showed that such strict energy error reduction does not hold in general even for (127). By introducing the concept of data oscillation and the interior node property, they proved convergence of the AFEM without restrictions on $\mathscr{T}_0$. The latter result, however, is valid only for $\boldsymbol{A}$ in (127) piecewise constant on $\mathscr{T}_0$. Inspired by the work of Chen and Feng [24], Mekchay and Nochetto [48] proved a contraction property for the total error, namely the sum of the energy error plus oscillation, for general second order elliptic operators such as those in Sect. 2.5.2. For non-symmetric $\mathscr{B}$ this requires a sufficient fineness of the initial grid $\mathscr{T}_0$. The total error will reappear in the study of convergence rates in Chap. 9.

Diening and Kreuzer proved a similar contraction property for the p-Laplacian replacing the energy norm by the so-called quasi-norm [31]. They were able to avoid marking for oscillation by using the fact that oscillation is dominated by the estimator. Most results for nonlinear problems utilize the equivalence of the energy error and error in the associated (nonlinear) energy; compare with Problem 49. This equivalence was first used by Veeser in a convergence analysis for the p-Laplacian [73] and later on by Siebert and Veeser for the obstacle problem [68].

The result of Diening and Kreuzer inspired the work by Cascón et al. [21], who proved a contraction property for the *quasi-error*:

$$|||u-U_k|||_\Omega^2+\gamma\mathscr{E}_k^2(U_k,\mathscr{T}_k),$$

where $\gamma>0$ is a suitable scaling constant. This approach hinges solely on a strict reduction of the mesh-size within refined elements, the upper a posteriori error bound, an orthogonality property natural for (127) in nested approximation spaces, and Dörfler marking. This appears to be the simplest approach currently available and is presented next.

8.1 The Modules of AFEM for the Model Problem

We assume Ω is triangulated by some initial grid $\mathscr{T}_0$. We suppose that $\boldsymbol{A}$ is uniformly SPD so that (127) is *coercive* and in addition we ask $\boldsymbol{A}$ to be piecewise Lipschitz over $\mathscr{T}_0$. We next describe the modules of the adaptive algorithm.

Module SOLVE. For any $\mathscr{T}\in\mathbb{T}$ we set $\mathbb{V}(\mathscr{T})=S^{n,0}(\mathscr{T})\cap H_0^1(\Omega)$ and suppose that

$$U_\mathscr{T}=\mathsf{SOLVE}\big(\mathbb{V}(\mathscr{T})\big)$$

outputs the *exact* Ritz-Galerkin approximation in $\mathbb{V}(\mathscr{T})$, namely,

$$U_\mathscr{T}\in\mathbb{V}(\mathscr{T}):\qquad \mathscr{B}[U_\mathscr{T},V]=\langle f,V\rangle\qquad\text{for all } V\in\mathbb{V}(\mathscr{T}).$$

This entails exact integration and linear algebra; see Remarks 9 and 10.

Module ESTIMATE. Given a grid $\mathscr{T}\in\mathbb{T}$ and the Ritz-Galerkin approximation $U_\mathscr{T}\in\mathbb{V}(\mathscr{T})$ the output

$$\{\mathscr{E}_\mathscr{T}(U_\mathscr{T},T)\}_{T\in\mathscr{T}}=\mathsf{ESTIMATE}\big(U_\mathscr{T},\mathscr{T}\big)$$

are the indicators of the residual estimator derived in Chap. 6. We recall that for a generic function $V\in\mathbb{V}(\mathscr{T})$ the *element* and *jump residuals* are defined by

$$\begin{aligned} r(V)|_T&=f+\operatorname{div}(\boldsymbol{A}\nabla V)=f && \text{for all } T\in\mathscr{T},\\ j(V)|_S&=[\![\boldsymbol{A}\nabla V]\!]_S && \text{for all } S\in\mathring{\mathscr{S}}\end{aligned}$$

and the element indicator evaluated in V is then

$$\mathscr{E}_\mathscr{T}^2(V,T)=h_T^2\,\|r(V)\|_{L^2(T)}^2+h_T\,\|j(V)\|_{L^2(\partial T\cap\Omega)}^2\quad\text{for all } T\in\mathscr{T}.$$

Module MARK. For any $\mathscr{T}\in\mathbb{T}$ and indicators $\{\mathscr{E}_\mathscr{T}(U_\mathscr{T},T)\}_{T\in\mathscr{T}}$ the module MARK selects elements for refinement using *Dörfler Marking*, i. e., using a fixed

parameter $\theta \in (0,1]$ the output

$$\mathscr{M} = \mathsf{MARK}\big(\{\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},T)\}_{T\in\mathscr{T}},\mathscr{T}\big)$$

satisfies

$$\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},\mathscr{M}) \geq \theta\,\mathscr{E}_{\mathscr{T}}(U_{\mathscr{T}},\mathscr{T}).$$

Dörfler Marking guarantees that the total estimator is controlled up the constant θ^{-1} by the estimator on the marked elements. This is a crucial property in our arguments. The choice of $\mathscr{M}$ does not have to be minimal at this stage, that is, the marked elements $T \in \mathscr{M}$ do not necessarily must be those with largest indicators. However, minimality of $\mathscr{M}$ will be crucial to derive rates of convergence in Chap. 9.

Module REFINE. We fix the number $b \in \mathbb{N}$ of bisections and consider the module REFINE from Sect. 4.4 to refine all marked elements b times. Then for any $\mathscr{T} \in \mathbb{T}$ the output

$$\mathscr{T}_* = \mathsf{REFINE}\big(\mathscr{T},\mathscr{M}\big)$$

satisfies $\mathscr{T}_* \in \mathbb{T}$. Furthermore, if $\mathscr{R}_{\mathscr{T}\to\mathscr{T}_*}$ is the set of refined elements of $\mathscr{T}$, then $\mathscr{M} \subset \mathscr{R}_{\mathscr{T}\to\mathscr{T}_*}$ and $h_{\mathscr{T}_*} \leq 2^{-b/d} h_{\mathscr{T}}$ inside all elements of $\mathscr{R}_{\mathscr{T}\to\mathscr{T}_*}$.

8.2 Properties of the Modules of AFEM

We next summarize some basic properties of the adaptive algorithm that emanate from the symmetry of the differential operator and features of the modules. In doing this, any explicit constant or hidden constant in $\lesssim$ must, apart from explicitly stated other dependencies, only depend on the uniform shape-regularity of $\mathbb{T}$, the dimension d, the polynomial degree n, and the (global) eigenvalues of $\boldsymbol{A}$, but not on a specific grid $\mathscr{T} \in \mathbb{T}$. Further on, u will always be the weak solution of (127).

Lemma 17 (Nesting of Spaces). *Any sequence $\{\mathbb{V}_k = \mathbb{V}(\mathscr{T}_k)\}_{k\geq 0}$ of discrete spaces generated by the basic adaptive loop* (4) *is nested, this is,*

$$\mathbb{V}_k \subset \mathbb{V}_{k+1} \qquad \text{for all } k \geq 0.$$

Proof. See Problem 47. □

The following property relies on the fact that $\mathscr{B}$ is coercive and symmetric, and so induces a scalar product in $\mathbb{V}$ equivalent to the H^1_0-scalar product.

Lemma 18 (Pythagoras). *Let $\mathscr{T},\mathscr{T}_* \in \mathbb{T}$ such that $\mathscr{T} \leq \mathscr{T}_*$. The respective Ritz-Galerkin solutions $U \in \mathbb{V}(\mathscr{T})$ and $U_* \in \mathbb{V}(\mathscr{T}_*)$ satisfy the following orthogonality property in the energy norm $|||\cdot|||_\Omega$*

$$|||u-U|||_\Omega^2 = |||u-U_*|||_\Omega^2 + |||U_*-U|||_\Omega^2\,. \tag{128}$$

Proof. See Problem 48. □

A by-product of (128) is the monotonicity property

$$|||U_* - U|||_\Omega \leq |||u - U|||_\Omega \,. \tag{129}$$

A perturbation of (128) is still valid for the general 2nd order elliptic operators of Sect. 2.5.2, as shown in [48], but not for non-coercive problems. Even for (127), property (128) is valid exclusively for the energy norm. This restricts the subsequent analysis to the energy norm, or equivalent norms, but does not extend to other, perhaps more practical, norms such as the maximum norm. This is an open problem.

We now continue the discussion of oscillation of Sect. 6.3.3. In view of (96), we denote by $\operatorname{osc}_{\mathscr{T}}(V,T)$ the *element oscillation* for any $V \in \mathbb{V}$

$$\operatorname{osc}_{\mathscr{T}}(V,T) = \|h(r(V) - \overline{r(V)})\|_{L^2(T)} + \|h^{1/2}(j(V) - \overline{j(V)})\|_{L^2(\partial T \cap \Omega)},$$

where $\overline{r(V)} = P_{2n-2}r(V)$ and $\overline{j(V)} = P_{2n-1}j(V)$ stand for L^2-projections of the residuals $r(V)$ and $j(V)$ onto the polynomials $\mathbb{P}_{2n-2}(T)$ and $\mathbb{P}_{2n-1}(S)$ defined on the element T or side $S \subset \partial T$, respectively. For variable $\boldsymbol{A}$, $\operatorname{osc}_{\mathscr{T}}(V,T)$ depends on the discrete function $V \in \mathbb{V}$, and its study is more involved than for piecewise constant $\boldsymbol{A}$. In the latter case, $\operatorname{osc}_{\mathscr{T}}(V,T)$ becomes *data oscillation* $\operatorname{osc}_{\mathscr{T}}(V,T) = \|h(f - \bar{f})\|_{L^2(T)}$; compare with Remark 25.

We now rewrite the a posteriori error estimates of Theorem 15 in the energy norm.

Lemma 19 (A Posteriori Error Estimates). *There exist constants $0 < C_2 \leq C_1$, such that for any $\mathscr{T} \in \mathbb{T}$ and the corresponding Ritz-Galerkin solution $U \in \mathbb{V}(\mathscr{T})$ there holds*

$$|||u - U|||_\Omega^2 \leq C_1\, \mathscr{E}_{\mathscr{T}}^2(U,\mathscr{T}) \tag{130a}$$

$$C_2\, \mathscr{E}_{\mathscr{T}}^2(U,\mathscr{T}) \leq |||u - U|||_\Omega^2 + \operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T}). \tag{130b}$$

The constants C_1 and C_2 depend on the smallest and largest global eigenvalues of $\boldsymbol{A}$. This dependence can be improved if the a posteriori analysis is carried out directly in the energy norm instead of the H_0^1-norm; see Problem 42. The definitions of $\overline{r(V)}$ and $\overline{j(V)}$, as well as the lower bound (130b), are immaterial for deriving a contraction property. However, they will be important for proving convergence rates in Chap. 9.

Lemma 20 (Lipschitz Property). *For any $\mathscr{T} \in \mathbb{T}$ and $T \in \mathscr{T}$, there holds*

$$|\mathscr{E}_{\mathscr{T}}(V,T) - \mathscr{E}_{\mathscr{T}}(W,T)| \lesssim \eta_{\mathscr{T}}(\boldsymbol{A},T)\|\nabla(V-W)\|_{L^2(\omega_T)} \quad \textit{for all } V,W \in \mathbb{V}(\mathscr{T}).$$

By ω_T we again denote the union of elements sharing a side with T, $\operatorname{div}\boldsymbol{A} \in \mathbb{R}^d$ *is the divergence of $\boldsymbol{A}$ computed by rows, and*

$$\eta_{\mathscr{T}}(\boldsymbol{A},T) := h_T\|\operatorname{div}\boldsymbol{A}\|_{L^\infty(T)} + \|\boldsymbol{A}\|_{L^\infty(\omega_T)}.$$

Proof. Recalling the definition of the indicators, the triangle inequality yields

$$|\mathscr{E}_{\mathscr{T}}(V,T)-\mathscr{E}_{\mathscr{T}}(W,T)| \le h_T\|r(V)-r(W)\|_{L^2(T)} + h_T^{1/2}\|j(V)-j(W)\|_{L^2(\partial T)}.$$

We set $E := V-W \in \mathbb{V}(\mathscr{T})$, and observe that

$$r(V)-r(W) = \operatorname{div}(\boldsymbol{A}\nabla E) = \operatorname{div}\boldsymbol{A}\cdot\nabla E + \boldsymbol{A} : D^2E,$$

where D^2E is the Hessian of E. Since E is a polynomial of degree $\le n$ in T, applying the inverse estimate $\|D^2E\|_{L^2(T)} \lesssim h_T^{-1}\|\nabla E\|_{L^2(T)}$, we deduce

$$h_T\|r(V)-r(W)\|_{L^2(T)} \lesssim \eta_{\mathscr{T}}(\boldsymbol{A},T)\|\nabla E\|_{L^2(T)}.$$

On the other hand, for any $S\subset\partial T$ applying the inverse estimate of Problem 50 gives

$$\|j(V)-j(W)\|_{L^2(S)} = \|j(E)\|_{L^2(S)} = \|\,[\![A\nabla E]\!]\,\|_{L^2(S)} \lesssim h_T^{-1/2}\|\nabla E\|_{L^2(\omega_T)}$$

where the hidden constant is proportional to $\eta_{\mathscr{T}}(\boldsymbol{A},T)$. This finishes the proof. □

One serious difficulty in dealing with AFEM is that one has access to the energy error $|||u-U|||_\Omega$ only through the estimator $\mathscr{E}_{\mathscr{T}}(U,\mathscr{T})$. The latter, however, fails to exhibit a monotonicity property such as (129) because it depends on the discrete solution $U\in\mathbb{V}(\mathscr{T})$ that changes with the mesh. We account for this change in the next lemma, which is a direct consequence of Lemma 20.

Lemma 21 (Estimator Reduction). *Let $\mathscr{T}\in\mathbb{T}$ be given with a subset $\mathscr{M}\subset\mathscr{T}$ of marked elements and let $\mathscr{T}_* = \mathsf{REFINE}(\mathscr{T},\mathscr{M})$.*

There exists a constant $\Lambda>0$, such that all $V\in\mathbb{V}(\mathscr{T})$, $V_*\in\mathbb{V}_*(\mathscr{T}_*)$ and any $\delta>0$ we have

$$\begin{aligned}\mathscr{E}^2_{\mathscr{T}_*}(V_*,\mathscr{T}_*) \le (1+\delta)\big(\mathscr{E}^2_{\mathscr{T}}(V,\mathscr{T}) - \lambda\,\mathscr{E}^2_{\mathscr{T}}(V,\mathscr{M})\big)\\ + (1+\delta^{-1})\,\Lambda\,\eta^2_{\mathscr{T}}(\boldsymbol{A},\mathscr{T})\,|||V_*-V|||^2_\Omega\,,\end{aligned}$$

where $\lambda = 1-2^{-b/d}$ and

$$\eta_{\mathscr{T}}(\boldsymbol{A},\mathscr{T}) := \max_{T\in\mathscr{T}}\eta_{\mathscr{T}}(\boldsymbol{A},T).$$

Proof. We proceed in several steps.

1 *Global Estimate.* We first observe that $V\in\mathbb{V}(\mathscr{T}_*)$ since the spaces are nested. We next invoke Lemma 20 for $T\in\mathscr{T}_*$ and $V,V_*\in\mathbb{V}(\mathscr{T}_*)$ to get

$$\mathscr{E}_{\mathscr{T}_*}(V_*,T) \le \mathscr{E}_{\mathscr{T}_*}(V,T) + C\,\eta_{\mathscr{T}_*}(\boldsymbol{A},T)\|V_*-V\|_{H^1(\omega_T)}.$$

Given $\delta>0$, we apply Young's inequality $(a+b)^2\le(1+\delta)a^2+(1+\delta^{-1})b^2$ and add over $T\in\mathscr{T}_*$ to arrive at

$$\mathscr{E}^2_{\mathscr{T}_*}(V_*,\mathscr{T}_*) \le (1+\delta)\,\mathscr{E}^2_{\mathscr{T}_*}(V,\mathscr{T}_*) + \Lambda\,(1+\delta^{-1})\eta^2_{\mathscr{T}}(\boldsymbol{A},\mathscr{T})\,|||V_*-V|||^2_\Omega\,. \quad (131)$$

Here, $\Lambda = (d+1)C/\alpha_1$ results from the finite overlapping property of sets ω_T and the relation between norms

$$\alpha_1 \|\nabla v\|^2_{L^2(\Omega)} \le |||v|||^2_\Omega \quad \text{for all } v \in \mathbb{V}.$$

In addition we have used the monotonicity property $\eta_{\mathscr{T}_*}(\boldsymbol{A}, \mathscr{T}_*) \le \eta_{\mathscr{T}}(\boldsymbol{A}, \mathscr{T})$.

2 *Accounting for* $\mathscr{M}$. We next decompose $\mathscr{E}^2_{\mathscr{T}_*}(V, \mathscr{T}_*)$ over elements $T \in \mathscr{T}$, and distinguish whether or not $T \in \mathscr{M}$. If $T \in \mathscr{M}$, then T is bisected at least b times and so T can be written as the union of elements $T' \in \mathscr{T}_*$ We denote this set of elements $\mathscr{T}_*(T)$ and observe $h_{T'} \le 2^{-b/d} h_T$ for all $T' \in \mathscr{T}_*(T)$. Therefore

$$\sum_{T' \in \mathscr{T}_*(T)} h^2_{T'} \|r(V)\|^2_{L^2(T')} \le 2^{-(2b)/d} h^2_T \|r(V)\|^2_{L^2(T)}$$

and

$$\sum_{T' \in \mathscr{T}_*(T)} h_{T'} \|j(V)\|^2_{L^2(\partial T' \cap \Omega)} \le 2^{-b/d} h_T \|j(V)\|^2_{L^2(\partial T \cap \Omega)}.$$

This implies

$$\mathscr{E}^2_{\mathscr{T}_*}(V,T) \le 2^{-b/d} \mathscr{E}^2_{\mathscr{T}}(V,T) \qquad \text{for all } T \in \mathscr{M}.$$

For the remaining elements $T \in \mathscr{T} \setminus \mathscr{M}$ we only know that mesh-size does not increased because $\mathscr{T} \le \mathscr{T}_*$, whence

$$\mathscr{E}^2_{\mathscr{T}_*}(V,T) \le \mathscr{E}^2_{\mathscr{T}}(V,T) \qquad \text{for all } T \in \mathscr{T} \setminus \mathscr{M}.$$

3 *Assembling.* Combining the two estimates we see that

$$\begin{aligned} \mathscr{E}^2_{\mathscr{T}_*}(V, \mathscr{T}_*) &\le 2^{-b/d} \mathscr{E}^2_{\mathscr{T}}(V, \mathscr{M}) + \mathscr{E}^2_{\mathscr{T}}(V, \mathscr{T} \setminus \mathscr{M}) \\ &= \mathscr{E}^2_{\mathscr{T}}(V, \mathscr{T}) - \big(1 - 2^{-b/d}\big) \mathscr{E}^2_{\mathscr{T}}(V, \mathscr{M}). \end{aligned}$$

Recalling the definition of $\lambda = 1 - 2^{-b/d}$ and replacing $\mathscr{E}^2_{\mathscr{T}_*}(V, \mathscr{T}_*)$ in (131) by the right hand side of this estimate yields the assertion. □

8.3 Contraction Property of AFEM

Recall that AFEM stands for the iteration loop (104) for the model problem. A key question to ask is what is (are) the quantity(ies) that AFEM may contract. In view of (129), an obvious candidate is the energy error $|||u - U_k|||_\Omega$. We show next that this may not be the case unless REFINE enforces several levels of refinement.

Example 2 (Interior Node). Let $\Omega = (0,1)^2$, $\boldsymbol{A} = I$, $f = 1$, and consider the sequence of meshes depicted in Fig. 15. If ϕ_0 denotes the basis function associated with the only interior node of $\mathscr{T}_0$, then

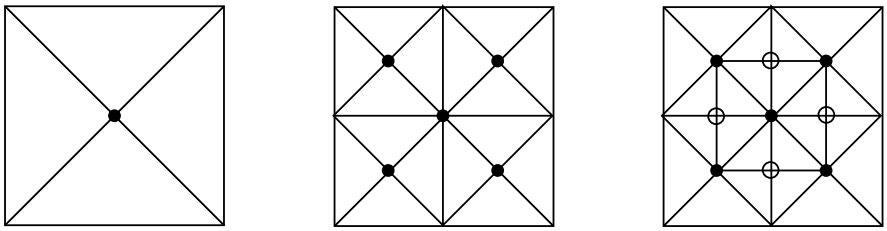

Fig. 15 Grids $\mathscr{T}_0$, $\mathscr{T}_1$, and $\mathscr{T}_2$ of the interior node example.

$$U_0 = U_1 = \frac{1}{12}\,\phi_0, \qquad U_2 \neq U_1.$$

The mesh $\mathscr{T}_1 \geq \mathscr{T}_0$ is produced by a standard 2-step bisection ($b = 2$) in $2d$. Since $U_0 = U_1$ we conclude that the energy error does not change

$$|||u - U_0|||_\Omega = |||u - U_1|||_\Omega$$

between consecutive steps of AFEM. This is no longer the case provided an interior node in each marked element is created, because then $U_2 \neq U_1$ and so $|||u - U_2|||_\Omega < |||u - U_1|||_\Omega$ (see (128)).

This example appeared first in [52, 53], and was used to justify the *interior node property: $\mathscr{T}_*$ must have one node in each side and interior of every $T \in \mathscr{M}$*. This property entails a minimal number of bisections that increases with the dimension d. The following heuristics explains why this property, closely related to a local discrete lower bound (see Problem (35)), is no longer needed in the present approach.

Heuristics. According to (128), the energy error is monotone, but the previous example shows that strict inequality may fail. However, in case $U_{k+1} = U_k$, the estimator reduction in Lemma 21 for $V_* = U_{k+1}$ and $V = U_k$ reveals a strict estimator reduction. We could thus expect that a suitable combination of them, the so-called *quasi error*

$$|||u - U_k|||_\Omega^2 + \gamma\,\mathscr{E}_k^2(U_k, \mathscr{T}_k),$$

may be contractive. This heuristics illustrates a distinct aspect of AFEM theory, the interplay between continuous quantities such the energy error $|||u - U_k|||_\Omega$ and discrete ones such as the estimator $\mathscr{E}_k(U_k, \mathscr{T}_k)$: no one alone has the requisite properties to yield a contraction between consecutive adaptive steps.

Theorem 17 (Contraction Property). *Let $\theta \in (0,1]$ be the Dörfler Marking parameter, and $\{\mathscr{T}_k, \mathbb{V}_k, U_k\}_{k=0}^\infty$ be a sequence of conforming meshes, finite element spaces and discrete solutions created by AFEM for the model problem* (127).

Then there exist constants $\gamma > 0$ and $0 < \alpha < 1$, additionally depending on the number b of bisections and θ, such that for all $k \geq 0$

$$|||u - U_{k+1}|||_\Omega^2 + \gamma\,\mathscr{E}_{k+1}^2(U_{k+1}, \mathscr{T}_{k+1}) \leq \alpha^2 \Big(|||u - U_k|||_\Omega^2 + \gamma\,\mathscr{E}_k^2(U_k, \mathscr{T}_k) \Big).$$

Proof. We split the proof into four steps. For convenience, we use the notation

$$e_k = |||u-U_k|||_\Omega\,,\ E_k = |||U_{k+1}-U_k|||_\Omega\,,\ \mathscr{E}_k = \mathscr{E}_k(U_k,\mathscr{T}_k),\ \mathscr{E}_k(\mathscr{M}_k) = \mathscr{E}_k(U_k,\mathscr{M}_k).$$

1 The error orthogonality (128) reads

$$e_{k+1}^2 = e_k^2 - E_k^2. \tag{132}$$

Employing Lemma 21 with $\mathscr{T} = \mathscr{T}_k$, $\mathscr{T}_* = \mathscr{T}_{k+1}$, $V = U_k$ and $V_* = U_{k+1}$ gives

$$\mathscr{E}_{k+1}^2 \le (1+\delta)\big(\mathscr{E}_k^2 - \lambda\,\mathscr{E}_k^2(\mathscr{M}_k)\big) + (1+\delta^{-1})\,\Lambda_0\,E_k^2, \tag{133}$$

where $\Lambda_0 = \Lambda\,\eta_{\mathscr{T}_0}^2(\boldsymbol{A},\mathscr{T}_0) \ge \Lambda\,\eta_{\mathscr{T}_k}^2(\boldsymbol{A},\mathscr{T}_k)$. After multiplying (133) by $\gamma>0$, to be determined later, we add (132) and (133) to obtain

$$e_{k+1}^2 + \gamma\,\mathscr{E}_{k+1}^2 \le e_k^2 + \big(\gamma(1+\delta^{-1})\Lambda_0 - 1\big)\,E_k^2 + \gamma(1+\delta)\,\big(\mathscr{E}_k^2 - \lambda\,\mathscr{E}_k^2(\mathscr{M}_k)\big).$$

2 We now choose the parameters δ,γ, the former so that

$$(1+\delta)\big(1-\lambda\,\theta^2\big) = 1 - \frac{\lambda\,\theta^2}{2},$$

and the latter to verify

$$\gamma(1+\delta^{-1})\,\Lambda_0 = 1.$$

Note that this choice of γ yields

$$e_{k+1}^2 + \gamma\,\mathscr{E}_{k+1}^2 \le e_k^2 + \gamma(1+\delta)\,\big(\mathscr{E}_k^2 - \lambda\,\mathscr{E}_k^2(\mathscr{M}_k)\big).$$

3 We next employ Dörfler Marking, namely $\mathscr{E}_k(\mathscr{M}_k) \ge \theta\,\mathscr{E}_k$, to deduce

$$e_{k+1}^2 + \gamma\,\mathscr{E}_{k+1}^2 \le e_k^2 + \gamma(1+\delta)(1-\lambda\,\theta^2)\mathscr{E}_k^2$$

which, in conjunction with the choice of δ, gives

$$e_{k+1}^2 + \gamma\,\mathscr{E}_{k+1}^2 \le e_k^2 + \gamma\left(1-\frac{\lambda\,\theta^2}{2}\right)\mathscr{E}_k^2 = e_k^2 - \frac{\gamma\lambda\,\theta^2}{4}\mathscr{E}_k^2 + \gamma\left(1-\frac{\lambda\,\theta^2}{4}\right)\mathscr{E}_k^2.$$

4 Finally, the upper bound (130a), namely $e_k^2 \le C_1\,\mathscr{E}_k^2$, implies that

$$e_{k+1}^2 + \gamma\,\mathscr{E}_{k+1}^2 \le \left(1-\frac{\gamma\lambda\,\theta^2}{4C_1}\right)e_k^2 + \gamma\left(1-\frac{\lambda\,\theta^2}{4}\right)\mathscr{E}_k^2.$$

This in turn leads to

$$e_{k+1}^2 + \gamma\,\mathscr{E}_{k+1}^2 \le \alpha^2\big(e_k^2 + \gamma\,\mathscr{E}_k^2\big),$$

with

$$\alpha^2 := \max\left\{1 - \frac{\gamma\lambda\theta^2}{4C_1}, 1 - \frac{\lambda\theta^2}{4}\right\},$$

and proves the theorem because $\alpha^2 < 1$. □

Remark 30 (Ingredients). This proof hinges on the following basic ingredients: Dörfler marking; symmetry of $\mathscr{B}$ and nesting of spaces, which imply the Pythagoras identity (Lemma 18); the a posteriori upper bound (Lemma 19); and the estimator reduction property (Lemma 21). It does not use the lower bound (130b) and does not require marking by oscillation, as previous proofs do [24, 48, 52, 53, 54]. The marking is driven by $\mathscr{E}_k$ exclusively, as it happens in all practical AFEM.

8.4 Example: Discontinuous Coefficients

We invoke the formulas derived by Kellogg [43] to construct an exact solution of an elliptic problem with piecewise constant coefficients and vanishing right-hand side f. We now write these formulas in the particular case $\Omega = (-1,1)^2$, $\boldsymbol{A} = a_1\boldsymbol{I}$ in the first and third quadrants, and $\boldsymbol{A} = a_2\boldsymbol{I}$ in the second and fourth quadrants. An exact weak solution u for $f \equiv 0$ is given in polar coordinates by $u(r,\theta) = r^\gamma \mu(\theta)$, where

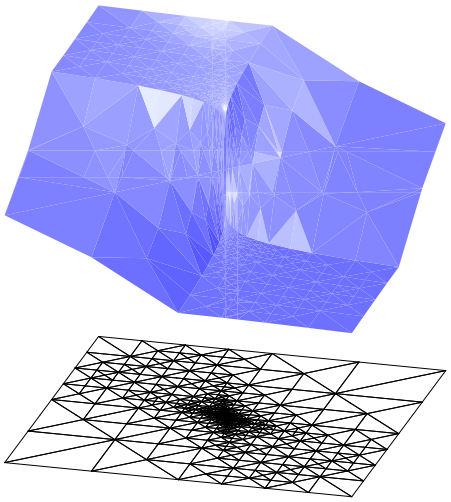

Fig. 16 Discontinuous coefficients in checkerboard pattern: Graph of the discrete solution, which is $u \approx r^{0.1}$, and underlying strongly graded grid. Notice the steep gradient of u at the origin.

$$\mu(\theta) = \begin{cases} \cos((\pi/2-\sigma)\gamma)\cdot\cos((\theta-\pi/2+\rho)\gamma) & \text{if } 0 \le \theta \le \pi/2, \\ \cos(\rho\gamma)\cdot\cos((\theta-\pi+\sigma)\gamma) & \text{if } \pi/2 \le \theta \le \pi, \\ \cos(\sigma\gamma)\cdot\cos((\theta-\pi-\rho)\gamma) & \text{if } \pi \le \theta < 3\pi/2, \\ \cos((\pi/2-\rho)\gamma)\cdot\cos((\theta-3\pi/2-\sigma)\gamma) & \text{if } 3\pi/2 \le \theta \le 2\pi, \end{cases}$$

and the numbers γ, ρ, σ satisfy the nonlinear relations

$$\begin{cases} R := a_1/a_2 = -\tan((\pi/2-\sigma)\gamma)\cdot\cot(\rho\gamma), \\ 1/R = -\tan(\rho\gamma)\cdot\cot(\sigma\gamma), \\ R = -\tan(\sigma\gamma)\cdot\cot((\pi/2-\rho)\gamma), \\ 0 < \gamma < 2, \\ \max\{0, \pi\gamma - \pi\} < 2\gamma\rho < \min\{\pi\gamma, \pi\}, \\ \max\{0, \pi - \pi\gamma\} < -2\gamma\sigma < \min\{\pi, 2\pi - \pi\gamma\}. \end{cases} \tag{134}$$

Since we want to test the algorithm AFEM in a worst case scenario, we choose $\gamma = 0.1$, which produces a very singular solution u that is barely in H^1; in fact $u \in H^s(\Omega)$ for $s < 1.1$ but still piecewise in $W_p^2(\Omega)$ for some $1 < p < \frac{20}{19}$ (see Figure 16). We then solve (134) for R, ρ, and σ using Newton's method to obtain within computing precision

$$R = a_1/a_2 \cong 161.4476387975881, \quad \rho = \pi/4, \quad \sigma \cong -14.92256510455152,$$

and finally choose $a_1 = R$ and $a_2 = 1$. A smaller γ would lead to a larger ratio R, but in principle γ may be as close to 0 as desired.

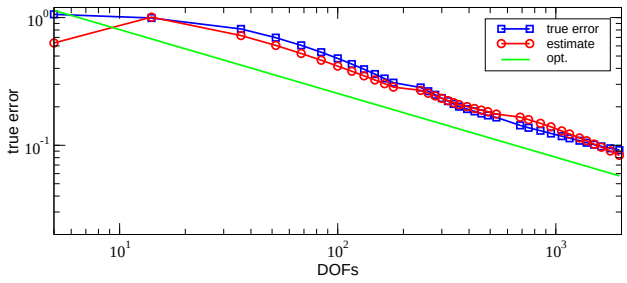

Fig. 17 Quasi-optimality of AFEM for discontinuous coefficients: estimate and true error. The optimal decay for piecewise linear elements in 2d is indicated by the green line with slope $-1/2$.

We realize from Fig. 17 that AFEM attains optimal decay rate for the energy norm. As we have seen in Sect. 5.4, this is consistent with adaptive approximation for functions piecewise in $W_p^2(\Omega)$, but nonobvious for AFEM which does not have direct access to u. We also notice from Fig. 18 that a graded mesh with mesh-size of order 10^{-10} at the origin is achieved with about 2×10^3 elements. To reach a similar resolution with a uniform mesh we would need $N \approx 10^{20}$ elements! This example

clearly reveals the advantages and potentials of adaptivity within the FEM even with modest computational resources.

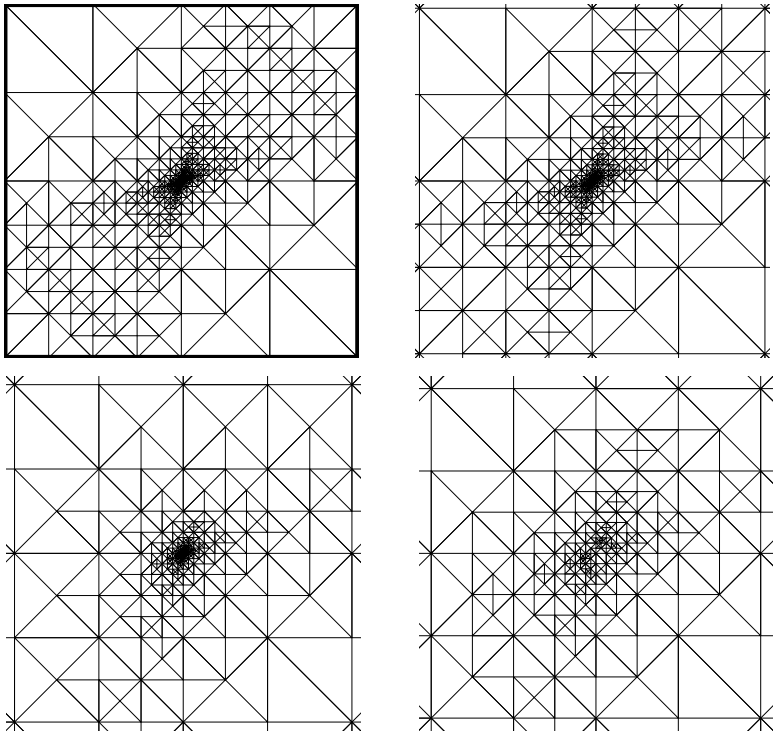

Fig. 18 Discontinuous coefficients in checkerboard pattern: Final grid (full grid with < 2000 nodes) (top left), zooms to $(-10^{-3}, 10^{-3})^2$ (top right), $(-10^{-6}, 10^{-6})^2$ (bottom left), and $(-10^{-9}, 10^{-9})^2$ (bottom right). The grid is highly graded towards the origin. For a similar resolution, a uniform grid would require $N \approx 10^{20}$ elements.

What is missing is an explanation of the recovery of optimal error decay $N^{-1/2}$ through mesh grading. This is the subject of Chap. 9, where we have to deal with the interplay between continuous and discrete quantities as already alluded to in the heuristics.

8.5 Problems

Problem 47 (Nesting of Spaces). If $\mathscr{T}_1, \mathscr{T}_2 \in \mathbb{T}$ satisfy $\mathscr{T}_1 \leq \mathscr{T}_2$, that is $\mathscr{T}_2$ is a refinement by bisection of $\mathscr{T}_1$, then the corresponding (Lagrange) finite element spaces are nested, i. e., $\mathbb{V}(\mathscr{T}_1) \subset \mathbb{V}(\mathscr{T}_2)$.

Problem 48 (Pythagoras). Let $\mathbb{V}_1 \subset \mathbb{V}_2 \subset \mathbb{V} = H^1_0(\Omega)$ be nested, conforming and closed subspaces. Let $u \in \mathbb{V}$ be the weak solution to (127), $U_1 \in \mathbb{V}_1$ and $U_2 \in \mathbb{V}_2$ the respective Ritz-Galerkin approximations to u. Prove the orthogonality property

$$|||u-U_1|||^2_\Omega = |||u-U_2|||^2_\Omega + |||U_2-U_1|||^2_\Omega\,. \tag{135}$$

Problem 49 (Error in Energy). Let $\mathbb{V}_1 \subset \mathbb{V}_2 \subset \mathbb{V}$ and U_1, U_2, u be as in Problem 48. Recalling Problem 7, we know that u, U_1, U_2 are the unique minimizer of the quadratic energy

$$I[v] := \tfrac{1}{2}\mathscr{B}[v,v] - \langle f, v\rangle$$

in $\mathbb{V}, \mathbb{V}_1, \mathbb{V}_2$ respectively. Show that (135) is equivalent to the identity

$$I[U_1] - I[u] = (I[U_2] - I[u]) + (I[U_1] - I[U_2]).$$

To this end prove

$$I[U_i] - I[u] = \tfrac{1}{2}\,|||U_i - u|||^2_\Omega \qquad \text{and} \qquad I[U_1] - I[U_2] = \tfrac{1}{2}\,|||U_1 - U_2|||^2_\Omega\,.$$

Problem 50. Let $S \in \mathring{\mathscr{S}}$ be a side of $T \in \mathscr{T}$, and let $\boldsymbol{A} \in W^1_\infty(T)$. Prove the following inverse estimate by a scaling argument to the reference element

$$\|\boldsymbol{A}\nabla V\|_S \lesssim h_S^{-1/2}\|\nabla V\|_T \quad \text{for all } V \in \mathbb{V}(\mathscr{T}),$$

where the hidden constant depends on the shape coefficient of $\mathscr{T}$, the dimension d, and $\|\boldsymbol{A}\|_{L^\infty(S)}$.

Problem 51. Let K be either a d or a $(d-1)$-simplex. For $\ell \in \mathbb{N}$ denote by $P^p_m\colon L^p(K,\mathbb{R}^\ell) \to \mathbb{P}_m(K,\mathbb{R}^\ell)$ the operator of best L^p-approximation in K. Then for all $v \in L^\infty(K,\mathbb{R}^\ell)$, $V \in \mathbb{P}_n(K,\mathbb{R}^\ell)$ and $m \geq n$, there holds

$$\|vV - P^2_m(vV)\|_{L^2(K)} \leq \|v - P^\infty_{m-n}v\|_{L^\infty(K)}\|V\|_{L^2(K)}.$$

Problem 52. Let $\boldsymbol{A} \in W^1_\infty(T)$ for all $T \in \mathscr{T}$. Prove the quasi-local Lipschitz property

$$|\operatorname{osc}_{\mathscr{T}}(V,T) - \operatorname{osc}_{\mathscr{T}}(W,T)| \lesssim \operatorname{osc}_{\mathscr{T}}(\boldsymbol{A},T)\,\|V-W\|_{H^1(\omega_T)} \qquad \text{for all } V,W \in \mathbb{V},$$

where $\operatorname{osc}_{\mathscr{T}}(\boldsymbol{A},T) = h_T\,\|\operatorname{div}\boldsymbol{A} - P^\infty_{n-1}(\operatorname{div}\boldsymbol{A})\|_{L^\infty(T)} + \|\boldsymbol{A} - P^\infty_n\boldsymbol{A}\|_{L^\infty(\omega_T)}$. Proceed as in the proof of Lemma 20 and use Problem 51.

Problem 53. Let $\mathscr{T}, \mathscr{T}_* \in \mathbb{T}$, with $\mathscr{T} \leq \mathscr{T}_*$. Use Problem 52 to prove that, for all $V \in \mathbb{V}(\mathscr{T})$ and $V_* \in \mathbb{V}(\mathscr{T}_*)$, there is a constant $\Lambda_1 > 0$ such that

$$\operatorname{osc}^2_{\mathscr{T}}(V, \mathscr{T}\cap\mathscr{T}_*) \leq 2\operatorname{osc}^2_{\mathscr{T}_*}(V_*, \mathscr{T}\cap\mathscr{T}_*) + \Lambda_1 \operatorname{osc}_{\mathscr{T}_0}(\boldsymbol{A},\mathscr{T}_0)^2\,|||V-V_*|||^2_\Omega\,.$$

9 Adaptivity: Convergence Rates

We have already realized in Chap. 5 that we can a priori accommodate the degrees of freedom in such a way that the finite element approximation retains optimal energy error decay for a class of singular functions. This presumes knowledge of the exact solution u. At the same time, we have seen numerical evidence in Sect. 8.4 that the AFEM of Chap. 8, achieves such a performance without direct access to the regularity of u. Practical experience strongly suggests that this is even true for a much larger class of problems and adaptive methods. The challenge ahead is to reconcile these two distinct aspects of AFEM. In doing this we have to restrict ourselves to the setting of Chap. 8. The mathematical foundation to justify the observed optimal error decay of adaptive methods in case of non-symmetric or non-coercive bilinear forms and other marking strategies is completely open.

One key to connect the two worlds for the simplest scenario, the Laplacian and f piecewise constant, is due to Stevenson [69]: *any marking strategy that reduces the energy error relative to the current value must contain a substantial bulk of $\mathscr{E}_{\mathscr{T}}(U,\mathscr{T})$, and so it can be related to Dörfler Marking*. This allows us to compare AFEM with an optimal mesh choice and to conclude optimal error decay.

The objective of this section is to study the model problem (127) for general data f and $\boldsymbol{A}$ and the AFEM from Chap. 8. In what follows it is important to use an error notion that is strictly reduced by the adaptive method. In this section we closely follow Cascón et al. [21] by utilizing the quasi-error as contracting quantity. This approach allows us to include variable data f and $\boldsymbol{A}$ and thus improves upon and extends Stevenson [69].

As in Chap. 8, u will always be the weak solution of (127) and, except when stated otherwise, any constant explicit or hidden constant in $\lesssim$ may depend on the uniform shape-regularity of $\mathbb{T}$, the dimension d, the polynomial degree n, the (global) eigenvalues of $\boldsymbol{A}$, and the oscillation $\operatorname{osc}_{\mathscr{T}_0}(\boldsymbol{A},\mathscr{T}_0)$ of $\boldsymbol{A}$ on the initial mesh $\mathscr{T}_0$, but not on a specific grid $\mathscr{T}\in\mathbb{T}$.

9.1 Approximation Class

Since AFEM selects elements for refinement based on information provided exclusively by the error indicators $\{\mathscr{E}_{\mathscr{T}}(U,T)\}_{T\in\mathscr{T}}$, it is natural that the measure of regularity and ensuing decay rate is closely related to the indicators. We explore this connection now.

The Total Error. We first introduce the concept of *total error* [48]

$$|||u-U|||_{\Omega}^2+\operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T}),$$

and next assert that it is equivalent to the quasi error, for the Galerkin function $U\in\mathbb{V}(\mathscr{T})$. In fact, in view of the upper and lower a posteriori error bounds (130a)

and (130b), and

$$\operatorname{osc}^2_{\mathscr{T}}(U,\mathscr{T}) \leq \mathscr{E}^2_{\mathscr{T}}(U,\mathscr{T})$$

we have

$$\begin{aligned} C_2\,\mathscr{E}^2_{\mathscr{T}}(U,\mathscr{T}) &\leq |||u-U|||^2_{\Omega} + \operatorname{osc}^2_{\mathscr{T}}(U,\mathscr{T}) \\ &\leq |||u-U|||^2_{\Omega} + \mathscr{E}^2_{\mathscr{T}}(U,\mathscr{T}) \leq (1+C_1)\,\mathscr{E}^2_{\mathscr{T}}(U,\mathscr{T}), \end{aligned}$$

whence

$$\mathscr{E}^2_{\mathscr{T}}(U,\mathscr{T}) \approx |||u-U|||^2_{\Omega} + \operatorname{osc}^2_{\mathscr{T}}. \tag{136}$$

We thus realize that the decay rate of AFEM must be characterized by the total error. Moreover, on invoking the upper bound once again, we also see that the total error is equivalent to the quasi error

$$|||u-U|||^2_{\Omega} + \operatorname{osc}^2_{\mathscr{T}}(U,\mathscr{T}) \approx |||u-U|||^2_{\Omega} + \mathscr{E}^2_{\mathscr{T}}(U,\mathscr{T}).$$

This is the quantity being strictly reduced by AFEM (Theorem 17). Finally, the total error satisfies the following Cea's type-lemma. In fact, if $\boldsymbol{A}$ is piecewise constant, then this is Cea's Lemma stated in Problem 11.

Lemma 22 (Quasi-Optimality of Total Error). *There exists a constant Λ_2, such that for any $\mathscr{T} \in \mathbb{T}$ and the corresponding Ritz–Galerkin solution $U \in \mathbb{V}(\mathscr{T})$ holds*

$$|||u-U|||^2_{\Omega} + \operatorname{osc}^2_{\mathscr{T}}(U,\mathscr{T}) \leq \Lambda_2 \inf_{V\in\mathbb{V}(\mathscr{T})} \Big(|||u-V|||^2_{\Omega} + \operatorname{osc}^2_{\mathscr{T}}(V,\mathscr{T}) \Big).$$

Proof. For $\varepsilon > 0$ choose $V_\varepsilon \in \mathbb{V}(\mathscr{T})$, with

$$|||u-V_\varepsilon|||^2_{\Omega} + \operatorname{osc}^2_{\mathscr{T}}(V_\varepsilon,\mathscr{T}) \leq (1+\varepsilon) \inf_{V\in\mathbb{V}(\mathscr{T})} \Big(|||u-V|||^2_{\Omega} + \operatorname{osc}^2_{\mathscr{T}}(V,\mathscr{T}) \Big).$$

Applying Problem 53 with $\mathscr{T}_* = \mathscr{T}$, $V = U$, and $V_* = V_\varepsilon$ yields

$$\operatorname{osc}^2_{\mathscr{T}}(U,\mathscr{T}) \leq 2\,\operatorname{osc}^2_{\mathscr{T}}(V_\varepsilon,\mathscr{T}) + C_3\,|||U-V_\varepsilon|||^2_{\Omega},$$

with

$$C_3 := \Lambda_1 \operatorname{osc}_{\mathscr{T}_0}(\boldsymbol{A},\mathscr{T}_0)^2.$$

Since $U \in \mathbb{V}(\mathscr{T})$ is the Galerkin solution, $U - V_\varepsilon \in \mathbb{V}(\mathscr{T})$ is orthogonal to $u-U$ in the energy norm, whence $|||u-U|||^2_{\Omega} + |||U-V_\varepsilon|||^2_{\Omega} = |||u-V_\varepsilon|||^2_{\Omega}$ and

$$\begin{aligned} |||u-U|||^2_{\Omega} + \operatorname{osc}^2_{\mathscr{T}}(U,\mathscr{T}) &\leq \big(1+C_3\big)\,|||u-V_\varepsilon|||^2_{\Omega} + 2\,\operatorname{osc}^2_{\mathscr{T}}(V_\varepsilon,\mathscr{T}) \\ &\leq (1+\varepsilon)\,\Lambda_2 \inf_{V\in\mathbb{V}(\mathscr{T})} \Big(|||u-U|||^2_{\Omega} + \operatorname{osc}^2_{\mathscr{T}}(V,\mathscr{T}) \Big), \end{aligned}$$

with $\Lambda_2 = \max\big\{2, 1+C_3\big\}$, and the assertion follows from $\varepsilon \to 0$. □

We next give a definition of an appropriate approximation class $\mathbb{A}_s$ that hinges on the concept of total error. We first let $\mathbb{T}_N \subset \mathbb{T}$ be the set of all possible conforming refinements of $\mathscr{T}_0$ with at most N elements more than $\mathscr{T}_0$, i. e.,

$$\mathbb{T}_N = \{\mathscr{T} \in \mathbb{T} \mid \#\mathscr{T} - \#\mathscr{T}_0 \le N\}.$$

The quality of the best approximation in $\mathbb{T}_N$ with respect to the total error is characterized by

$$\sigma(N;u,f,\boldsymbol{A}) := \inf_{\mathscr{T}\in\mathbb{T}_N} \inf_{V\in\mathbb{V}(\mathscr{T})} \big(|||u-V|||_\Omega^2 + \operatorname{osc}_{\mathscr{T}}^2(V,\mathscr{T})\big)^{1/2},$$

and the approximation class $\mathbb{A}_s$ for $s>0$ is defined by

$$\mathbb{A}_s := \Big\{ (v,f,\boldsymbol{A}) \mid |v,f,\boldsymbol{A}|_s := \sup_{N>0} \big(N^s\, \sigma(N;v,f,\boldsymbol{A})\big) < \infty \Big\}.$$

Thanks to Lemma 22, the solution u with data $(f,\boldsymbol{A})$ satisfies

$$\sigma(N;u,f,\boldsymbol{A}) \approx \inf_{\mathscr{T}\in\mathbb{T}_N} \big\{ \mathscr{E}_{\mathscr{T}}(U,\mathscr{T}) \mid U = \mathsf{SOLVE}(\mathbb{V}(\mathscr{T}))\big\}. \tag{137}$$

We point out the upper bound $s \le n/d$ for polynomial degree $n \ge 1$; this can be seen with full regularity and uniform refinement (see (69)). Note that if $(v,f,\boldsymbol{A}) \in \mathbb{A}_s$ then for all $\varepsilon > 0$ there exist $\mathscr{T}_\varepsilon \ge \mathscr{T}_0$ conforming and $V_\varepsilon \in \mathbb{V}(\mathscr{T}_\varepsilon)$ such that (see Problem 54)

$$|||v-V_\varepsilon|||_\Omega^2 + \operatorname{osc}_{\mathscr{T}_\varepsilon}^2 \le \varepsilon^2 \qquad \text{and} \qquad \#\mathscr{T}_\varepsilon - \#\mathscr{T}_0 \le |v,f,\boldsymbol{A}|_s^{1/s}\, \varepsilon^{-1/s}. \tag{138}$$

For the subsequent discussion we recall Lemma 6: the overlay $\mathscr{T}_1 \oplus \mathscr{T}_2 \in \mathbb{T}$ of two meshes $\mathscr{T}_1, \mathscr{T}_2 \in \mathbb{T}$ is the smallest common refinement of $\mathscr{T}_1$ and $\mathscr{T}_2$ and

$$\#\mathscr{T}_1 \oplus \mathscr{T}_2 \le \#\mathscr{T}_1 + \#\mathscr{T}_2 - \#\mathscr{T}_0. \tag{139}$$

We first investigate the class $\mathbb{A}_s$ for piecewise constant coefficient matrix $\boldsymbol{A}$ with respect to $\mathscr{T}_0$. In this simplified scenario, the oscillation $\operatorname{osc}_{\mathscr{T}}(U,\mathscr{T})$ reduces to *data oscillation* (see Remark 25):

$$\operatorname{osc}_{\mathscr{T}} = \|h_{\mathscr{T}}(f - P_{2n-2}\, f)\|_{L^2(\Omega)}.$$

We then have the following characterization of $\mathbb{A}_s$ in terms of the standard approximation classes [13, 14, 69]:

$$\mathscr{A}_s := \Big\{ v \in \mathbb{V} \mid \ |v|_{\mathscr{A}_s} := \sup_{N>0} \big(N^s \inf_{\mathscr{T}\in\mathbb{T}_N} \inf_{V\in\mathbb{V}(\mathscr{T})} |||v-V|||_\Omega\big) < \infty \Big\},$$

$$\bar{\mathscr{A}_s} := \Big\{ g \in L^2(\Omega) \mid \ |g|_{\bar{\mathscr{A}_s}} := \sup_{N>0} \big(N^s \inf_{\mathscr{T}\in\mathbb{T}_N} \|h_{\mathscr{T}}(g - P_{2n-2}\, g)\|_{L^2(\Omega)}\big) < \infty \Big\}.$$

Lemma 23 (Equivalence of Classes). *Let $\boldsymbol{A}$ be piecewise constant over $\mathscr{T}_0$. Then $(u,f,\boldsymbol{A}) \in \mathbb{A}_s$ if and only if $(u,f) \in \mathscr{A}_s \times \mathscr{A}_s^{\bar{}}$ and*

$$|u,f,\boldsymbol{A}|_s \approx |u|_{\mathscr{A}_s} + |f|_{\mathscr{A}_s^{\bar{}}}. \tag{140}$$

Proof. It is obvious that $(u,f,\boldsymbol{A}) \in \mathbb{A}_s$ implies $(u,f) \in \mathscr{A}_s \times \mathscr{A}_s^{\bar{}}$ as well as the bound $|u|_{\mathscr{A}_s} + |f|_{\mathscr{A}_s^{\bar{}}} \lesssim |u,f,\boldsymbol{A}|_s$.

In order to prove the reverse inequality, let $(u,f) \in \mathscr{A}_s \times \mathscr{A}_s^{\bar{}}$. Then there exist $\mathscr{T}_1, \mathscr{T}_2 \in \mathbb{T}_N$ so that $|||u-U|||_\Omega \le |u|_{\mathscr{A}_s} N^{-s}$ where $U \in \mathbb{V}(\mathscr{T}_1)$ is the best approximation and $\|h_{\mathscr{T}_2}(f - P^2_{2n-2} f)\|_{L^2(\Omega)} \le |f|_{\mathscr{A}_s^{\bar{}}} N^{-s}$.

The overlay $\mathscr{T} = \mathscr{T}_1 \oplus \mathscr{T}_2 \in \mathbb{T}_{2N}$ according to (139), and

$$|||u-U|||^2_\Omega + \operatorname{osc}^2_{\mathscr{T}} \le |||u-U|||^2_\Omega + \operatorname{osc}^2_{\mathscr{T}_2} \le 2^s \big(|u|^2_{\mathscr{A}_s} + |f|^2_{\mathscr{A}_s^{\bar{}}} \big) (2N)^{-s}.$$

This yields $(u,f,\boldsymbol{A}) \in \mathbb{A}_s$ together with the bound $|u,f,\boldsymbol{A}|_s \lesssim |u|_{\mathscr{A}_s} + |f|_{\mathscr{A}_s^{\bar{}}}$. □

We next turn to the special case of linear finite elements.

Corollary 11 (Membership in $\mathbb{A}_{1/2}$). *Let $d = 2$, polynomial degree $n = 1$, $f \in L^2(\Omega)$, and $\boldsymbol{A}$ piecewise constant with respect to $\mathscr{T}_0$. If $u \in W^2_p(\Omega;\mathscr{T}_0)$ for some $p > 1$, then $(u,f,\boldsymbol{A}) \in \mathbb{A}_{1/2}$ and*

$$|u,f,\boldsymbol{A}|_{1/2} \lesssim \|D^2 u\|_{L^p(\Omega;\mathscr{T}_0)} + \|f\|_{L^2(\Omega)}$$

Proof. We start with the data oscillation $\operatorname{osc}_{\mathscr{T}}$, and realize that

$$\operatorname{osc}_{\mathscr{T}} = \|h_{\mathscr{T}}(f - P_0 f)\|_{L^2(\Omega)} \le h_{\max}(\mathscr{T}) \|f\|_{L^2(\Omega)} \lesssim (\#\mathscr{T})^{-1/2} \|f\|_{L^2(\Omega)},$$

for any uniform refinement $\mathscr{T} \in \mathbb{T}$. This implies $f \in \mathscr{A}_{1/2}^{\bar{}}$ with $|f|_{\mathscr{A}_{1/2}^{\bar{}}} \lesssim \|f\|_{L^2(\Omega)}$.

For $u \in W^2_p(\Omega;\mathscr{T}_0)$ we learn from Corollary 7 and Remark 21 that $u \in \mathscr{A}_{1/2}$ and $|u|_{\mathscr{A}_{1/2}} \lesssim \|D^2 u\|_{L^2(\Omega;\mathscr{T}_0)}$. The assertion then follows from Lemma 23. □

Example 3 (Pre-asymptotics). Corollary 11 shows that oscillation decays at least with rate $1/2$ for $f \in L^2(\Omega)$. Since the decay rate of the total error is $s \le 1/2$, oscillation can be ignored asymptotically. However, Remark 26 shows that oscillation may dominate the total error, or equivalently the class $\mathbb{A}_s$ may fail to describe the behavior of $|||u-U_k|||_\Omega$, in the early stages of adaptivity. In fact, we recall that $\operatorname{osc}_k(U_k,\mathscr{T}_k) = \|h_k(f - P_0 f)\|_{L^2(\Omega)}$, the discrete solution $U_k = 0$, and $|||u-U_k|||_\Omega \approx 2^{-K}$ is constant for as many steps $k \le K$ as desired. In contrast, $\mathscr{E}_k(U_k,\mathscr{T}_k) = \operatorname{osc}_k(U_k,\mathscr{T}_k) = \|h_k f\|_{L^2(\Omega)}$ reduces strictly for $k \le K$ but overestimates $|||u-U_k|||_\Omega$. The fact that the preasymptotic regime $k \le K$ for the energy error could be made arbitrarily long would be problematic if we focus exclusively on $|||u-U_k|||_\Omega$. In practice, this effect is typically less dramatic because f is not orthogonal to $\mathbb{V}(\mathscr{T}_k)$. Figure 19 displays the behavior of AFEM for the smooth solution $u = u_S$ given by

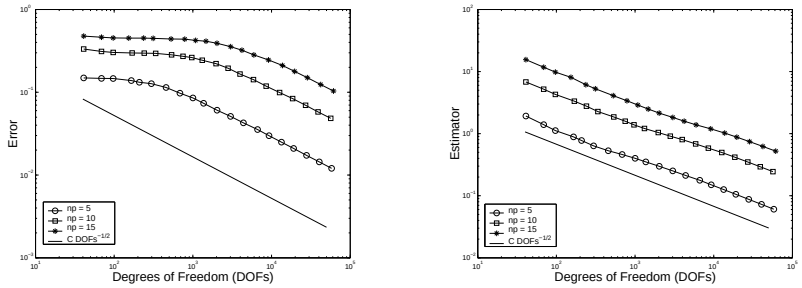

Fig. 19 Decay of the energy error (left) and the estimator (right) for the smooth solution u_S of (141) with frequencies $\kappa = 5, 10$, and 15. The energy error exhibits a frequency-dependent plateau in the preasymptotic regime and later an optimal decay. This behavior is allowed by $\mathbb{A}_s$.

$$u_S(x,y) = 10^{-2} a_i^{-1}(x^2+y^2)\sin^2(\kappa\pi x)\sin^2(\kappa\pi y), \qquad 1 \le i \le 4. \tag{141}$$

of the problem in Sect. 8.4 with discontinuous coefficients $\{a_i\}_{i=1}^4$ in checkerboard pattern and frequencies $\kappa = 5, 10$, and 15. We can see that the error exhibits a frequency-dependent plateau in the preasymptotic regime and later an optimal decay. In contrast, the estimator decays always with the optimal rate. Since all decisions of the AFEM are based on the estimator, this behavior has to be expected and is consistent with our notion of approximation class $\mathbb{A}_s$, which can be characterized just by the estimator according to (137).

We next turn to the nonlinear interaction encoded in $\operatorname{osc}_{\mathscr{T}}(U,\mathscr{T})$ via the product $\boldsymbol{A}\nabla U$. It is this interaction which makes the class $\mathbb{A}_s$ a non-standard object in approximation theory that deserves further scrutiny.

Lemma 24 (Decay Rate of Oscillation). *Let $\boldsymbol{A}$ be piecewise Lipschitz with respect to $\mathscr{T}_0$, $f \in L^2(\Omega)$, and polynomial degree $n = 1$. If $U \in \mathbb{V}(\mathscr{T})$ is the Ritz-Galerkin solution, then oscillation $\operatorname{osc}_{\mathscr{T}}(U,\mathscr{T})$ has at least a decay rate of order $-1/d$*

$$\inf_{\mathscr{T}\in\mathbb{T}_N} \operatorname{osc}_{\mathscr{T}}(U,\mathscr{T}) \lesssim \Big(\|f\|_{L^2(\Omega)} + \|\boldsymbol{A}\|_{W^1_\infty(\Omega;\mathscr{T}_0)}\Big) N^{-1/d}.$$

Proof. Let $\mathscr{T} \in \mathbb{T}_N$ be a uniform refinement of $\mathscr{T}_0$ with $\#\mathscr{T} \approx N$. By applying Problem 52 with $V = U$ and $W = 0$, we obtain

$$\operatorname{osc}_{\mathscr{T}}(U,T) \lesssim h_T \|f - P_0^2 f\|_{L^2(T)} + \operatorname{osc}_{\mathscr{T}}(\boldsymbol{A},T)\|U\|_{H^1(\omega_T)}$$

with $h_T\|f - P_0^2 f\|_{L^2(T)} \le h_T\|f\|_{L^2(T)}$ and

$$\operatorname{osc}_{\mathscr{T}}(\boldsymbol{A},T) = h\|\operatorname{div}\boldsymbol{A} - P_0^\infty(\operatorname{div}\boldsymbol{A})\|_{L^\infty(T)} + \|\boldsymbol{A} - P_1^\infty\boldsymbol{A}\|_{L^\infty(\omega_T)} \lesssim h_T\|\boldsymbol{A}\|_{W^1_\infty(\omega_T;\mathscr{T}_0)}.$$

Uniform refinement yields the relation $h_T \approx N^{-1/d}$ for all $T \in \mathscr{T}$, whence

$$\operatorname{osc}^2_{\mathscr{T}}(U,\mathscr{T}) = \sum_{T\in\mathscr{T}} \operatorname{osc}^2_{\mathscr{T}}(U,T) \lesssim \Big(\|f\|^2_{L^2(\Omega)} + \|\boldsymbol{A}\|^2_{W^1_\infty(\Omega;\mathscr{T}_0)}\Big) N^{-2/d},$$

because $\|U\|_{H^1(\Omega)} \leq \alpha_1^{-1}\|f\|_{L^2(\Omega)}$ according to (40). □

Remark 31 (Asymptotic Order of Oscillation). Let's assume the following piecewise regularity of data $(f,\boldsymbol{A})$ with respect to a conforming refinement $\mathscr{T}_*$ of $\mathscr{T}_0$:

$$f \in H^1(\Omega;\mathscr{T}_*), \quad \boldsymbol{A} \in W^2_\infty(\Omega;\mathscr{T}_*).$$

The proof of Lemma 24, in conjunction with Proposition 3(a), shows that for $n = 1$

$$\inf_{\mathscr{T}\in\mathbb{T}_N:\mathscr{T}\geq\mathscr{T}_*} \operatorname{osc}_{\mathscr{T}}(U,\mathscr{T}) \lesssim \Big(\|f\|_{H^1(\Omega;\mathscr{T}_*)} + \|\boldsymbol{A}\|_{W^2_\infty(\Omega;\mathscr{T}_*)}\Big)N^{2/d},$$

and the rate in Lemma 24 can be improved. Since the energy error decay is never better than $N^{-1/d}$, according to (69), we realize that oscillation is of higher order than the energy error asymptotically as $N \uparrow \infty$; compare with Remark 23.

Corollary 12 (Membership in $\mathbb{A}_{1/2}$). *Let $d = 2$, polynomial degree $n = 1$, $\boldsymbol{A} \in W^1_\infty(\Omega;\mathscr{T}_0)$, and $f \in L^2(\Omega)$. If $u \in W^2_p(\Omega;\mathscr{T}_0)$ for some $p > 1$, then $(u,f,\boldsymbol{A}) \in \mathbb{A}_{1/2}$ and*

$$|u,f,\boldsymbol{A}|_{1/2} \lesssim \|D^2u\|_{L^p(\Omega;\mathscr{T}_0)} + \|\boldsymbol{A}\|_{W^1_\infty(\Omega;\mathscr{T}_0)} + \|f\|_{L^2(\Omega)}.$$

Proof. Repeat the proof of Corollary 11 with the help of Lemma 24. □

A complete characterization of $\mathbb{A}_s$ for general d and n is still missing. It is important to realize that the nonlinear interaction between data $\boldsymbol{A}$ and U must be accounted for, thereby leading to a new concept of approximation class $\mathbb{A}_s$, which generalizes those in [13, 14, 69]. It is worth mentioning that a near characterization of the standard approximation class $\mathscr{A}_s$ in terms of Besov spaces for $d = 2$ can be found in [13, 14, 37]: $u \in \mathscr{A}_s$ implies that $u \in B^{2s+1}_p(L^p(\Omega))$ for $p = \frac{2}{2s+1}$ [13, Theorem 9.3]; $u \in B^{2s+1}_p(L^p(\Omega))$ for $p > \frac{2}{2s+1}$ implies that $u \in \mathscr{A}_s$ [13, Theorem 9.1]. Note that $p < 1$ for $s > 1/2$; see Remark 22.

9.2 Cardinality of $\mathscr{M}_k$

To assess the performance of AFEM in terms of degrees of freedom $\#\mathscr{T}_k$, we need to impose further restrictions on the modules of AFEM beyond those of Sect. 8.1. We recall that $C_2 \leq C_1$ are the constants in (130a) and (130b) and $C_3 = \Lambda_1 \operatorname{osc}^2_{\mathscr{T}_0}(\boldsymbol{A},\mathscr{T}_0)$ is the constant in Problem 53 and Lemma 22.

Assumption 2 (Assumptions for Optimal Decay Rate). *We assume the following additional properties of the marking procedure* MARK *and the initial grid $\mathscr{T}_0$:*

(a) *The marking parameter θ of Dörfler Marking satisfies $\theta \in (0,\theta^*)$ with*

$$\theta_*^2 = \frac{C_2}{1+C_1(1+C_3)};$$

(b) *MARK outputs a set $\mathscr{M}$ with minimal cardinality;*

(c) *The initial triangulation $\mathscr{T}_0$ satisfies Assumption 1.*

A few comments are now in order.

- *Threshold $\theta_* < 1$*: We first point out that, according to (130a) and (130b), the ratio $C_2/C_1 \leq 1$ is a quality measure of the estimator $\mathscr{E}_{\mathscr{T}}(U,\mathscr{T})$: the closer to 1 the better! It is thus natural to be cautious in marking if the reliability constant C_1 and efficiency constant C_2 are very disparate. The additional factor C_3 accounts for the effect of a function dependent oscillation (see Problem 53), and is zero if the oscillation just depends on data f because then $\operatorname{osc}_{\mathscr{T}_0}(\boldsymbol{A},\mathscr{T}_0) = 0$.
- *Minimal $\mathscr{M}_k$*: According to Remark 24 about the significance of the local lower a posteriori error estimate for relatively small oscillation, it is natural to mark elements with largest error indicators. This leads to a minimal set $\mathscr{M}_k$ and turns out to be crucial to link AFEM with optimal meshes and approximation classes.
- *Initial Triangulation*: The initial labeling of the element's vertices on $\mathscr{T}_0$ stated in Assumption 1 of Sect.4.2 is rather restrictive for dimension $d > 2$ but guarantees the complexity estimate of Theorem 10 for our module REFINE. Any other refinement ensuing the same complexity estimate can replace REFINE together with the assumption on $\mathscr{T}_0$.

We stress that we cannot expect local upper bounds between the continuous solution u and discrete solution U due to the global nature of the underlying PDE: the error in a region may be dictated by pollution effects arising somewhere else. The following crucial result shows, however, that this is a matter of scale: if $\mathscr{T}_* \geq \mathscr{T}$, then what determines the error between Galerkin solutions $U \in \mathbb{V}(\mathscr{T})$ and $U_* \in \mathbb{V}(\mathscr{T}_*)$ is the refined set $\mathscr{R}_{\mathscr{T}\to\mathscr{T}_*}$, namely the region of Ω where the scale of resolution differs from $\mathscr{T}$ to $\mathscr{T}_*$. This is not, of course, in contradiction with the previous statement because one needs an infinitely fine scale to reach the exact solution u.

Lemma 25 (Localized Upper Bound). *Let $\mathscr{T},\mathscr{T}_* \in \mathbb{T}$ satisfy $\mathscr{T}_* \geq \mathscr{T}$ and define $\mathscr{R} := \mathscr{R}_{\mathscr{T}\to\mathscr{T}_*}$ to be the set of refined elements in $\mathscr{T}$. If $U \in \mathbb{V}(\mathscr{T})$ and $U_* \in \mathbb{V}(\mathscr{T}_*)$ are the corresponding Galerkin solutions, then*

$$|||U_* - U|||_\Omega^2 \leq C_1\, \mathscr{E}_{\mathscr{T}}^2(U,\mathscr{R}).$$

where $C_1 > 0$ is the same constant as in (130a)*.*

Proof. Problem 55. □

The following result reveals the importance of Dörfler's marking in the present context. The original result, established by Stevenson [69], referred to the energy error alone. We follow [21] in this analysis.

Lemma 26 (Optimal Marking). *Let the marking parameter θ satisfy Assumption 2(a) and set $\mu := \frac{1}{2}(1 - \frac{\theta^2}{\theta_*^2}) > 0$. For $\mathscr{T}_* \geq \mathscr{T}$ let the corresponding Galerkin solution $U \in \mathbb{V}(\mathscr{T})$ and $U_* \in \mathbb{V}(\mathscr{T}_*)$ satisfy*

$$|||u-U_*|||_\Omega^2+\operatorname{osc}_{\mathscr{T}_*}^2(U_*,\mathscr{T}_*)\le\mu\big(|||u-U|||_\Omega^2+\operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T})\big). \tag{142}$$

Then the set $\mathscr{R}=\mathscr{R}_{\mathscr{T}\to\mathscr{T}_*}$ of refined elements of $\mathscr{T}$ satisfies the Dörfler property

$$\mathscr{E}_{\mathscr{T}}(U,\mathscr{R})\ge\theta\,\mathscr{E}_{\mathscr{T}}(U,\mathscr{T}). \tag{143}$$

Proof. We split the proof into four steps.
1 In view of the global lower bound (130b)

$$C_2\,\mathscr{E}_{\mathscr{T}}^2(U,\mathscr{T})\le|||u-U|||_\Omega^2+\operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T})$$

and (142), we can write

$$\begin{aligned}(1-2\mu)C_2\,\mathscr{E}_{\mathscr{T}}^2(U,\mathscr{T})&\le(1-2\mu)\big(|||u-U|||_\Omega^2+\operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T})\big)\\&\le\big(|||u-U|||_\Omega^2-2\,|||u-U_*|||_\Omega^2\big)+\big(\operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T})-2\operatorname{osc}_{\mathscr{T}_*}^2(U_*,\mathscr{T}_*)\big).\end{aligned}$$

2 Combining the orthogonality relation (128)

$$|||u-U|||_\Omega^2-|||u-U_*|||_\Omega^2=|||U-U_*|||_\Omega^2.$$

with the localized upper bound Lemma 25 yields

$$|||u-U|||_\Omega^2-2\,|||u-U_*|||_\Omega^2\le C_1\,\mathscr{E}_{\mathscr{T}}^2(U,\mathscr{R}).$$

3 To deal with oscillation we decompose the elements of $\mathscr{T}$ into two disjoint sets: $\mathscr{R}$ and $\mathscr{T}\setminus\mathscr{R}$. In the former case, we have

$$\operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{R})-2\operatorname{osc}_{\mathscr{T}_*}^2(U_*,\mathscr{R})\le\operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{R})\le\mathscr{E}_{\mathscr{T}}^2(U,\mathscr{R}),$$

because $\operatorname{osc}_{\mathscr{T}}(U,T)\le\mathscr{E}_{\mathscr{T}}(U,T)$ for all $T\in\mathscr{T}$. On the other hand, we use that $\mathscr{T}\setminus\mathscr{R}=\mathscr{T}\cap\mathscr{T}_*$ and apply Problem 53 in conjunction with Lemma 25 to arrive at

$$\operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T}\setminus\mathscr{R})-2\operatorname{osc}_{\mathscr{T}_*}^2(U_*,\mathscr{T}\setminus\mathscr{R})\le C_3\,|||U-U_*|||_\Omega^2\le C_1C_3\mathscr{E}_{\mathscr{T}}^2(U,\mathscr{R}).$$

Adding these two estimates gives

$$\operatorname{osc}_{\mathscr{T}}^2(U,\mathscr{T})-2\operatorname{osc}_{\mathscr{T}_*}^2(U_*,\mathscr{T}_*)\le(1+C_1C_3)\mathscr{E}_{\mathscr{T}}^2(U,\mathscr{R}).$$

4 Returning to 1 we realize that

$$(1-2\mu)C_2\,\mathscr{E}_{\mathscr{T}}^2(U,\mathscr{T})\le\big(1+C_1(1+C_3)\big)\,\mathscr{E}_{\mathscr{T}}^2(U,\mathscr{R}),$$

which is (143) in disguise. In fact, recalling that $\theta_*^2=C_2/\big(1+C_1(1+C_3)\big)$ then $\theta^2=(1-2\mu)\theta_*^2<\theta_*^2$ as asserted. □

We are now ready to explore the cardinality of $\mathscr{M}_k$. To this end, we must relate AFEM with the approximation class $\mathbb{A}_s$. This might appear like an undoable task. However, the key to unravel this connection is given by Lemma 26.

Lemma 27 (Cardinality of $\mathscr{M}_k$). *Let Assumptions 2(a) and 2(b) be satisfied. If $(u,f,\boldsymbol{A}) \in \mathbb{A}_s$ then*

$$\#\mathscr{M}_k \lesssim |u,f,\boldsymbol{A}|_s^{1/s} \big(|||u-U_k|||_\Omega + \operatorname{osc}_k(U_k,\mathscr{T}_k)\big)^{-1/s} \quad \text{for all } k \geq 0. \tag{144}$$

Proof. We split the proof into three steps.
1 We set $\varepsilon^2 := \mu\Lambda_2^{-1}\big(|||u-U_k|||_\Omega^2 + \operatorname{osc}_k^2(U_k,\mathscr{T}_k)\big)$ with $\mu = \frac{1}{2}\big(1-\frac{\theta^2}{\theta_*^2}\big) > 0$ as in Lemma 26 and Λ_2 given Lemma 22. Since $(u,f,\boldsymbol{A}) \in \mathbb{A}_s$, in view of (138) there exists $\mathscr{T}_\varepsilon \in \mathbb{T}$ and $U_\varepsilon \in \mathbb{V}(\mathscr{T}_\varepsilon)$ such that

$$|||u-U_\varepsilon|||_\Omega^2 + \operatorname{osc}_\varepsilon^2(U_\varepsilon,\mathscr{T}_\varepsilon) \leq \varepsilon^2 \qquad \text{and} \qquad \#\mathscr{T}_\varepsilon - \#\mathscr{T}_0 \lesssim |u,f,\boldsymbol{A}|_s^{1/2}\,\varepsilon^{-1/s}.$$

Since $\mathscr{T}_\varepsilon$ may be totally unrelated to $\mathscr{T}_k$ we introduce the overlay

$$\mathscr{T}_* = \mathscr{T}_k \oplus \mathscr{T}_\varepsilon.$$

2 We claim that the total error over $\mathscr{T}_*$ reduces by a factor μ relative to that one over $\mathscr{T}_k$. In fact, since $\mathscr{T}_* \geq \mathscr{T}_\varepsilon$ and so $\mathbb{V}(\mathscr{T}_*) \supset \mathbb{V}(\mathscr{T}_\varepsilon)$, we use Lemma 22 to obtain

$$\begin{aligned} |||u-U_*|||_\Omega^2 + \operatorname{osc}_{\mathscr{T}_*}^2(U_*,\mathscr{T}_*) &\leq \Lambda_2\Big(|||u-U_\varepsilon|||_\Omega^2 + \operatorname{osc}_\varepsilon^2(U_\varepsilon,\mathscr{T}_\varepsilon)\Big) \\ &\leq \Lambda_2\varepsilon^2 = \mu\,\big(|||u-U_k|||_\Omega^2 + \operatorname{osc}_k^2(U_k,\mathscr{T}_k)\big). \end{aligned}$$

Upon applying Lemma 26 we conclude that the set $\mathscr{R} = \mathscr{R}_{\mathscr{T}_k \to \mathscr{T}_*}$ of refined elements satisfies a Dörfler marking (143) with parameter $\theta < \theta_*$.
3 According to Assumption 2(b) MARK selects a minimal set $\mathscr{M}_k$ satisfying this property. Therefore, we deduce

$$\#\mathscr{M}_k \leq \#\mathscr{R} \leq \#\mathscr{T}_* - \#\mathscr{T}_k \leq \#\mathscr{T}_\varepsilon - \#\mathscr{T}_0 \lesssim |u,f,\boldsymbol{A}|_s^{1/s}\,\varepsilon^{-1/s},$$

where we have employed Lemma 6 for the overlay. Now recalling the definition of ε we end up with the asserted estimate (144). □

Remark 32 (Blow-up). The constant hidden in (144) blows up as $\theta \uparrow \theta_*$ because $\mu \downarrow 0$.

9.3 Quasi-Optimal Convergence Rates

We are ready to prove the main result of this section, which combines Theorem 17 and Lemma 27.

Theorem 18 (Quasi-Optimality). *Let Assumption 2 be satisfied. If* $(u,f,\boldsymbol{A}) \in \mathbb{A}_s$ *then AFEM gives rise to a sequence* $\{\mathscr{T}_k,\mathbb{V}_k,U_k\}_{k=0}^{\infty}$ *such that*

$$|||u-U_k|||_\Omega + \operatorname{osc}_k(U_k,\mathscr{T}_k) \lesssim |u,f,\boldsymbol{A}|_s\,(\#\mathscr{T}_k - \#\mathscr{T}_0)^{-s} \quad \text{for all } k \geq 1.$$

Proof. [1] Since no confusion arises, we use the notation $\operatorname{osc}_j = \operatorname{osc}_j(U_j,\mathscr{T}_j)$ and $\mathscr{E}_j = \mathscr{E}_j(U_j,\mathscr{T}_j)$. In light of Assumption 2(c) and (144) we have

$$\#\mathscr{T}_k - \#\mathscr{T}_0 \lesssim \sum_{j=0}^{k-1} \#\mathscr{M}_j \lesssim |u,f,\boldsymbol{A}|_s^{1/s} \sum_{j=0}^{k-1} \big(\,|||u-U_j|||_\Omega^2 + \operatorname{osc}_j^2\big)^{-1/(2s)}.$$

[2] Let $\gamma>0$ be the scaling factor in the (contraction) Theorem 17. The lower bound (130b) along with $\operatorname{osc}_j \leq \mathscr{E}_j$ implies

$$|||u-U_j|||_\Omega^2 + \gamma\operatorname{osc}_j^2 \leq |||u-U_j|||_\Omega^2 + \gamma\mathscr{E}_j^2 \leq \Big(1+\frac{\gamma}{C_2}\Big)\big(\,|||u-U_j|||_\Omega^2 + \operatorname{osc}_j^2\big).$$

[3] Theorem 17 yields for $0 \leq j < k$

$$|||u-U_k|||_\Omega^2 + \gamma\mathscr{E}_k^2 \leq \alpha^{2(k-j)}\,\big(\,|||u-U_j|||_\Omega^2 + \gamma\mathscr{E}_j^2\big),$$

whence

$$\#\mathscr{T}_k - \#\mathscr{T}_0 \lesssim |u,f,\boldsymbol{A}|_s^{1/s}\,\big(\,|||u-U_k|||_\Omega^2 + \gamma\mathscr{E}_k^2\big)^{-1/(2s)} \sum_{j=0}^{k-1} \alpha^{(k-j)/s}.$$

Since $\sum_{j=0}^{k-1}\alpha^{(k-j)/s} = \sum_{j=1}^{k}\alpha^{j/s} < \sum_{j=1}^{\infty}\alpha^{j/s} < \infty$ because $\alpha<1$, the assertion follows immediately. □

Corollary 13 (Estimator Decay). *Let Assumption 2 be satisfied. If* $(u,f,\boldsymbol{A}) \in \mathbb{A}_s$ *then the estimator* $\mathscr{E}_k(U_k,\mathscr{T}_k)$ *satisfies*

$$\mathscr{E}_k(U_k,\mathscr{T}_k) \lesssim |u,f,\boldsymbol{A}|_s^{1/s}\,(\#\mathscr{T}_k - \#\mathscr{T}_0)^{-s}.$$

Proof. Use (136) and Theorem 18. □

Corollary 14 (W_p^2-Regularity). *Let* $d=2$*, the polynomial degree* $n=1$*,* $f \in L^2(\Omega)$*, and let* $\boldsymbol{A}$ *be piecewise constant over* $\mathscr{T}_0$*. If* $u \in W_p^2(\Omega;\mathscr{T}_0)$ *for* $p>1$*, then AFEM gives rise to a sequence* $\{\mathscr{T}_k,\mathbb{V}_k,U_k\}_{k=0}^{\infty}$ *satisfying* $\operatorname{osc}_k = \|h_k(f-P_0f)\|_{L^2(\Omega)}$ *and*

$$|||u-U_k|||_\Omega + \operatorname{osc}_k \lesssim \Big(\|D^2u\|_{L^p(\Omega;\mathscr{T}_0)} + \|f\|_{L^2(\Omega)}\Big)\,(\#\mathscr{T}_k - \#\mathscr{T}_0)^{-1/2}$$

for all $k \geq 1$.

Proof. Combine Corollary 11 with Theorem 18. □

Corollary 15 (W_p^2-Regularity). *Besides the assumptions of Corollary 14, let $\boldsymbol{A}$ be piecewise Lipschitz over the initial grid $\mathscr{T}_0$. Then AFEM gives rise to a sequence $\{\mathscr{T}_k, \mathbb{V}_k, U_k\}_{k=0}^{\infty}$ satisfying for all $k \geq 1$*

$$|||u - U_k|||_{\Omega} + \operatorname{osc}_k(U_k, \mathscr{T}_k) \lesssim \left(\|D^2 u\|_{L^p(\Omega;\mathscr{T}_0)} + \|f\|_{L^2(\Omega)} + \|\boldsymbol{A}\|_{W^1_\infty(\Omega;\mathscr{T}_0)} \right) (\#\mathscr{T}_k - \#\mathscr{T}_0)^{-1/2}.$$

Proof. Combine Corollary 12 with Theorem 18. □

So far we have assumed that the module SOLVE gives the *exact* Galerkin solution U_k and in doing this we have ignored the effects of numerical integration and inexact solution of the linear system; recall Remarks 9 and 10. The two issues above are important for AFEM to be fully practical. If one could control a posteriori the errors due to inexactness of SOLVE, then it would still be possible to prove a contraction property, as in Chap. 8, and examine the number of operations of AFEM in terms of $\#\mathscr{T}_k$ for a desired accuracy, following the steps of Sect. 9.2 and Sect. 9.3. We refer to Stevenson [69], who explores this endeavor for problem (127) with $\boldsymbol{A}$ piecewise constant.

9.4 Marking vs Optimality

We conclude with a brief discussion of processes that optimize more than one quantity at once and the critical role of marking, i. e., we consider adaptive algorithms that mark in each iteration for different error contributions separately. For instance, in earlier work on adaptivity error indicator and oscillation are treated independently [52, 53, 48]. Furthermore, when dealing with systems one is easily tempted to mark separately for the different components; compare for instance with [40]. It is worth observing that Binev et al. [13], Stevenson [69] and also Cascón et al. [21] avoided to use separate marking in their algorithms when proving optimal error decay. When dealing with the Poisson problem, oscillation becomes data oscillation and allows one to first approximate data sufficiently well and then reduce the energy error. This is done in different ways in [13] and [69]. However, for variable $\boldsymbol{A}$ the oscillation depends on the discrete solution, as discussed in Sect. 9.1, and the above splitting does not apply. Nonetheless marking solely for the estimator gives an optimal decay rate according to Sect. 9.3.

The design of adaptive algorithms that rely on separate marking is extremely delicate when aiming for optimal decay rates. To shed light on this issue we first present some numerical experiments based on separate marking, and next analyze the effect of separate marking in a simplified setting.

9.4.1 Separate Marking

The procedure ESTIMATE of Morin, Nochetto and Siebert, used in previous convergence proofs [52, 53, 48], calculates both the error and oscillation indicators $\{\mathscr{E}_k(U_k,T),\mathrm{osc}_k(U_k,T)\}_{T\in\mathscr{T}_k}$ (see Remarks 23 and 31), and the procedure MARK uses Dörfler marking for both the estimator and oscillation. More precisely, the routine MARK is of the form: *given parameters* $0<\theta_{\mathrm{est}},\theta_{\mathrm{osc}}<1$,

$$\text{mark any subset } \mathscr{M}_k\subset\mathscr{T}_k \text{ such that } \mathscr{E}_k(U_k,\mathscr{M}_k)\geq\theta_{\mathrm{est}}\mathscr{E}_k(U_k,\mathscr{T}_k); \tag{145a}$$

$$\text{if necessary enlarge } \mathscr{M}_k \text{ to satisfy } \mathrm{osc}_k(U_k,\mathscr{M}_k)\geq\theta_{\mathrm{osc}}\,\mathrm{osc}_k(U_k,\mathscr{T}_k). \tag{145b}$$

Since oscillation is generically of higher order than the estimator, the issue at stake is whether elements added by oscillation, even though immaterial relative to the error, could ruin the optimal cardinality observed in experiments. If $\mathscr{E}_k(U_k,\mathscr{T}_k)$ has large indicators in a small area, then Dörfler marking for the estimator (145a) could select a set $\mathscr{M}_k$ with a small number of elements relative to $\mathscr{T}_k$. However, if $\mathrm{osc}_k(U_k,\mathscr{T}_k)$ were globally distributed in $\mathscr{T}_k$, then separate marking would require additional marking of a large percentage of all elements to satisfy (145b); i.e., $\#\mathscr{M}_k$ could be large relative to $\#\mathscr{T}_k$.

To explore this idea computationally, we consider a simple modification of the Example of Sect. 8.4 with exact solution that we denote hereafter by u_R. Let u_S be the smooth solution of (141), which is of comparable magnitude with u_R, while the corresponding $f=-\mathrm{div}\boldsymbol{A}\nabla u_S$ exhibits an increasing amount of data oscillation away from the origin. Let $u=u_R+u_S$ be the modified exact solution and let f be the corresponding forcing function. Procedure MARK takes the usual value of $\theta_{\mathrm{est}}=0.5$ [32, 52, 53, 63], and procedure REFINE subdivides all elements in $\mathscr{M}_k$ by using two bisections.

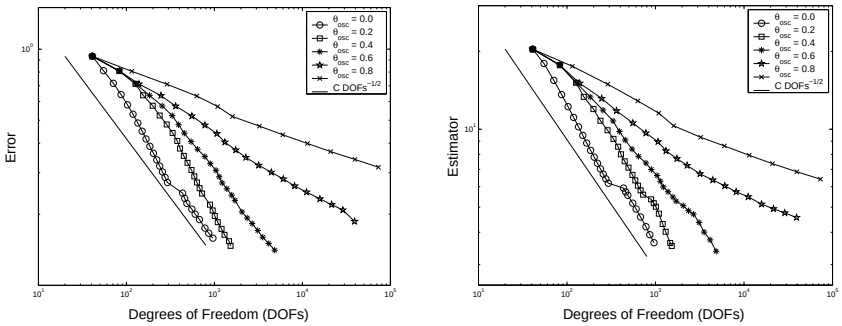

Fig. 20 Decay of the error (left) and the estimator (right) vs. degrees of freedom for $\theta_{\mathrm{est}}=0.5$ and values $\theta_{\mathrm{osc}}=0.0,0.2,0.4,0.6$, and 0.8. For values of $\theta_{\mathrm{osc}}\leq 0.4$ the rate of convergence is quasi-optimal, but for $\theta_{\mathrm{osc}}>0.4$ the curves flatten out, thereby indicating lack of optimality.

The behavior of separate marking for several values of θ_{osc} is depicted in Figure 20. We can visualize its sensitivity with respect to parameter θ_{osc}. For values of

$\theta_{\mathrm{osc}} \leq 0.4$ the rate of convergence appears to be quasi-optimal. However, beyond this threshold the curves for both the error and the estimator flatten out, thereby indicating a lack of optimality. The threshold value $\theta_{\mathrm{osc}} = 0.4$, even though consistent with practice, is tricky to find in general since it is problem-dependent. Therefore, marking by oscillation (145b) is questionable.

9.4.2 Analysis of Separate Marking

In order to gain mathematical insight on the key issues related to separate marking, we examine the adaptive approximation of two given functions in an idealized scenario. We show that separate marking, similar to (145), may lead to suboptimal meshes in general. However, a suitable choice of marking parameters may restore optimality. The numerical experiments of Sect. 9.4.1 confirm this theoretical insight in a realistic environment.

For the discussion, we assume that we have two functions u_i, $i = 1,2$, and have access to their local approximation error

$$e_{\mathscr{T}}(u_i;T) = |u_i - I_{\mathscr{T}} u_i|_{i;T} \qquad \forall T \in \mathscr{T}$$

and global error $e^2_{\mathscr{T}}(u_i) = \sum_{T\in\mathscr{T}} e^2_{\mathscr{T}}(u_i;T)$; hereafter $|\cdot|_i$ are unspecified norms, and $I_{\mathscr{T}}$ is a local interpolation operator over $\mathscr{T} \in \mathbb{T}$. We define the *total error* to be

$$e^2_{\mathscr{T}} := e^2_{\mathscr{T}}(u_1) + e^2_{\mathscr{T}}(u_2)$$

and are interested in its asymptotic decay. If $\mathscr{T} = \mathscr{T}_k$, then we denote $e_k = e_{\mathscr{T}_k}$.

To explore the use of (145), we examine the effect of separate marking for $e_k(u_i)$ on a sequence of meshes $\mathscr{T}^i$ for $i = 1,2$. We put ourselves in an idealized, but plausible, situation governed by the following three simplifying assumptions:

Independence: $\mathscr{T}^1_k$ and $\mathscr{T}^2_k$ are generated from $\mathscr{T}_0$ and are independent of each other; (146a)

Marking: Separate Dörfler marking with parameters $\theta_i \in (0,1)$ implies that $e_k(u_i) \approx \alpha_i^k$ on $\mathscr{T}^i_k$, with $\alpha_i \in (0,1)$; (146b)

Approximability: $e_k(u_i) \approx (\#\mathscr{T}^i_k - \#\mathscr{T}_0)^{-s_i}$, with $s_1 \leq s_2$ maximal. (146c)

We are interested in the decay of the total error e_k on the overlay $\mathscr{T}_k := \mathscr{T}^1_k \oplus \mathscr{T}^2_k$. This scenario is a simplification of the more realistic approximation of u_1 and u_2 with separate Dörfler marking on the same sequence of grids $\mathscr{T}_k$ but avoids the complicated interaction of the two marking procedures.

Lemma 28 (Separate Marking). *Let assumptions* (146) *be satisfied. Then the decay of the total error e_k on the overlay $\mathscr{T}_k = \mathscr{T}^1_k \oplus \mathscr{T}^2_k$ for separate marking is always suboptimal except when α_1 and α_2 satisfy*

$$\alpha_2 \leq \alpha_1 \leq \alpha_2^{s_1/s_2}.$$

Proof. [1] Assumption (146b) on the average reduction rate implies for the total error that

$$e_k \approx e_k(u_1) + e_k(u_2) \approx \max\{e_k(u_1), e_k(u_2)\} \approx \max\{\alpha_1^k, \alpha_2^k\}. \tag{147}$$

Combining (146b) and (146c) yields $\alpha_i^k \approx (\#\mathscr{T}_k^i - \#\mathscr{T}_0)^{-s_i}$, whence

$$\#\mathscr{T}_k^1 - \#\mathscr{T}_0 \approx \alpha_1^{-k/s_1} = \beta^k \alpha_2^{-k/s_2} \approx \beta^k (\#\mathscr{T}_k^2 - \#\mathscr{T}_0), \tag{148}$$

with $\beta = \alpha_1^{-1/s_1} \alpha_2^{1/s_2}$. In view of Lemma 6, this gives for the overlay $\mathscr{T}_k$

$$\#\mathscr{T}_k - \#\mathscr{T}_0 \approx \begin{cases} \#\mathscr{T}_k^1 - \#\mathscr{T}_0, & \beta \geq 1, \\ \#\mathscr{T}_k^2 - \#\mathscr{T}_0, & \beta < 1. \end{cases} \tag{149}$$

The optimal decay of total error e_k corresponds to $e_k \approx (\#\mathscr{T}_k - \#\mathscr{T}_0)^{-s_1}$ because $s_1 \leq s_2$. In analyzing the relation of e_k to the number of elements $\#\mathscr{T}_k$ in the overlay $\mathscr{T}_k$, we distinguish three cases and employ (147), (148), and (149).

[2] *Case*: $\alpha_1 < \alpha_2$. We note that $\alpha_1 < \alpha_2$ and $s_1 \leq s_2$ yields $\beta \geq 1$. We thus deduce

$$\begin{aligned} e_k &\approx \max\{\alpha_1^k, \alpha_2^k\} = \alpha_2^k = (\alpha_2/\alpha_1)^k \alpha_1^k \\ &\approx (\alpha_2/\alpha_1)^k (\#\mathscr{T}_k^1 - \#\mathscr{T}_0)^{-s_1} \approx (\alpha_2/\alpha_1)^k (\#\mathscr{T}_k - \#\mathscr{T}_0)^{-s_1}. \end{aligned}$$

Since $\alpha_2/\alpha_1 > 1$, the approximation of e_k on $\mathscr{T}_k$ is suboptimal.

[3] *Case*: $\alpha_1 \geq \alpha_2$ and $\beta < 1$. We obtain

$$\begin{aligned} e_k &\approx \max\{\alpha_1^k, \alpha_2^k\} = \alpha_1^k \approx (\#\mathscr{T}_k^1 - \#\mathscr{T}_0)^{-s_1} \\ &\approx \beta^{-ks_1} (\#\mathscr{T}_k^2 - \#\mathscr{T}_0)^{-s_1} \approx \beta^{-ks_1} (\#\mathscr{T}_k - \#\mathscr{T}_0)^{-s_1}, \end{aligned}$$

whence the approximation of the total error on $\mathscr{T}_k$ is again suboptimal.

[4] *Case*: $\alpha_1 \geq \alpha_2$ and $\beta \geq 1$. We infer that

$$e_k \approx \max\{\alpha_1^k, \alpha_2^k\} = \alpha_1^k \approx (\#\mathscr{T}_k^1 - \#\mathscr{T}_0)^{-s_1} \approx (\#\mathscr{T}_k - \#\mathscr{T}_0)^{-s_1}$$

and that $\mathscr{T}_k$ exhibits optimal cardinality. This exceptional case corresponds to the assertion and concludes the proof. □

We learn from Lemma 28 that separate marking requires a critical choice of parameters θ_i to retain optimal error decay with respect to the total error e_k. In light of Lemma 28, we could identify the AFEM estimator $\mathscr{E}_k$ with the error $e_k(u_1)$ and the AFEM oscillation osc_k with the error $e_k(u_2)$. We observe that $\operatorname{osc}_k \leq \mathscr{E}_k$ combined with (146b) implies that $\alpha_2 \leq \alpha_1$ and that osc_k is generically of higher order than $\mathscr{E}_k$, thereby yielding $s_1 < s_2$.

We wonder whether or not the optimality condition $\alpha_1 \leq \alpha_2^{s_1/s_2}$ is valid. Note that $\alpha_2^{s_1/s_2}$ increases as the gap between s_1 and s_2 increases. Since the oscillation reduction estimate of [52] reveals that α_2 increases as θ_{osc} decreases, we see that

separate marking may be optimal for a wide range of marking parameters $\theta_{\mathrm{est}}, \theta_{\mathrm{osc}}$; this is confirmed by the numerical experiments in Sect. 9.4.1 even though it is unclear whether $\mathscr{E}_k$ and osc_k satisfy (146). However, choosing marking parameters $\theta_{\mathrm{est}}, \theta_{\mathrm{osc}}$ is rather tricky in practice because neither the explicit dependence of average reduction rates α_1, α_2 on $\theta_{\mathrm{est}}, \theta_{\mathrm{osc}}$ nor the optimal exponents s_1, s_2 are known. In contrast to [52, 53, 48], the standard AFEM of Chap. 8 marks solely according to the estimator $\mathscr{E}_k(U_k, \mathscr{T}_k)$ and thus avoids separate marking.

9.5 Problems

Problem 54. Show that $(v, f, \boldsymbol{A}) \in \mathbb{A}_s$ if and only there exists a constant $\Lambda > 0$ such that for all $\varepsilon > 0$ there exist $\mathscr{T}_\varepsilon \geq \mathscr{T}_0$ conforming and $V_\varepsilon \in \mathbb{V}(\mathscr{T}_\varepsilon)$ such that

$$|||v - V_\varepsilon|||_\Omega^2 + \mathrm{osc}_{\mathscr{T}_\varepsilon}^2 \leq \varepsilon^2 \qquad \text{and} \qquad \#\mathscr{T}_\varepsilon - \#\mathscr{T}_0 \leq \Lambda^{1/s}\, \varepsilon^{-1/s};$$

in this case $|v, f, \boldsymbol{A}|_s \leq \Lambda$. Hint: Let $\mathscr{T}_\varepsilon$ be minimal for $|||v - V_\varepsilon|||_\Omega^2 + \mathrm{osc}_{\mathscr{T}_\varepsilon}^2 \leq \varepsilon^2$. This means that for all $\mathscr{T} \in \mathbb{T}$ such that $\#\mathscr{T} = \#\mathscr{T}_\varepsilon - 1$ we have $|||v - V_\varepsilon|||_\Omega^2 + \mathrm{osc}_{\mathscr{T}_\varepsilon}^2 > \varepsilon$.

Problem 55. Prove Lemma 25: if $\mathscr{T}, \mathscr{T}_* \in \mathbb{T}$ satisfy $\mathscr{T}_* \geq \mathscr{T}$, $\mathscr{R} := \mathscr{R}_{\mathscr{T} \to \mathscr{T}_*}$ is the refined set to go from $\mathscr{T}$ to $\mathscr{T}_*$, and $U \in \mathbb{V}$, $U_* \in \mathbb{V}_*$ are the corresponding Galerkin solutions, then

$$|||U_* - U|||_\Omega^2 \leq C_1\, \mathscr{E}_{\mathscr{T}}^2(U, \mathscr{R}).$$

To this end, write the equation fulfilled by $U_* - U \in \mathbb{V}_*$ and use as a test function the local quasi-interpolant $I_{\mathscr{T}}(U_* - U)$ of $U_* - U$ introduced in Proposition 3(b) and compare with Remark 20.

Problem 56. Trace the dependence as $\theta \to \theta_*$ and $s \to 0$ in Lemma 27 and Theorem 18.

Problem 57. Let $d = 2$ and $n = 1$. Let f be piecewise W_1^1 over the initial mesh $\mathscr{T}_0$, namely $f \in W_1^1(\Omega; \mathscr{T}_0)$. Show that

$$\inf_{\mathscr{T} \in \mathbb{T}_N} \|h_{\mathscr{T}}(f - P_0 f)\|_{L^2(\Omega)} \lesssim \|f\|_{W_1^1(\Omega; \mathscr{T}_0)} N^{-1}.$$

This shows the same decay rate of data oscillation as in Remark 31 but with weaker regularity.

References

1. Ainsworth, M., Oden, J.T.: A posteriori error estimation in finite element analysis. Wiley (2000)
2. Arnold, D.N., Mukherjee, A., Pouly, L.: Locally adapted tetrahedral meshes using bisection. SIAM J. Sci. Comput. **22**(2), 431–448 (2000)
3. Atalay, F.B., Mount, D.M.: The cost of compatible refinement of simplex decomposition trees. In: Proc. International Meshing Roundtable 2006 (IMR 2006), pp. 57–69 (2006). Birmingham, AL
4. Babuška, I., Kellogg, R.B., Pitkäranta, J.: Direct and inverse error estimates for finite elements with mesh refinements. Numer. Math. **33**(4), 447–471 (1979)
5. Babuška, I., Rheinboldt, W.: Error estimates for adaptive finite element computations. SIAM J. Numer. Anal. **15**, 736–754 (1978)
6. Babuška, I., Strouboulis, T.: The finite element method and its reliability. Numerical Mathematics and Scientific Computation. The Clarendon Press Oxford University Press, New York (2001)
7. Babuška, I., Vogelius, M.: Feedback and adaptive finite element solution of one-dimensional boundary value problems. Numer. Math. **44**, 75–102 (1984)
8. Babuška, I.: Error-bounds for finite element method. Numer. Math. **16**, 322–333 (1971)
9. Babuška, I., Aziz, A.K.: Survey lectures on the mathematical foundations of the finite element method. with the collaboration of G. Fix and R. B. Kellogg. Math. Found. Finite Elem. Method Appl. Part. Differ. Equations, Sympos. Univ. Maryland, Baltimore 1972, 1-359 (1972). (1972)
10. Bänsch, E.: Local mesh refinement in 2 and 3 dimensions. IMPACT Comput. Sci. Engrg. **3**, 181–191 (1991)
11. Bebendorf, M.: A note on the Poincaré inequality for convex domains. Z. Anal. Anwendungen **22**(4), 751–756 (2003)
12. Beck, R., Hiptmair, R., Hoppe, R.H., Wohlmuth, B.: Residual based a posteriori error estimators for eddy current computation. Math. Model. Numer. Anal. **34**(1), 159–182 (2000)
13. Binev, P., Dahmen, W., DeVore, R.: Adaptive finite element methods with convergence rates. Numer. Math. **97**, 219–268 (2004)
14. Binev, P., Dahmen, W., DeVore, R., Petrushev, P.: Approximation classes for adaptive methods. Serdica Math. J. **28**(4), 391–416 (2002). Dedicated to the memory of Vassil Popov on the occasion of his 60th birthday
15. Braess, D.: Finite Elements. Theory, fast solvers, and applications in solid mechanics, 2nd edition edn. Cambridge University Press (2001)
16. Brenner, S., Scott, R.: The Mathematical Theory of Finite Element Methods. Springer Texts in Applied Mathematics 15 (2008)
17. Brezzi, F.: On the existence, uniqueness and approximation of saddle-point problems arising from lagrange multipliers. R.A.I.R.O. Anal. Numer. **R2**, T 129–151 (1974)
18. Brezzi, F., Fortin, M.: Mixed and Hybrid Finite Element Methods. Springer Series in Computational Mathematics 15 (1991)
19. Carroll, R., Duff, G., Friberg, J., Gobert, J., Grisvard, P., Nečas, J., Seeley, R.: Equations aux dérivées partielles. No. 19 in Seminaire de mathematiques superieures. Les Presses de l'Université de Montréal (1966)
20. Carstensen, C., Funken, S.A.: Fully reliable localized error control in the FEM. SIAM J. Sci. Comput. **21**(4), 1465–1484 (electronic) (1999/00)
21. Cascón, J.M., Kreuzer, C., Nochetto, R.H., Siebert, K.G.: Quasi-optimal convergence rate for an adaptive finite element method. Preprint 009/2007, Universität Augsburg (2007)
22. Cea, J.: Approximation variationnelle des problèmes aux limites. Ann. Inst. Fourier **14**(2), 345–444 (1964)
23. Chen, L., Nochetto, R.H., Xu, J.: Adaptive multilevel methods on graded bisection grids. to appear (2009)
24. Chen, Z., Feng, J.: An adaptive finite element algorithm with reliable and efficient error control for linear parabolic problems. Math. Comp. **73**, 1167–1042 (2006)

25. Ciarlet, P.G.: The finite element methods for elliptic problems. Classics in Applied Mathematics, 40, SIAM (2002)
26. Clément, P.: Approximation by finite element functions using local regularization. R.A.I.R.O. **9**, 77–84 (1975)
27. Dahlke, S., DeVore, R.A.: Besov regularity for elliptic boundary value problems. Commun. Partial Differ. Equations **22**(1-2), 1–16 (1997)
28. DeVore, R.A.: Nonlinear approximation. In: A. Iserles (ed.) Acta Numerica, vol. 7, pp. 51–150. Cambridge University Press (1998)
29. DeVore, R.A., Lorentz, G.G.: Constructive approximation, *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*, vol. 303. Springer-Verlag, Berlin (1993)
30. DeVore, R.A., Popov, V.A.: Interpolation of Besov spaces. Trans. Amer. Math. Soc. **305**(1), 397–414 (1988)
31. Diening, L., Kreuzer, C.: Convergence of an adaptive finite element method for the p-Laplacian equation. SIAM J. Numer. Anal. **46**(2), 614–638 (2008)
32. Dörfler, W.: A convergent adaptive algorithm for Poisson's equation. SIAM J. Numer. Anal. **33**, 1106–1124 (1996)
33. Dupont, T., Scott, R.: Polynomial approximation of functions in Sobolev spaces. Math. Comp. **34**(150), 441–463 (1980)
34. Eriksson, K., Johnson, C.: Adaptive finite element methods for parabolic problems I: A linear model problem. SIAM J. Numer. Anal. **28**, 43–77 (1991)
35. Evans, L.C.: Partial Differential Equations. Graduate Studies in Mathematics. 19. Providence, AMS (1998)
36. Galdi, G.P.: An introduction to the mathematical theory of the Navier-Stokes equations. Vol. 1: Linearized steady problems. Springer Tracts in Natural Philosophy, 38 (1994)
37. Gaspoz, F., Morin, P.: Approximation classes for adaptive higher order finite element approximation. (in preparation) (2009)
38. Gilbarg, D., Trudinger, N.S.: Elliptic partial differential equations of second order. Classics in Mathematics, Springer (2001)
39. Grisvard, P.: Elliptic problems in nonsmooth domains, *Monographs and Studies in Mathematics*, vol. 24. Pitman (Advanced Publishing Program), Boston, MA (1985)
40. Hintermüller, M., Hoppe, R.H., Iliash, Y., Kieweg, M.: An a posteriori error analysis of adaptive finite element methods for distributed elliptic control problems with control constraints. ESAIM, Control Optim. Calc. Var. **14**(3), 540–560 (2008)
41. Jarausch, H.: On an adaptive grid refining technique for finite element approximations. SIAM J. Sci. Stat. Comput. **7**, 1105–1120 (1986)
42. Kato, T.: Estimation of iterated matrices, with application to the von Neumann condition. Numer. Math. **2**, 22–29 (1960)
43. Kellogg, R.B.: On the Poisson equation with intersecting interfaces. Applicable Anal. **4**, 101–129 (1974/75). Collection of articles dedicated to Nikolai Ivanovich Muskhelishvili
44. Kossaczký, I.: A recursive approach to local mesh refinement in two and three dimensions. J. Comput. Appl. Math. **55**, 275–288 (1994)
45. Lax, P., Milgram, A.: Parabolic equations. Ann. Math. Stud. **33**, 167–190 (1954)
46. Liu, A., Joe, B.: Quality local refinement of tetrahedral meshes based on bisection. SIAM J. Sci. Comput. **16**, 1269–1291 (1995)
47. Maubach, J.M.: Local bisection refinement for n-simplicial grids generated by reflection. SIAM J. Sci. Comput. **16**, 210–227 (1995)
48. Mekchay, K., Nochetto, R.H.: Convergence of adaptive finite element methods for general second order linear elliptic PDEs. SIAM J. Numer. Anal. **43**(5), 1803–1827 (2005)
49. Mitchell, W.F.: Unified multilevel adaptive finite element methods for elliptic problems. Ph.D. thesis, Department of Computer Science, University of Illinois, Urbana (1988)
50. Mitchell, W.F.: A comparison of adaptive refinement techniques for elliptic problems. ACM Trans. Math. Softw. **15**, 326–347 (1989)
51. Monk, P.: Finite element methods for Maxwell's equations. Numerical Mathematics and Scientific Computation. Oxford University Press. xiv, 450 p. (2003)

52. Morin, P., Nochetto, R.H., Siebert, K.G.: Data oscillation and convergence of adaptive FEM. SIAM J. Numer. Anal. **38**, 466–488 (2000)
53. Morin, P., Nochetto, R.H., Siebert, K.G.: Convergence of adaptive finite element methods. SIAM Review **44**, 631–658 (2002)
54. Morin, P., Nochetto, R.H., Siebert, K.G.: Local problems on stars: A posteriori error estimators, convergence, and performance. Math. Comp. **72**, 1067–1097 (2003)
55. Morin, P., Siebert, K.G., Veeser, A.: A basic convergence result for conforming adaptive finite elements. Math. Models Methods Appl. 18 (2008) **18**, 707–737 (2008)
56. Necas, J.: Sur une méthode pour resoudre les équations aux dérivées partielles du type elliptique, voisine de la variationnelle. Ann. Sc. Norm. Super. Pisa, Sci. Fis. Mat., III. Ser. **16**, 305–326 (1962)
57. Nochetto, R.H., Paolini, M., Verdi, C.: An adaptive finite element method for two-phase Stefan problems in two space dimensions. I. Stability and error estimates. Math. Comp. **57**(195), 73–108, S1–S11 (1991)
58. Oswald, P.: On function spaces related to finite element approximation theory. Z. Anal. Anwendungen **9**(1), 43–64 (1990)
59. Otto, F.: On the Babuška–Brezzi condition for the Taylor–Hood element. Diploma thesis Universität Bonn (1990). In German
60. Payne, L.E., Weinberger, H.F.: An optimal Poincaré-inequality for convex domains. Archive Rat. Mech. Anal. **5**, 286–292 (1960)
61. Rivara, M.C.: Mesh refinement processes based on the generalized bisection of simplices. SIAM J. Numer. Anal. **21**(3), 604–613 (1984)
62. Sacchi, R., Veeser, A.: Locally efficient and reliable a posteriori error estimators for Dirichlet problems. Math. Models Methods Appl. Sci. **16**(3), 319–346 (2006)
63. Schmidt, A., Siebert, K.G.: Design of adaptive finite element software. The finite element toolbox ALBERTA. Lecture Notes in Computational Science and Engineering 42, Springer (2005)
64. Schöberl, J.: A posteriori error estimates for Maxwell equations. Math. Comp. **77**(262), 633–649 (2008)
65. Scott, L.R., Zhang, S.: Finite element interpolation of nonsmooth functions satisfying boundary conditions. Mathematics of Computation **54**(190), 483–493 (1990)
66. Sewell, E.G.: Automatic generation of triangulations for piecewise polynomial approximation. Ph.D. dissertation, Purdue Univ., West Lafayette, Ind., 1972
67. Siebert, K.G.: A convergence proof for adaptive finite elements without lower bound (2009). Preprint Universität Duisburg-Essen and Universität Freiburg No. 1/2009
68. Siebert, K.G., Veeser, A.: A unilaterally constrained quadratic minimization with adaptive finite elements. SIAM J. Optim. **18**(1), 260–289 (2007)
69. Stevenson, R.: Optimality of a standard adaptive finite element method. Found. Comput. Math. **7**(2), 245–269 (2007)
70. Stevenson, R.: The completion of locally refined simplicial partitions created by bisection. Math. Comput. **77**(261), 227–241 (2008)
71. Storoženko, È.A., Oswald, P.: Jackson's theorem in the spaces $L^p(\mathbf{R}^k)$, $0 < p < 1$. Sibirsk. Mat. Ž. **19**(4), 888–901, 956 (1978)
72. Traxler, C.T.: An algorithm for adaptive mesh refinement in n dimensions. Computing **59**(2), 115–137 (1997)
73. Veeser, A.: Convergent adaptive finite elements for the nonlinear Laplacian. Numer. Math. **92**(4), 743–770 (2002)
74. Veeser, A., Verfürth, R.: Explicit upper bounds for dual norms of residuals. SIAM J. Numer. Anal. (to appear)
75. Verfürth, R.: A posteriori error estimators for the Stokes equations. Numer. Math. **55**, 309–325 (1989)
76. Verfürth, R.: A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques. Adv. Numer. Math. John Wiley, Chichester, UK (1996)
77. Wu, H., Chen, Z.: Uniform convergence of multigrid V-cycle on adaptively refined finite element meshes for second order elliptic problems. Sci. China Ser. A **49**(10), 1405–1429 (2006)

78. Xu, J., Zikatanov, L.: Some observations on Babuška and Brezzi theories. Numer. Math. **94**(1), 195–202 (2003)