

# Adaptive Finite Element Methods

Andrea Bonito

*Department of Mathematics, Texas A&M University,  
College Station, Texas 77843, USA  
E-mail: bonito@tamu.edu*

Claudio Canuto

*Dipartimento di Scienze Matematiche, Politecnico di Torino,  
Corso Duca degli Abruzzi 24, 10129 Torino, Italy  
E-mail: claudio.canuto@polito.it*

Ricardo H. Nochetto

*Department of Mathematics and Institute for Physical Science and Technology,  
University of Maryland, College Park, MD 20742, USA  
E-mail: rhn@umd.edu*

Andreas Veeger

*Dipartimento di Matematica, Università degli Studi di Milano,  
Via Saldini 50, 20133 Milano, Italy  
E-mail: andreas.veeger@unimi.it*

This is a survey on the theory of adaptive finite element methods (AFEMs), which are fundamental in modern computational science and engineering but whose mathematical assessment is a formidable challenge. We present a self-contained and up-to-date discussion of AFEMs for linear second order elliptic PDEs and dimension  $d > 1$ , with emphasis on foundational issues. After a brief review of functional analysis and basic finite element theory, including piecewise polynomial approximation in graded meshes, we present the core material for coercive problems. We start with a novel a posteriori error analysis applicable to rough data, which delivers estimators fully equivalent to the solution error. They are used in the design and study of three AFEMs depending on the structure of data. We prove linear convergence of these algorithms and rate-optimality provided the solution and data belong to suitable approximation classes. We also address the relation between approximation and regularity classes. We finally extend this theory to discontinuous Galerkin methods as prototypes of non-conforming AFEMs and beyond coercive problems to inf-sup stable AFEMs.

## CONTENTS

1	Introduction: Overview of AFEMs	2
2	Linear Elliptic Boundary Value Problems: Basic Theory	10
3	A Priori Approximation Theory	23
4	A Posteriori Error Analysis	54
5	Convergence of AFEM for Coercive Problems	122
6	Convergence Rates of AFEM for Coercive Problems	161
7	Data Approximation	190
8	Mesh Refinement: The Bisection Method	216
9	Discontinuous Galerkin Methods	232
10	AFEMs for Inf-Sup Stable Problems	263
	Index	310
	References	315

### 1 Introduction: Overview of AFEMs

This is a survey on the theory of adaptive finite element methods (AFEMs), which are fundamental in modern computational science and engineering. We present a self-contained and up-to-date discussion of AFEMs for linear second order elliptic PDE in dimension  $d > 1$ , with emphasis on foundational issues rather than applications of AFEMs. This paper builds on and expands the older surveys [Nochetto, Siebert and Veerer 2009, Nochetto and Veerer 2012]. In fact, we decided to incorporate several new aspects to the theory described below.

The paper develops the theory of AFEMs gradually and is meant to be accessible to advanced students and researchers interested in learning the fundamental aspects of adaptivity and why AFEMs outperform classical FEMs. We quantify the superior performance of AFEMs with precise mathematical statements rather than simulations. We present very few numerical experiments to illustrate some key (and new) algorithmic ideas and methods, but the paper is otherwise a tour in the numerical analysis of adaptive approximation of linear elliptic PDEs.

By design this paper goes deep into some foundational aspects of AFEMs theory, provides full discussions and proofs, as well as pointers to the main literature. We consider the following model problem on a polyhedral domain  $\Omega \subset \mathbb{R}^d$  with  $d \geq 2$

$$L[u] := -\operatorname{div}(A\nabla u) + cu = f \quad (1.1)$$

with general variable coefficients  $(A, c)$ , forcing  $f \in H^{-1}(\Omega)$  and homogeneous Dirichlet boundary conditions  $u = 0$  on  $\partial\Omega$  mostly, but not exclusively. If  $\mathbb{V} := H_0^1(\Omega)$  and  $\mathcal{B} : \mathbb{V} \times \mathbb{V} \rightarrow \mathbb{R}$  is the bilinear form associated with (1.1), the weak form reads:

$$u \in \mathbb{V} : \quad \mathcal{B}[u, v] = \langle f, v \rangle \quad \forall v \in \mathbb{V}. \quad (1.2)$$

Given a conforming and shape regular partition  $\mathcal{T}$  of  $\Omega$ , created by successive refinement of a coarse mesh  $\mathcal{T}_0$ , let  $\mathbb{V}_{\mathcal{T}}$  denote the space of continuous piecewise

polynomial functions of degree  $n \geq 1$  over  $\mathcal{T}$  vanishing on  $\partial\Omega$ . The Galerkin approximation  $u_{\mathcal{T}}$  of  $u$  solves

$$u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}} : \quad \mathcal{B}[u_{\mathcal{T}}, v] = \langle f, v \rangle \quad \forall v \in \mathbb{V}_{\mathcal{T}}. \quad (1.3)$$

This is a conforming approximation because  $\mathbb{V}_{\mathcal{T}} \subset \mathbb{V}$ . The aim of this paper is to:

- Design and analyze practical ways to estimate the error  $|u - u_{\mathcal{T}}|_{H_0^1(\Omega)} := \|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)}$  in terms of so-called a posteriori error estimators, which are computable quantities depending on the discrete solution  $u_{\mathcal{T}}$  and data  $\mathcal{D} = (\mathbf{A}, c, f)$ .
- Design adaptive algorithms that equidistribute the local errors  $\|\nabla(u - u_{\mathcal{T}})\|_{L^2(T)}$  for all elements  $T \in \mathcal{T}$ , thereby optimizing the computational effort; this is a key step that makes complex 3d situations accessible computationally.
- Show that this strategy delivers a performance comparable with the best possible in terms of degrees of freedom, which is a measure of computational complexity. This is a delicate matter because it entails dealing with approximation classes and their relation to regularity classes in terms of Besov and Lipschitz spaces.
- Present and analyze the bisection method for mesh refinement, one of the most versatile techniques for local mesh refinement that guarantees shape regularity and optimal complexity; the latter is instrumental for the previous point. Our study includes conforming meshes as well as certain non-conforming meshes.
- Extend the theory to a range of important problems that fail to be conforming or coercive. The first class is discontinuous Galerkin methods and the second one is inf-sup stable FEMs. The former is a notorious example of non-conforming approximation, whereas the second is non-coercive.

In achieving these goals we provide several new ideas and methods. We also refer to the pertinent literature but we do not give a full list of references nor get into comparisons of various approaches. It is not our intention to be comprehensive but rather to cover basic aspects of adaptivity in depth at the expense of important topics we do not touch upon. Some of them are:

- Adaptive eigenvalue approximation
- Goal oriented error analysis
- Non-conforming discretizations (except for discontinuous Galerkin)
- Coarsening or aggregation
- Anisotropic refinements
- $hp$ -adaptivity
- Tree approximation
- Other PDEs: convection-diffusion equations, nonlinear and evolution equations.

We devote the rest of this introduction to provide a roadmap to the rest of the paper. In doing so, we introduce notation that will be used later and present some topics in their most primitive form to render an early idea about how they fit and interrelate.

*A posteriori error analysis.* We refer to the books [Ainsworth and Oden 2000, Verfürth 2013] for the classical theory. However, in contrast to most of the existing literature, the current theory deals with forcing  $f \in H^{-1}(\Omega)$ . This allows for rough data useful in applications, such as line Dirac masses, but also encompasses a new approach to error estimation that leads to error-dominated estimator and oscillation and prevents error overestimation; this extends [Kreuzer and Veeser 2021] to (1.1) and polynomial degree  $n \geq 1$ . The new twist is the construction of a projection operator  $P_{\mathcal{T}} : H^{-1}(\Omega) \rightarrow \mathbb{F}_{\mathcal{T}}$  into a space of piecewise polynomials in  $\mathcal{T}$  and on its skeleton  $\mathcal{F}$ , namely the set of all internal faces. Such an operator happens to be locally stable on stars (or patches)  $\omega_z$  of  $\mathcal{T}$  for all vertices  $z \in \mathcal{V}$  of  $\mathcal{T}$ :

$$\|P_{\mathcal{T}}\ell\|_{H^{-1}(\omega_z)} \leq C_{\text{IStb}}\|\ell\|_{H^{-1}(\omega_z)} \quad \forall \ell \in H^{-1}(\omega_z). \quad (1.4)$$

An important property of  $P_{\mathcal{T}}$  and its range  $\mathbb{F}_{\mathcal{T}}$  is that for piecewise polynomial coefficients  $(A, c)$ , or in short discrete coefficients,  $P_{\mathcal{T}}$  is invariant in the subspace  $L[\mathbb{V}_{\mathcal{T}}]$  of  $H^{-1}(\Omega)$  or equivalently

$$P_{\mathcal{T}}(L[v]) = L[v] \in \mathbb{F}_{\mathcal{T}} \quad \forall v \in \mathbb{V}_{\mathcal{T}}. \quad (1.5)$$

It is worth realizing that  $L[v]$  is made of two distinct parts. The first one is absolutely continuous relative to the Lebesgue measure, namely  $-\text{div}(A\nabla v) + cv$  in every element  $T \in \mathcal{T}$ . The second part is singular and supported in the skeleton  $\mathcal{F}$ , namely  $[[A\nabla v]] \cdot \mathbf{n}|_F \delta_F$  for every face  $F \in \mathcal{F}$ , where  $[[\cdot]]$  is the jump across  $F$ ,  $\mathbf{n}$  is a unit normal to  $F$ , and  $\delta_F$  is the Dirac mass on  $F$ .

These two properties of  $P_{\mathcal{T}}$  have the following crucial consequences. Let

$$R_{\mathcal{T}} := L[u - u_{\mathcal{T}}] = f - L[u_{\mathcal{T}}] \in H^{-1}(\Omega) \quad (1.6)$$

be the residual of the Galerkin approximation of (1.2). Using (1.5) yields

$$R_{\mathcal{T}} - P_{\mathcal{T}}R_{\mathcal{T}} = f - P_{\mathcal{T}}f.$$

This shows that  $R_{\mathcal{T}}$  decomposes into a discrete, thus finite dimensional and computable, PDE part  $P_{\mathcal{T}}R_{\mathcal{T}} = P_{\mathcal{T}}f - L[u_{\mathcal{T}}]$  and an infinite dimensional component  $f - P_{\mathcal{T}}f$ , the so-called data oscillation that depends on  $f$  and can only be evaluated with additional knowledge of  $f$ .

The nonlocal  $H^{-1}$ -norm of  $R_{\mathcal{T}}$  splits into local contributions on stars, whence <sup>1</sup>

$$|u - u_{\mathcal{T}}|_{H_0^1(\Omega)}^2 \approx \|R_{\mathcal{T}}\|_{H^{-1}(\Omega)}^2 \approx \sum_{z \in \mathcal{V}} \|R_{\mathcal{T}}\|_{H^{-1}(\omega_z)}^2.$$

The discrete nature of  $P_{\mathcal{T}}R_{\mathcal{T}}$  allows us to derive a computable  $L^2$ -weighted PDE estimator  $\eta_{\mathcal{T}}(u_{\mathcal{T}}, z)$  equivalent to  $\|R_{\mathcal{T}}\|_{H^{-1}(\omega_z)}$ , which together with the remaining

<sup>1</sup> Throughout this work,  $A \lesssim B$  signifies  $A \leq CB$  with a constant  $C$  independent of the discretization parameters, and  $A \approx B$  stands for  $A \lesssim B$  and  $B \lesssim A$ .

data oscillation  $\text{osc}_{\mathcal{T}}(f, z)_{-1} := \|f - P_{\mathcal{T}}f\|_{H^{-1}(\omega_z)}$  gives the upper bound

$$|u - u_{\mathcal{T}}|_{H_0^1(\Omega)}^2 \lesssim \sum_{z \in \mathcal{V}} \left( \eta_{\mathcal{T}}(u_{\mathcal{T}}, z)^2 + \text{osc}_{\mathcal{T}}(f, z)_{-1}^2 \right).$$

It turns out that this estimate is sharp or, in other words, that there is no overestimation of the error. To see this important and unique property of these new estimators, we invoke (1.4) to write the local lower bounds

$$\begin{aligned} \eta_{\mathcal{T}}(u_{\mathcal{T}}, z) &\approx \|P_{\mathcal{T}}R_{\mathcal{T}}\|_{H^{-1}(\omega_z)} \leq C_{\text{Istb}} \|R_{\mathcal{T}}\|_{H^{-1}(\omega_z)}, \\ \text{osc}_{\mathcal{T}}(f, z)_{-1} &= \|R_{\mathcal{T}} - P_{\mathcal{T}}R_{\mathcal{T}}\|_{H^{-1}(\omega_z)} \leq (1 + C_{\text{Istb}}) \|R_{\mathcal{T}}\|_{H^{-1}(\omega_z)}. \end{aligned}$$

Section 4 constructs the operator  $P_{\mathcal{T}}$ , and derives several important properties such as its local quasi-best approximation and the above error-dominated a posteriori bounds. The former guarantees the inequality for the local  $L^2$ -projection  $\Pi_{\mathcal{T}}$

$$\|f - P_{\mathcal{T}}f\|_{H^{-1}(\omega_z)} \lesssim \|f - \Pi_{\mathcal{T}}f\|_{H^{-1}(\omega_z)} \lesssim \|h(f - \Pi_{\mathcal{T}}f)\|_{L^2(\omega_z)},$$

which is the typical form of data oscillation provided  $f \in L^2(\Omega)$ . However, this  $L^2$ -weighted oscillation is not bounded above by the error and is thus responsible for potential overestimation. Section 4 proves further properties of  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$  such as its reduction upon refinement and its localized discrete upper bound, as well as quasi-monotonicity of  $\text{osc}_{\mathcal{T}}(f)_{-1}$  upon refinement. These properties, known for the standard  $L^2$ -weighted estimator and oscillation, are thus retained by the new construction.

Section 4 also deals with the alternative error estimators that result from solving local problems, using hierarchy, or imposing flux equilibration. We show that all of them lead, essentially, to estimators equivalent to  $\|R_{\mathcal{T}}\|_{H^{-1}(\omega_z)}$ . Moreover, we present an optimal framework to deal with non-homogeneous Dirichlet boundary conditions as well as with Robin and Neumann boundary conditions.

*Linear convergence of AFEMs.* Local a posteriori error indicators are usually employed to mark elements (or set of elements) with largest indicators for refinement. We are concerned with the most popular *Dörfler marking* (or bulk chasing): given a parameter  $\theta \in (0, 1]$ , select a set  $\mathcal{M} \subset \mathcal{T}$  such that

$$\eta_{\mathcal{T}}(u_{\mathcal{T}}, \mathcal{M}) \geq \theta \eta_{\mathcal{T}}(u_{\mathcal{T}}); \quad (1.7)$$

hereafter we define  $\eta_{\mathcal{T}}(u_{\mathcal{T}}, \mathcal{M})^2 := \sum_{T \in \mathcal{M}} \eta_{\mathcal{T}}(u_{\mathcal{T}}, T)^2$  where  $\eta_{\mathcal{T}}(u_{\mathcal{T}}, T)$  is the PDE indicator associated with a generic element  $T \in \mathcal{T}$ . Note that  $\theta = 1$  corresponds to uniform refinement. In Section 5, we present three AFEMs in increasing order of complexity regarding data  $\mathcal{D} = (A, c, f)$  and prove their linear convergence.

The simplest algorithm, so-called GALERKIN, works for discrete  $\mathcal{D}$  and is the usual adaptive loop

$$\text{SOLVE} \rightarrow \text{ESTIMATE} \rightarrow \text{MARK} \rightarrow \text{REFINE}.$$

We assume SOLVE computes the exact Galerkin solution  $u_{\mathcal{T}}$ , so we refrain from addressing linear algebra issues. The module ESTIMATE computes the a posteriori

error indicator and the module MARK implements (1.7); in most of the paper we deal with weighted  $L^2$ -error indicators  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$  but we also address linear convergence for alternative estimators. The module REFINE bisects marked elements and perhaps a few more to keep meshes conforming (or  $\Lambda$ -admissible if they are non-conforming). We denote by  $\|u - u_{\mathcal{T}}\|_{\Omega}$  the energy error associated with the bilinear form  $\mathcal{B}$ . This error is monotone with refinement but may stagnate. We thus exploit the estimator reduction property with refinement, typical of  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$ , to show that the combined quantity

$$\zeta_{\mathcal{T}}(u_{\mathcal{T}})^2 := \|u - u_{\mathcal{T}}\|_{\Omega}^2 + \gamma \eta_{\mathcal{T}}(u_{\mathcal{T}})^2 \quad (1.8)$$

contracts in every iteration of GALERKIN for a suitable scaling parameter  $\gamma > 0$ . This readily leads to linear convergence of both  $|u - u_{\mathcal{T}}|_{H_0^1(\Omega)}$  and  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$ .

We next keep the coefficients  $(A, c)$  discrete but allow for a general  $f \in H^{-1}(\Omega)$ . This is to prevent the *multiplicative interaction* between  $(A, c)$  and  $u$  that occurs in (1.1) if we were to approximate  $(A, c)$ . In contrast, the effect of  $f$  is *linear* in (1.1). We show examples where  $\|u - u_{\mathcal{T}}\|_{\Omega}$  may stagnate because the adaptive process is dominated by oscillation  $\text{osc}_{\mathcal{T}}(f)_{-1}$  (preasymptotic regime). To compensate for this fact, we design a one-step AFEM with switch as in [Kreuzer, Veeser and Zangotti n.d.], the so-called AFEM-SW, that proceeds as GALERKIN provided  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$  dominates and otherwise reduces  $\text{osc}_{\mathcal{T}}(f)_{-1}$  separately. We show that, for suitable a parameter  $\gamma > 0$ , the combined quantity

$$\zeta_{\mathcal{T}}(u_{\mathcal{T}})^2 := \|u - u_{\mathcal{T}}\|_{\Omega}^2 + \gamma \eta_{\mathcal{T}}(u_{\mathcal{T}})^2 + \text{osc}_{\mathcal{T}}(f)_{-1}^2 \quad (1.9)$$

contracts in every loop of AFEM-SW. This yields linear convergence of the error  $|u - u_{\mathcal{T}}|_{H_0^1(\Omega)}$  and the estimator  $\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f) = \eta_{\mathcal{T}}(u_{\mathcal{T}}) + \text{osc}_{\mathcal{T}}(f)_{-1}$ .

The third algorithm is the two-step AFEM, the so-called AFEM-TS, which allows for general data  $\mathcal{D} = (A, c, f)$ . To handle the aforementioned nonlinear effect of  $(A, c)$  and also deal with general  $f$ , all data  $\mathcal{D}$  are first approximated by a routine DATA to a desired level of accuracy, which is adjusted at every step of AFEM-TS, and then fed to GALERKIN which handles discrete data. Suitably combining the accuracies of each intermediate module leads to linear convergence and optimal complexity of GALERKIN within each loop. The structure of AFEM-TS is flexible enough to easily handle discontinuous coefficients  $(A, c)$  with discontinuities that may not be aligned with the mesh. This is because the approximation of  $(A, c)$  by discontinuous polynomials takes place in  $L^p(\Omega)$  for  $p < \infty$ .

It is worth stressing two important points. First the approximation of  $\mathcal{D}$  is carried out by a GREEDY algorithm, which is shown to perform optimally starting from any refinement of  $\mathcal{T}_0$ . Second, the discontinuous piecewise polynomial approximations  $(\hat{A}, \hat{c})$  of  $(A, c)$  may not respect, for polynomial degree  $\geq 1$ , the positivity bounds associated with the coefficients. This requires a nonlinear correction of the output of GREEDY that restores positivity and does not deteriorate accuracy beyond a modest multiplicative constant. We postpone the discussion of these two delicate and technical processes to Section 7, which can be omitted in a first reading.

*Rate-optimality of AFEMs.* Showing that AFEMs outperform classical FEMs is a difficult but important matter. This reduces to proving a superior relation between the required degrees of freedom (or number of elements) for a desired accuracy; the former is in fact an acceptable measure of complexity. Showing that AFEMs deliver a performance comparable with the best entails the following basic ingredients:

- *Nonlinear approximation classes:* they classify functions in terms of the best possible algebraic decay rate of approximation  $e_N(v)_X$  of a given function  $v$  in a given norm  $X$  with  $N$  number of elements; roughly speaking, we say  $v \in \mathbb{A}_s$  if  $e_N(v)_X \lesssim N^{-s}$ . These classes are related to regularity classes (Sobolev, Besov, and Lipschitz) along Sobolev embedding lines.
- *Dörfler marking:* If the oscillation  $\text{osc}_{\mathcal{T}}(f)_{-1}$  is dominated by the PDE estimator  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$  for a given mesh  $\mathcal{T}$ , then any conforming refinement  $\mathcal{T}_* \geq \mathcal{T}$  of  $\mathcal{T}$  that reduces  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$  by a substantial amount induces a refined set  $\mathcal{R} := \mathcal{T} \setminus \mathcal{T}_*$  to modify  $\mathcal{T}$  into  $\mathcal{T}_*$  satisfying (1.7), namely  $\eta_{\mathcal{T}}(u_{\mathcal{T}}, \mathcal{R}) \geq \theta \eta_{\mathcal{T}}(u_{\mathcal{T}})$ .
- *Minimality of  $\mathcal{M}$ :* If the subset  $\mathcal{M} \subset \mathcal{T}$  in (1.7) is minimal, then the cardinality of  $\mathcal{M}$  compares favorably to the cardinality of the best mesh with a comparable error accuracy, thereby leading to rate-optimality of AFEM.

This is altogether the topic of Section 6. It is important to notice that membership in  $\mathbb{A}_s$  is never used explicitly by AFEM to learn about problem (1.1) and improve its resolution. The fundamental reason behind the superior performance of AFEM relative to FEM lies in nonlinear approximation theory. We illustrate now this point with the following insightful approximation example for  $d = 1$  and  $X = L^\infty(0, 1)$  [ DeVore 1998, Kahane 1961]: let  $\Omega = (0, 1)$ ,  $\mathcal{T}_N = \{[x_{j-1}, x_j]\}_{j=1}^N$  be a partition of  $\Omega$ , with

$$0 = x_0 < x_1 < \cdots < x_j < \cdots < x_N = 1,$$

and let  $v : \Omega \rightarrow \mathbb{R}$  be an absolutely continuous function to be approximated by a piecewise constant function  $v_N$  over  $\mathcal{T}_N$ . To quantify the difference between  $v$  and  $v_N$  we resort to the *maximum norm* and study two cases depending on the regularity of  $v$ . We define  $v_N(x) := v(x_{j-1})$  for all  $x_{j-1} \leq x < x_j$  and note that

$$|v(x) - v_N(x)| = |v(x) - v(x_{j-1})| \leq \int_{x_{j-1}}^x |v'(t)| dt.$$

- *Case 1:  $W_\infty^1$ -Regularity.* If  $u \in W_\infty^1(0, 1)$  and  $x_{j-1} \leq x < x_j$ , then

$$|v(x) - v_N(x)| \leq h_j \|v'\|_{L^\infty(x_{j-1}, x_j)} \quad \Rightarrow \quad \|v - v_N\|_{L^\infty(\Omega)} \leq \frac{1}{N} \|v'\|_{L^\infty(\Omega)}$$

for a *uniform* mesh. We thus deduce a rate  $N^{-1}$  using the same integrability  $L^\infty$  on both sides of the error estimate.

- *Case 2:  $W_1^1$ -Regularity.* Let  $\|v'\|_{L^1(\Omega)} = 1$  and  $\mathcal{T}_N$  be a *graded* partition so that

$\int_{x_{j-1}}^{x_j} |v'(t)| dt = \frac{1}{N}$ . Then, for  $x \in [x_{j-1}, x_j]$ ,

$$|v(x) - v(x_{j-1})| \leq \int_{x_{j-1}}^{x_j} |v'(t)| dt = \frac{1}{N} \quad \Rightarrow \quad \|v - v_N\|_{L^\infty(\Omega)} \leq \frac{1}{N} \|v'\|_{L^1(\Omega)}.$$

We thus conclude that we could achieve the same rate of convergence  $N^{-1}$  for rougher functions with just  $\|u'\|_{L^1(\Omega)} < \infty$ . Three comments are now in order. First, the contrast between Cases 1 and 2 is more dramatic for  $v(x) = x^\alpha$  with  $\alpha \in (0, 1)$  because Case 1 only yields the suboptimal rate  $\|v - v_N\|_{L^\infty(\Omega)} \leq N^{-\alpha}$ . Second,  $\mathcal{T}_N$  in Case 2 *equidistributes* the max-error, a concept that will permeate our discussions later. Third, the optimal rate of Case 2 is due to the exchange of differentiability with integrability along the critical Sobolev embedding line between the left and right-hand sides of the error estimate (nonlinear Sobolev scale), while Case 1 relies on the linear Sobolev scale with constant integrability.

We exploit and further elaborate these concepts in Section 6 to show rate-optimality of the three algorithms GALERKIN, AFEM-SW, AFEM-TS, discussed in Section 5, provided  $u$  and  $\mathcal{D}$  belong to suitable approximation classes. We also investigate the relation between these approximation classes with regularity classes, allowing for discontinuous coefficients, and present a rather complete discussion.

*Mesh refinement.* A key component of any adaptive algorithm, such as the three AFEMs already described, is the routine REFINE which refines a current mesh  $\mathcal{T}$  into  $\mathcal{T}_*$  to improve resolution. In Section 8 we study the *bisection method*, which is the most popular method to refine simplicial meshes in  $\mathbb{R}^d$  for  $d \geq 1$ . For simplicity we focus our attention on this method, but most results apply to other refinement strategies such as quad-trees (for quadrilaterals), octrees (for hexagons) and red-green (for simplicial meshes). We do not insist on these extensions but refer to [Bonito and Nochetto 2010] for details.

Given an initial grid  $\mathcal{T}_0$  with a suitable labeling, the bisection method splits a given simplex into two children. The rules for successive cutting of simplices, for instance newest vertex bisection for  $d = 2$ , are such that the ensuing meshes are shape regular (with a uniform constant only depending on  $\mathcal{T}_0$  and  $d$ ). However, bisection may not be completely local to keep conformity. The analysis of propagation of refinement is a delicate combinatorial problem. It is easy to see by example, that bisecting one element of large generation (i.e. number of bisections needed to produce it) may require a chain of elements with length similar to the generation. Therefore, the number of refined elements in one step cannot be bounded by the number of marked elements. The following amazing estimate by [Binev, Dahmen and DeVore 2004] for  $d = 2$  and [Stevenson 2008] for  $d > 2$  shows that the cumulative effect of bisection counting from  $\mathcal{T}_0$  all the marked elements  $\mathcal{M}_j$  is quasi-optimal: there exists a constant  $D > 0$ , depending on  $\mathcal{T}_0$  and  $d$ , such that

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \leq D \sum_{j=0}^{k-1} \#\mathcal{M}_j. \quad (1.10)$$



This estimate is crucial for the study of rate-optimality of AFEM and is proved in Section 8 for  $d = 2$  and for both conforming refinement and  $\Lambda$ -admissible refinement. The latter is a systematic way to handle non-conforming meshes that goes back to [Beirão da Veiga, Canuto, Nochetto, Vacca and Verani 2023]. It associates a computable global index to hanging nodes and imposes a restriction to them not to exceed a preassigned value  $\Lambda \geq 0$ ; if  $\Lambda = 0$  then the mesh is conforming. Section 8 also discusses several interesting geometric properties of  $\Lambda$ -admissible meshes which turn out to be crucial for discontinuous Galerkin methods. Since Section 8 is quite technical, it can be skipped in a first reading.

*Discontinuous Galerkin methods.* These methods, so-called dG, are the natural first step to investigate the role of non-conformity in adaptivity, namely that the discrete space of discontinuous piecewise polynomials  $\mathbb{V}_{\mathcal{T}}$  is no longer a subspace of  $H_0^1(\Omega)$ . To this end, we study the symmetric interior penalty dG method in Section 9 on  $\Lambda$ -admissible partitions  $\mathcal{T}$  of  $\mathcal{T}_0$ . Such dG methods exhibit some characteristic and novel features with respect to conforming FEMs: the most notable one is the presence of weighted jumps that stabilize the method and compensate for the lack of  $H^1$ -conformity. We consider the formulation with lifting, which allows for minimal regularity  $u \in H_0^1(\Omega)$ , and forcing  $f \in H^{-1}(\Omega)$  despite that  $\mathbb{V}_{\mathcal{T}}$  is not a subspace of  $H_0^1(\Omega)$ . The latter is possible because, within the framework of AFEM-TS,  $f$  is approximated by a piecewise polynomial  $P_{\mathcal{T}}f$  for which the pairing with functions in  $\mathbb{V}_{\mathcal{T}}$  is meaningful.

The fact that jumps are not monotone upon refinement constitutes one of the main obstructions for studying adaptivity for dG. To circumvent this issue we follow [Bonito and Nochetto 2010], who in turn modified the original approach of [Karakashian and Pascal 2007], and introduce the largest conforming subspace  $\mathbb{V}_{\mathcal{T}}^0$  of  $\mathbb{V}_{\mathcal{T}}$ . It turns out that, despite being coarser,  $\mathbb{V}_{\mathcal{T}}^0$  exhibits a local resolution comparable with  $\mathbb{V}_{\mathcal{T}}$  because of key geometric properties of  $\Lambda$ -admissible meshes that control the degree of non-conformity of  $\mathcal{T}$ . In addition,  $\mathbb{V}_{\mathcal{T}}^0$  is responsible for the scaled jumps to be bounded by the PDE estimator  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$ . Exploiting properties of  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$ , similar to the conforming case, leads to a quasi-orthogonality estimate for the dG norm, a dG variant of the Pythagoras equality. This is instrumental to prove a contraction property for the error plus scaled estimator and deduce convergence for both GALERKIN and AFEM-TS. Moreover, we derive rate-optimality for both algorithms provided  $u$  and  $\mathcal{D}$  belong to suitable approximation classes. Such classes are the same as for conforming AFEMs: in fact we prove that the approximation classes for  $u$  using continuous and discontinuous piecewise polynomials on  $\Lambda$ -admissible meshes coincide.

*Inf-sup stable AFEMs.* The convergence and optimality theories developed in Sections 5 and 6 rely on the bilinear form in (1.2) being *coercive*. We remove this strong restriction in Section 10 and consider uniformly *inf-sup stable* FEMs on conforming refinements  $\mathcal{T}_j$  of  $\mathcal{T}_0$ . The lack of an energy norm and its monotone behavior upon refinement has been an obstacle for the study of this class of prob-

lems. We follow the recent work by M. Feischl [Feischl 2022], who introduced the following form of *quasi-orthogonality* between consecutive Galerkin solutions  $u_j \in \mathbb{V}_j$ , originally proposed in [Carstensen, Feischl, Page and Praetorius 2014] as part of an abstract set of axioms of adaptivity:

$$\sum_{k=j}^{j+N} \|u_{k+1} - u_k\|_{\mathbb{V}}^2 \leq C(N) \|u - u_j\|_{\mathbb{V}}^2 \quad j \geq 0, \quad (1.11)$$

where  $C(N)/N \rightarrow 0$  as  $N \rightarrow \infty$ . This is our departing point to develop a *variational approach* to prove linear convergence of  $u_j$  provided data  $\mathcal{D}$  is discrete; the latter is reflected in an equivalence property between error and estimator (without oscillation). This is the context of a GALERKIN routine, which is next used as a building block together with DATA for a AFEM-TS that handles general data  $\mathcal{D}$ . Moreover, we prove rate-optimality for both algorithms, thereby extending Sections 5 and 6.

This discussion is rather abstract. We specialize it to the Stokes equations for viscous incompressible fluids and mixed formulations of (1.1) using Raviart-Thomas-Nedelec and Brezzi-Douglas-Marini elements. We thereby obtain convergence and rate-optimality for AFEM-SW for the Stokes equations and AFEM-TS for mixed methods with variable and possibly discontinuous coefficients  $(A, c)$ .

We conclude with a complete proof of (1.11) following [Feischl 2022]. This is a tour de force in applied linear algebra and is rather technical. It can be omitted in a first reading.

*Acknowledgments.* We are grateful to R. DeVore, P. Morin, and A. Salgado for discussions about regularity classes and to Ch. Kreuzer for pointing out a gap in an earlier version of the proof of Theorem 5.43. We are also thankful to G. Vacca for providing simulations and D. Fassino for assistance with the manuscript.

This research was supported by the NSF grants DMS-2110811 (AB) and DMS-1908267 (RHN), the MUR-PRIN grants 20227K44ME (CC) and 201752HKH (AV), and the INdAM research group GNCS (CC and AV). RHN was also partially supported by NSF grant DMS-1929284 while he was in residence at the Institute for Computational and Experimental Research in Mathematics in Providence, RI, during the Numerical PDEs: Analysis, Algorithms, and Data Challenges semester program.

## 2 Linear Elliptic Boundary Value Problems: Basic Theory

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 two model boundary value problems that are instrumental in our subsequent analysis. We next present the so-called inf-sup theory that characterizes existence, uniqueness and stability of variational problems, and apply it to coercive and saddle point problems. These two classes will play essential roles later.

## 2.1 Sobolev spaces: scaling and embedding

Let  $\Omega \subset \mathbb{R}^d$  with  $d > 1$  be a Lipschitz and bounded domain, and let  $k \in \mathbb{N}, 1 \leq p \leq \infty$ . The Sobolev space  $W_p^k(\Omega)$  is defined by

$$W_p^k(\Omega) := \{v : \Omega \rightarrow \mathbb{R} \mid D^\alpha v \in L^p(\Omega) \ \forall |\alpha| \leq k\},$$

and is a Banach space with the norm

$$\|v\|_{W_p^k(\Omega)} = \left( \sum_{j=1}^k |v|_{W_p^j(\Omega)}^p \right)^{\frac{1}{p}}, \quad |v|_{W_p^j(\Omega)} = \left( \int_{\Omega} |D^j v|^p \right)^{\frac{1}{p}}.$$

The space  $W_p^k(\Omega; \mathbb{R}^m)$  is the space  $W_p^k(\Omega)$  of vector or matrix-valued functions. If  $p = 2$  we write  $H^k(\Omega) = W_2^k(\Omega)$  and note that this is a Hilbert space. We denote by  $H_0^1(\Omega) \subset H^1(\Omega)$  the completion of  $C_0^\infty(\Omega)$  within  $H^1(\Omega)$ .

Sobolev spaces  $W_p^k(\Omega)$  of fractional order  $k > 0$  can be defined as well, by applying the real interpolation method between  $W_p^{[k]}(\Omega)$  and  $W_p^{[k]+1}(\Omega)$  (see [Bergh and L fstr m 1976] and [Adams and Fournier 2003] for the details). The subsequent definitions and properties hold for Sobolev spaces of integer or fractional order.

The *Sobolev number* of  $W_p^k(\Omega)$  is given by

$$\text{sob}(W_p^k) := k - \frac{d}{p}. \quad (2.1)$$

This number governs the scaling properties of the semi-norm  $|v|_{W_p^k(\Omega)}$ , because rescaling variables  $\widehat{x} = \frac{1}{h}x$  for all  $x \in \Omega$ , transforms  $\Omega$  into  $\widehat{\Omega}$  and  $v$  into  $\widehat{v}$ , while the corresponding norms scale as

$$|\widehat{v}|_{W_p^k(\widehat{\Omega})} = h^{\text{sob}(W_p^k)} |v|_{W_p^k(\Omega)}.$$

## 2.2 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 [Evans 2010, Gilbarg and Trudinger 2001, Grisvard 2011] for details. We use the notation  $\hookrightarrow$  to denote a continuous embedding.

**Lemma 2.1 (Sobolev embedding)** *Let  $m > k \geq 0$  and assume  $\text{sob}(W_q^m) > \text{sob}(W_p^k)$ . Then the embedding  $W_q^m(\Omega) \hookrightarrow W_p^k(\Omega)$  is compact and*

$$\|v\|_{W_p^k(\Omega)} \leq C \|v\|_{W_q^m(\Omega)} \quad \forall v \in W_q^m(\Omega),$$

where  $C = C(m, k, q, p, \Omega, d)$ .

We say that two Sobolev spaces are in the same nonlinear Sobolev scale if they have the same Sobolev number; see Figure 2.1. We thus note that for compactness

the space  $W_q^m(\Omega)$  must be above the Sobolev scale of  $W_p^k(\Omega)$ , i.e.  $\text{sob}(W_q^m) > \text{sob}(W_p^k)$ .

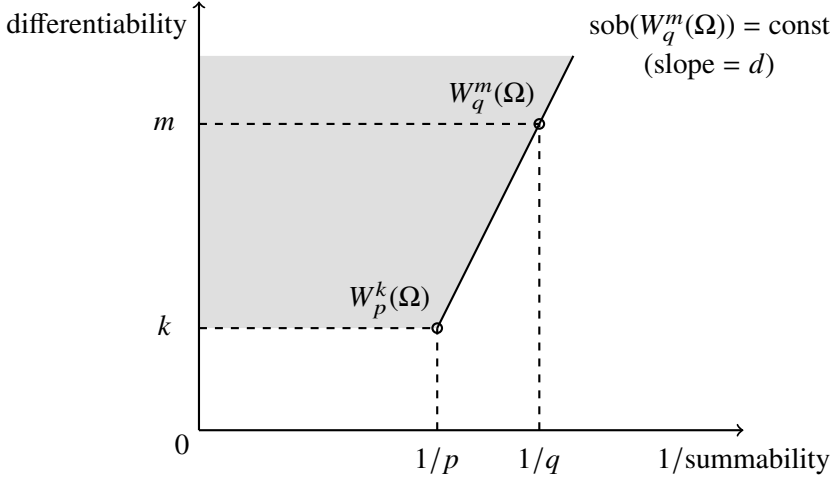


Figure 2.1. DeVore diagram [DeVore 1998]. The space  $W_p^k(\Omega)$  is represented by the point  $(1/p, k)$  in the first quadrant. The line  $\text{sob}(W_q^m) = \text{const} = \text{sob}(W_p^k)$ , with slope  $d$ , may be called the (critical) Sobolev embedding line for  $W_p^k(\Omega)$ . It represents all Sobolev spaces having the same Sobolev number as  $W_p^k(\Omega)$ . Sobolev spaces corresponding to points inside the gray region and on its boundary on the vertical axis are compactly embedded in  $W_p^k(\Omega)$ . Spaces on the oblique and horizontal lines emanating from  $W_p^k(\Omega)$  are generally continuously, but not compactly embedded, in  $W_p^k(\Omega)$  with exceptions such as  $p = \infty$ . Note that indices  $k$  and  $m$  may take non-integer values, corresponding to fractional Sobolev spaces.

The assumption on the Sobolev number cannot be relaxed. To see this, consider  $\Omega$  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  $v \in W_d^1(\Omega) \setminus L^\infty(\Omega)$ , but

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

Therefore, equality cannot be expected in the embedding theorem. On the other hand, consider  $d = 1$  and the spaces  $W_1^1(\Omega)$  and  $L^\infty(\Omega)$ . We see that  $\text{sob}(W_1^1) = \text{sob}(L^\infty) = 0$  but  $W_1^1(\Omega)$  is compactly embedded in  $L^\infty(\Omega)$  in this case. This shows that these two spaces are in the same nonlinear Sobolev scale and that the above inequality between Sobolev numbers for a compact embedding is only sufficient.

Moreover, if  $0 < \alpha = \text{sob}(W_p^k) \leq 1$ , then functions of  $W_p^k(\Omega)$  are Hölder- $\alpha$  and

$$|v|_{C^{0,\alpha}(\overline{\Omega})} \leq C |v|_{W_p^k(\Omega)} \quad \forall v \in W_p^k(\Omega).$$

This allows for the use of the standard Lagrange interpolation operator. We will exploit this fact later in Section 3.

**Lemma 2.2 (first Poincaré inequality)** *Let  $\Omega \subset \mathbb{R}^d$  be a bounded and Lipschitz domain. Then there is a constant  $C_P = C_d |\Omega|^{1/d}$  such that*

$$\|v\|_{L^2(\Omega)} \leq C_P \|\nabla v\|_{L^2(\Omega)} \quad \forall v \in H_0^1(\Omega). \quad (2.2)$$

*The same inequality is valid in  $W_p^1(\Omega)$  for any  $1 \leq p \leq \infty$  provided  $v$  has vanishing trace [Gilbarg and Trudinger 2001, p.158].*

**Lemma 2.3 (second Poincaré inequality)** *There exists  $C_P$  depending on  $\Omega$  such that*

$$\|v - \bar{v}\|_{L^2(\Omega)} \leq C_P \|\nabla v\|_{L^2(\Omega)} \quad \forall v \in H^1(\Omega), \quad (2.3)$$

*where  $\bar{v} := |\Omega|^{-1} \int_{\Omega} v$ . The best constant within the class of convex domains is  $C_P = \frac{1}{\pi} \text{diam}(\Omega)$  [Payne and Weinberger 1960]. The same inequality is valid in  $W_p^1(\Omega)$  for  $1 \leq p \leq \infty$ . [Gilbarg and Trudinger 2001].*

**Lemma 2.4 (traces)** *Let  $\Omega$  be Lipschitz. There exists a unique linear operator  $T: H^1(\Omega) \rightarrow L^2(\partial\Omega)$  such that*

$$\begin{aligned} \|Tv\|_{L^2(\partial\Omega)} &\leq c(\Omega) \|v\|_{H^1(\Omega)} & \forall v \in H^1(\Omega), \\ Tv &= v & \forall v|_{\partial\Omega} \in C^0(\bar{\Omega}) \cap H^1(\Omega). \end{aligned}$$

*The operator  $T$  is also well defined on  $W_p^1(\Omega)$  for  $1 \leq p \leq \infty$ . [Evans 2010, Grisvard 1985]*

Since  $Tv = v|_{\partial\Omega}$  for continuous functions, we write  $v$  for  $Tv$ . The image of  $T$  is a strict subspace of  $L^2(\partial\Omega)$ , the so-called  $H^{1/2}(\partial\Omega)$ . This is a Hilbert space if equipped with the norm  $\|g\|_{H^{1/2}(\partial\Omega)} := \inf\{\|v\|_{H^1(\Omega)} \mid Tv = g\}$ , and  $T$  is continuous with respect to this norm, i.e.,

$$\|Tv\|_{H^{1/2}(\partial\Omega)} \leq \|v\|_{H^1(\Omega)} \quad \forall v \in H^1(\Omega). \quad (2.4)$$

The definition of  $H_0^1(\Omega)$  can be reconciled with that of traces because

$$H_0^1(\Omega) = \{v \in H^1(\Omega) \mid v = 0 \text{ on } \partial\Omega\}.$$

We point out that, in view of Lemma 2.2, the semi-norm  $|v|_{H_0^1(\Omega)} := \|\nabla v\|_{L^2(\Omega)}$  is a norm in  $H_0^1(\Omega)$ . We let  $H^{-1}(\Omega)$  be the dual of  $H_0^1(\Omega)$ , with corresponding norm  $\|f\|_{H^{-1}(\Omega)} := \sup_{v \in H_0^1(\Omega)} \frac{\langle f, v \rangle}{|v|_{H_0^1(\Omega)}}$ . These definitions extend to any  $p \in (1, \infty)$ .

**Lemma 2.5 (Gauss divergence theorem)** *If  $\mathbf{n}$  is the outer unit normal to  $\Omega$ , then*

$$\int_{\Omega} \text{div } \mathbf{w} = \int_{\partial\Omega} \mathbf{w} \cdot \mathbf{n} \quad \forall \mathbf{w} \in W_1^1(\Omega; \mathbb{R}^d).$$

**Lemma 2.6 (Green's formula)** *The integration by parts formula holds*

$$\int_{\Omega} \operatorname{div} \mathbf{w} v = - \int_{\Omega} \mathbf{w} \cdot \nabla v + \int_{\partial\Omega} v \mathbf{w} \cdot \mathbf{n} \quad \forall v \in H^1(\Omega), \mathbf{w} \in H^1(\Omega; \mathbb{R}^d).$$

### 2.3 Examples of boundary value problems

We consider two model elliptic problems in this paper. We start with the *scalar diffusion-reaction equation* with variable coefficients

$$\begin{aligned} L[u] &:= -\operatorname{div}(A \nabla u) + cu = f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{2.5}$$

where  $\Omega \subset \mathbb{R}^d$  is a bounded domain with Lipschitz boundary,  $A \in L^\infty(\Omega; \mathbb{R}^{d \times d})$  is a diffusion tensor uniformly symmetric positive definite (SPD) over  $\Omega$ , i. e., there exist constants  $0 < \alpha_1 \leq \alpha_2$  such that

$$\alpha_1 |\xi|^2 \leq \xi^T A(x) \xi \leq \alpha_2 |\xi|^2 \quad \forall x \in \Omega, \xi \in \mathbb{R}^d, \tag{2.6}$$

$c \in L^\infty(\Omega)$ ,  $c \geq 0$  is a reaction term, and  $f \in L^2(\Omega)$  is a scalar load term.

To derive the variational formulation of (2.5) we let  $\mathbb{V} = H_0^1(\Omega)$  and  $\mathbb{V}^* = H^{-1}(\Omega)$ . Since  $H_0^1(\Omega)$  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 (2.5). We next multiply (2.5) with a test function  $v \in H_0^1(\Omega)$ , integrate over  $\Omega$  and use Lemma 2.6 (Green's formula), provided  $\mathbf{w} = -A \nabla u \in H^1(\Omega; \mathbb{R}^d)$ , to obtain

$$u \in \mathbb{V} : \quad \mathcal{B}[u, v] = \langle f, v \rangle \quad \forall v \in \mathbb{V}. \tag{2.7}$$

Here,  $\mathcal{B} : \mathbb{V} \times \mathbb{V} \rightarrow \mathbb{R}$  stands for the bilinear form

$$\mathcal{B}[w, v] := \int_{\Omega} \nabla v \cdot A \nabla w + c v w \quad \forall v, w \in \mathbb{V}, \tag{2.8}$$

and  $\langle \cdot, \cdot \rangle$  stands for the  $L^2(\Omega)$ -scalar product and also for a duality pairing  $H^{-1}(\Omega) - H_0^1(\Omega)$ . Since  $A$  is assumed to be symmetric, the bilinear form  $\mathcal{B}$  is also symmetric; however  $A$  does not have to be symmetric in general. Note that the weak form (2.7) allows for fluxes  $\mathbf{w} \in L^2(\Omega; \mathbb{R}^d)$  and forcing  $f \in H^{-1}(\Omega)$ . We examine existence, uniqueness and stability of (2.7) in Section 2.4 below.

We assume *homogeneous Dirichlet* boundary condition in (2.5) for simplicity and because this will be our basic setting later. However, we could allow a *non-homogeneous Dirichlet* condition in the sense of traces

$$Tu = g \quad \text{on } \partial\Omega, \tag{2.9}$$

for any given function  $g \in H^{\frac{1}{2}}(\partial\Omega)$ . To write the companion variational formulation to (2.7), we first introduce the subspace  $\mathbb{V}(g) \subset H^1(\Omega)$  of functions  $v$  with trace  $Tv = g$  on  $\partial\Omega$ , and then rewrite (2.7) as follows:

$$u \in \mathbb{V}(g) : \quad \mathcal{B}[u, v] = \langle f, v \rangle \quad \forall v \in \mathbb{V}(0). \tag{2.10}$$

Moreover, we could consider a *Robin* boundary condition for given functions  $g$  and  $p$  on  $\partial\Omega$

$$\mathbf{A}\nabla u \cdot \mathbf{n} + pu = g \quad \text{on } \partial\Omega, \quad (2.11)$$

where  $\mathbf{n}$  is the outer unit normal to  $\Omega$ . To figure out the variational formulation, we now multiply the PDE in (2.5) by a test function  $v \in H^1(\Omega)$  and integrate by parts to find the following variant of (2.7)

$$u \in H^1(\Omega) : \quad \mathcal{B}[u, v] = \ell(v) \quad \forall v \in H^1(\Omega), \quad (2.12)$$

where for all  $v, w \in H^1(\Omega)$

$$\mathcal{B}[w, v] := \int_{\Omega} \nabla v \cdot \mathbf{A} \nabla w + c v w + \int_{\partial\Omega} p v w, \quad \ell(v) := \langle f, v \rangle + \int_{\partial\Omega} g v. \quad (2.13)$$

We realize that (2.13) makes sense for  $p \in L^\infty(\partial\Omega)$ ,  $p \geq 0$  and  $g \in H^{-\frac{1}{2}}(\partial\Omega)$ , the dual space of  $H^{\frac{1}{2}}(\partial\Omega)$ , whence the last integral in (2.13) means a duality pairing. If  $p = 0$ , then (2.13) reduces to the weak form of the *Neumann* boundary condition.

The second model problem is the *Stokes system*, which is the simplest model of a stationary viscous incompressible fluid. Given an external force  $\mathbf{f} \in L^2(\Omega; \mathbb{R}^d)$ , let the velocity-pressure pair  $(\mathbf{u}, p)$  satisfy the momentum and incompressibility equations with no-slip boundary condition:

$$\begin{aligned} -\Delta \mathbf{u} + \nabla p &= \mathbf{f} && \text{in } \Omega, \\ \operatorname{div} \mathbf{u} &= 0 && \text{in } \Omega, \\ \mathbf{u} &= 0 && \text{on } \partial\Omega. \end{aligned} \quad (2.14)$$

For the variational formulation we consider two Hilbert spaces  $\mathbb{V} = H_0^1(\Omega; \mathbb{R}^d)$  and  $\mathbb{Q} = L_0^2(\Omega)$ , where  $L_0^2(\Omega)$  is the space of  $L^2$  functions with zero mean value. The space  $H_0^1(\Omega; \mathbb{R}^d)$  takes care of the no-slip boundary values of the velocity. We first multiply the momentum equation in (2.14) by  $\mathbf{v} \in \mathbb{V}$ , assume  $\mathbf{u} \in H^2(\Omega; \mathbb{R}^d)$  and use Lemma 2.6 (Green's formula) component-wise. We next multiply the incompressibility equation in (2.14) by  $q \in \mathbb{Q}$  and integrate over  $\Omega$ . We end up with the following variational formulation: find  $(\mathbf{u}, p) \in \mathbb{V} \times \mathbb{Q}$  such that

$$\begin{aligned} a[\mathbf{u}, \mathbf{v}] + b[p, \mathbf{v}] &= \langle \mathbf{f}, \mathbf{v} \rangle && \forall \mathbf{v} \in \mathbb{V}, \\ b[q, \mathbf{u}] &= 0 && \forall q \in \mathbb{Q}. \end{aligned} \quad (2.15)$$

Here the bilinear forms  $a: \mathbb{V} \times \mathbb{V} \rightarrow \mathbb{R}$  and  $b: \mathbb{Q} \times \mathbb{V} \rightarrow \mathbb{R}$  read

$$a[\mathbf{w}, \mathbf{v}] := \int_{\Omega} \nabla \mathbf{v} : \nabla \mathbf{w} = \sum_{i=1}^d \int_{\Omega} \nabla v_i \cdot \nabla w_i \quad \forall \mathbf{v}, \mathbf{w} \in \mathbb{V}.$$

and

$$b[q, \mathbf{v}] := - \int_{\Omega} q \operatorname{div} \mathbf{v} \quad \forall q \in \mathbb{Q}, \mathbf{v} \in \mathbb{V}.$$

We observe that  $a[\mathbf{w}, \mathbf{v}]$  does not require  $\mathbf{w} \in H^2(\Omega; \mathbb{R}^d)$  and that (2.15) makes

sense for  $\mathbf{f} \in H^{-1}(\Omega; \mathbb{R}^d)$ ; note that the second equation in (2.15) is always satisfied for constant  $q$  due to Gauss' theorem, which explains the choice  $q \in \mathbb{Q}$ . Furthermore, (2.15) can be reformulated as (2.7), namely

$$(\mathbf{u}, p) \in \mathbb{V} \times \mathbb{Q} : \quad \mathcal{B}[(\mathbf{u}, p), (\mathbf{v}, q)] = \langle \mathbf{f}, \mathbf{v} \rangle \quad \forall (\mathbf{v}, q) \in \mathbb{V} \times \mathbb{Q}, \quad (2.16)$$

with

$$\mathcal{B}[(\mathbf{u}, p), (\mathbf{v}, q)] := a[\mathbf{u}, \mathbf{v}] + b[p, \mathbf{v}] + b[q, \mathbf{u}].$$

We discuss existence, uniqueness and stability of (2.16) in Section 2.4.

We could formulate the Stokes system with other boundary conditions. First, we may allow a non-homogeneous Dirichlet condition  $\mathbf{u} = \mathbf{g}$  for a given function  $\mathbf{g} \in H^{\frac{1}{2}}(\partial\Omega; \mathbb{R}^d)$  satisfying the compatibility condition  $\int_{\Omega} \mathbf{g} \cdot \mathbf{n} = 0$  imposed by Gauss' theorem, and proceed as in the scalar case (2.10). Second, to deal with a Neumann boundary condition, we introduce the stress tensor  $\boldsymbol{\sigma}(\mathbf{u}, p) := \frac{1}{2}(\nabla \mathbf{u} + \nabla \mathbf{u}^T) - p\mathbf{I}$  and the symmetric part of the velocity gradient  $\boldsymbol{\varepsilon}(\mathbf{u}) := \frac{1}{2}(\nabla \mathbf{u} + \nabla \mathbf{u}^T)$ . Then instead of (2.14) we could write the strong form of the Neumann problem as

$$\begin{aligned} -\operatorname{div} \boldsymbol{\sigma}(\mathbf{u}, p) &= \mathbf{f}, \quad \operatorname{div} \mathbf{u} = 0, \quad \text{in } \Omega \\ \boldsymbol{\sigma}(\mathbf{u}, p)\mathbf{n} &= \mathbf{g} \quad \text{on } \partial\Omega, \end{aligned} \quad (2.17)$$

and its weak form as (2.15) but with

$$\mathbb{V} := \{\mathbf{v} \in H^1(\Omega; \mathbb{R}^d) : \int_{\Omega} \mathbf{v} = \mathbf{0}\}, \quad \mathbb{Q} := L^2(\Omega), \quad (2.18)$$

as well as bilinear form  $a$  and right-hand side

$$a[\mathbf{u}, \mathbf{v}] := \int_{\Omega} \boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v}), \quad \ell(\mathbf{v}) := \langle \mathbf{f}, \mathbf{v} \rangle + \int_{\partial\Omega} \mathbf{g} \cdot \mathbf{v} \quad (2.19)$$

for all  $\mathbf{v} \in \mathbb{V}$ . Again, the last integral in (2.19) is to be interpreted as a duality pairing for  $\mathbf{g} \in H^{-\frac{1}{2}}(\partial\Omega; \mathbb{R}^d)$ .

## 2.4 Inf-sup theory

We present a functional framework for existence, uniqueness, and stability of variational problems of the form (2.7) or (2.16). 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

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

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

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



The following result relates a continuous bilinear form  $\mathcal{B}: \mathbb{V} \times \mathbb{W} \rightarrow \mathbb{R}$  with an operator  $B \in L(\mathbb{V}; \mathbb{W})$  [Nečas 1962, Babuška 1971].

**Theorem 2.7 (Banach-Nečas)** *Let  $\mathcal{B}: \mathbb{V} \times \mathbb{W} \rightarrow \mathbb{R}$  be a continuous bilinear form with norm*

$$\|\mathcal{B}\| := \sup_{v \in \mathbb{V}} \sup_{w \in \mathbb{W}} \frac{\mathcal{B}[v, w]}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}}. \quad (2.20)$$

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

$$\langle Bv, w \rangle_{\mathbb{W}} = \mathcal{B}[v, w] \quad \forall v \in \mathbb{V}, w \in \mathbb{W}$$

*with operator norm*

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

*Moreover, the bilinear form  $\mathcal{B}$  satisfies the two conditions*

$$i) \text{ there exists } \alpha > 0 \text{ such that } \alpha \|v\|_{\mathbb{V}} \leq \sup_{w \in \mathbb{W}} \frac{\mathcal{B}[v, w]}{\|w\|_{\mathbb{W}}} \text{ for all } v \in \mathbb{V}; \quad (2.21a)$$

$$ii) \text{ for every } 0 \neq w \in \mathbb{W} \text{ there exists } v \in \mathbb{V} \text{ such that } \mathcal{B}[v, w] \neq 0; \quad (2.21b)$$

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

$$\|B^{-1}\|_{L(\mathbb{W}, \mathbb{V})} \leq \alpha^{-1}. \quad (2.22)$$

We consider now the abstract variational problem

$$u \in \mathbb{V}: \quad \mathcal{B}[u, v] = \langle f, v \rangle \quad \forall v \in \mathbb{W}, \quad (2.23)$$

The following result, due to Nečas [Nečas 1962, Theorem 3.3], characterizes properties of the bilinear form  $\mathcal{B}$  that imply that (2.23) is well-posed.

**Theorem 2.8 (Nečas)** *Let  $\mathcal{B}: \mathbb{V} \times \mathbb{W} \rightarrow \mathbb{R}$  be a continuous bilinear form. Then the variational problem (2.23) 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  $\mathcal{B}$  satisfies one of the equivalent inf-sup conditions:*

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

$$i) \quad \sup_{w \in \mathbb{W}} \frac{\mathcal{B}[v, w]}{\|w\|_{\mathbb{W}}} \geq \alpha \|v\|_{\mathbb{V}} \quad \text{for some } \alpha > 0; \quad (2.24a)$$

$$ii) \text{ For every } 0 \neq w \in \mathbb{W} \text{ there exists } v \in \mathbb{V} \text{ such that } \mathcal{B}[v, w] \neq 0. \quad (2.24b)$$

(2) *There holds*

$$\inf_{v \in \mathbb{V}} \sup_{w \in \mathbb{W}} \frac{\mathcal{B}[v, w]}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}} > 0, \quad \inf_{w \in \mathbb{W}} \sup_{v \in \mathbb{V}} \frac{\mathcal{B}[v, w]}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}} > 0. \quad (2.25)$$

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

$$\inf_{v \in \mathbb{V}} \sup_{w \in \mathbb{W}} \frac{\mathcal{B}[v, w]}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}} = \inf_{w \in \mathbb{W}} \sup_{v \in \mathbb{V}} \frac{\mathcal{B}[v, w]}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}} = \alpha. \quad (2.26)$$

In addition, the solution  $u$  of (2.23) satisfies the stability estimate

$$\|u\|_{\mathbb{V}} \leq \alpha^{-1} \|f\|_{\mathbb{W}^*}. \quad (2.27)$$

The equality in (2.26) 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, (2.24) is simpler to verify than (2.26) and  $\alpha$  of (2.26) is the largest possible  $\alpha$  in (2.24a). Since (2.22) shows that  $\|B^{-1}\|_{L(\mathbb{W},\mathbb{V})}$  is the best inf-sup constant  $\alpha$  in (2.21a), we readily obtain the following result.

**Corollary 2.9 (well posedness implies inf-sup)** *Assume that the variational problem (2.23) admits a unique solution  $u \in \mathbb{V}$  for all  $f \in \mathbb{W}^*$  so that*

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

*Then  $\mathcal{B}$  satisfies the inf-sup condition (2.26) with  $\alpha \geq C^{-1}$ .*

We next apply these abstract results to two special but important cases. The first class are problems with coercive bilinear form and the second one comprises problems of saddle point type.

**Coercive Problems.** An existence and uniqueness result for coercive bilinear forms was established by Lax and Milgram eight years prior to the result by Nečas [Lax and Milgram 1954]. Coercivity of  $\mathcal{B}$  is a sufficient condition for existence and uniqueness but it is not necessary. In this case, we assume  $\mathbb{V} = \mathbb{W}$ .

**Corollary 2.10 (Lax-Milgram)** *Let  $\mathcal{B}: \mathbb{V} \times \mathbb{V} \rightarrow \mathbb{R}$  be a continuous bilinear form that is coercive, namely there exists  $\alpha > 0$  such that*

$$\mathcal{B}[v, v] \geq \alpha \|v\|_{\mathbb{V}}^2 \quad \forall v \in \mathbb{V}. \quad (2.28)$$

*Then (2.23) has a unique solution that satisfies the stability estimate (2.27).*

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

$$\mathcal{B}[v, w] = \mathcal{B}[w, v] \quad \forall v, w \in \mathbb{V},$$

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

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

For the reaction-diffusion equation (2.5) the bilinear form given in (2.8) satisfies

$$0 < \alpha_1 \leq \alpha \leq \|\mathcal{B}\| \leq \alpha_2 + \|c\|_{L^\infty(\Omega)} C_P^2, \quad (2.29)$$

where  $C_P$  is the constant in Lemma 2.2 (first Poincaré inequality). Coercivity and continuity of  $\mathcal{B}$ , with constants  $c_{\mathcal{B}} = \alpha_1$  and  $C_{\mathcal{B}} = \|\mathcal{B}\|$ , in turn imply that the *energy norm*  $\|\cdot\|_{\Omega}$  is equivalent to the natural norm  $\|\cdot\|_{\mathbb{V}}$  in  $\mathbb{V} = H_0^1(\Omega)$ :

$$c_{\mathcal{B}} \|v\|_{\mathbb{V}}^2 \leq \|v\|_{\Omega}^2 \leq C_{\mathcal{B}} \|v\|_{\mathbb{V}}^2 \quad \forall v \in \mathbb{V}. \quad (2.30)$$

Moreover, it is rather easy to show that for symmetric and coercive  $\mathcal{B}$  the solution

$u$  of (2.23) is the unique minimizer of the quadratic energy

$$J[v] := \frac{1}{2} \mathcal{B}[v, v] - \langle f, v \rangle \quad \forall v \in \mathbb{V},$$

i. e.,  $u = \operatorname{argmin}_{v \in \mathbb{V}} J[v]$ . In particular, the energy norm and the quadratic energy play a relevant role in Sections 5, 6 and 9.

This framework applies to the scalar diffusion-reaction equation (2.7) with *homogeneous Dirichlet* condition. Since the full  $H^1$ -norm  $\|\cdot\|_{H^1(\Omega)}$  and the seminorm  $|\cdot|_{H_0^1(\Omega)}$  are equivalent in the space  $\mathbb{V} = H_0^1(\Omega)$ , according to Lemma 2.2 (first Poincaré inequality), the bilinear form  $\mathcal{B}$  in (2.8) is coercive and continuous in view of (2.6) and  $c \geq 0$ . This framework applies as well to the *non-homogeneous Dirichlet* problem (2.10), upon extending  $g$  to a function  $g \in H^1(\Omega)$  and rewriting the problem in terms of  $w = u - g \in H_0^1(\Omega)$  and forcing  $\ell = f - L[g] \in H^{-1}(\Omega)$ .

On the other hand, the bilinear form  $\mathcal{B}$  in (2.13) associated with a *Robin* boundary condition is coercive provided  $p \geq p_0$  on  $\partial\Omega$  (or at least on an open subset of  $\partial\Omega$ ) with some constant  $p_0 > 0$ . This is a consequence of the norm equivalence

$$\|v\|_{H^1(\Omega)}^2 \approx |v|_{H_0^1(\Omega)}^2 + \|v\|_{L^2(\partial\Omega)}^2 \quad \forall v \in H^1(\Omega). \quad (2.31)$$

For the *Neumann* problem, instead, we have  $p = 0$  in  $\partial\Omega$ , whence  $\mathcal{B}$  is coercive whenever  $c > 0$  in  $\Omega$  (or at least in an open subset of  $\Omega$ ). If  $c = 0$  in  $\Omega$ , then  $\mathcal{B}$  is only coercive in the subspace of  $H^1(\Omega)$  of functions with vanishing meanvalue, according to Lemma 2.3 (second Poincaré inequality). Existence, uniqueness and stability is thus guaranteed by Corollary 2.10 (Lax-Milgram) provided the forcing in (2.13) satisfies the compatibility condition  $\ell(1) = 0$ .

*Saddle Point Problems.* We consider now an abstract problem a bit more general than (2.15), so the following results apply to the Stokes system (2.14).

Given a pair of Hilbert spaces  $(\mathbb{V}, \mathbb{Q})$ , we consider two continuous bilinear forms  $a: \mathbb{V} \times \mathbb{V} \rightarrow \mathbb{R}$  and  $b: \mathbb{Q} \times \mathbb{V} \rightarrow \mathbb{R}$ . 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*

$$a[u, v] + b[p, v] = \langle f, v \rangle \quad \forall v \in \mathbb{V}, \quad (2.32a)$$

$$b[q, u] = \langle g, q \rangle \quad \forall q \in \mathbb{Q}. \quad (2.32b)$$

Problem (2.32) is variational and can be rewritten in the form (2.23)

$$(u, p) \in \mathbb{V} \times \mathbb{Q} : \quad \mathcal{B}[(u, p), (v, q)] = \langle f, v \rangle + \langle g, q \rangle \quad \forall (v, q) \in \mathbb{V} \times \mathbb{Q}, \quad (2.33)$$

where  $\mathcal{B}$  is the bilinear form

$$\mathcal{B}[(u, p), (v, q)] := a[u, v] + b[p, v] + b[q, u]. \quad (2.34)$$

Therefore, the saddle point problem (2.32) is well-posed if and only if  $\mathcal{B}$  satisfies the inf-sup condition (2.26). Since  $\mathcal{B}$  is defined via the bilinear forms  $a$  and  $b$ , and (2.32) has a degenerate structure, it is not that simple to show (2.26) directly.

However, the result is a consequence of the inf-sup theorem for saddle point problems given by Brezzi in 1974 [Brezzi 1974].

**Theorem 2.11 (Brezzi)** *The saddle point problem (2.32) 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*

$$\inf_{v \in \mathbb{V}_0} \sup_{w \in \mathbb{V}_0} \frac{a[v, w]}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{V}}} = \inf_{w \in \mathbb{V}_0} \sup_{v \in \mathbb{V}_0} \frac{a[v, w]}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{V}}} = \alpha > 0, \quad (2.35a)$$

$$\inf_{q \in \mathbb{Q}} \sup_{v \in \mathbb{V}} \frac{b[q, v]}{\|q\|_{\mathbb{Q}} \|v\|_{\mathbb{V}}} = \beta > 0, \quad (2.35b)$$

where

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

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

$$(\|u\|_{\mathbb{V}}^2 + \|p\|_{\mathbb{Q}}^2)^{1/2} \leq \gamma (\|f\|_{\mathbb{V}^*}^2 + \|g\|_{\mathbb{Q}^*}^2)^{1/2}. \quad (2.36)$$

Combining Theorem 2.11 (Brezzi) with Corollary 2.9 (well posedness implies inf-up) we infer the inf-sup condition for the bilinear form  $\mathcal{B}$  in (2.34).

**Corollary 2.12 (inf-sup of  $\mathcal{B}$ )** *Let the bilinear form  $\mathcal{B}: \mathbb{W} \rightarrow \mathbb{W}$  be defined by (2.34). Then there holds*

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

where  $\gamma$  is the stability constant from Theorem 2.11.

Assume that  $a: \mathbb{V} \times \mathbb{V} \rightarrow \mathbb{R}$  is symmetric and let  $(u, p)$  be the solution to (2.32). 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

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

*Stokes system.* Theorem 2.11 (Brezzi) applies to the Stokes system (2.14) and (2.15) once we verify the inf-sup property (2.35b) for the bilinear form  $b[q, v] = -\int_{\Omega} q \operatorname{div} v$ . This turns out to be equivalent to the following problem: for any  $q \in L_0^2(\Omega)$  there exists a  $w \in H_0^1(\Omega; \mathbb{R}^d)$  such that

$$-\operatorname{div} w = q \quad \text{in } \Omega \quad \text{and} \quad \|w\|_{H^1(\Omega; \mathbb{R}^d)} \leq C(\Omega) \|q\|_{L^2(\Omega)}. \quad (2.37)$$

This non-trivial result goes back to Nečas [Carroll, Duff, Friberg, Gobert, Grisvard, Nečas and Seeley 1966] and a proof can for instance be found in [Galdi 1994,

Theorem III.3.1]. This implies

$$\sup_{\mathbf{v} \in H_0^1(\Omega; \mathbb{R}^d)} \frac{b[q, \mathbf{v}]}{|\mathbf{v}|_{H^1(\Omega; \mathbb{R}^d)}} \geq \frac{b[q, \mathbf{w}]}{|\mathbf{w}|_{H^1(\Omega; \mathbb{R}^d)}} = \frac{\|q\|_{L^2(\Omega)}^2}{|\mathbf{w}|_{H^1(\Omega; \mathbb{R}^d)}} \geq C(\Omega)^{-1} \|q\|_{L^2(\Omega)}.$$

Therefore, (2.35b) holds with  $\beta \geq C(\Omega)^{-1}$ .

The inf-sup condition is satisfied also for the spaces  $\mathbb{V}$  and  $\mathbb{Q}$  defined in (2.18), which are appropriate for the weak formulation (2.17) of the Neumann boundary value problem. Indeed, given any  $q \in L^2(\Omega)$ , we can split it as  $q = (q - \hat{q}) + \hat{q}$  with  $\hat{q} := |\Omega|^{-1} \int_{\Omega} q$ . Let  $\mathbf{w}_0 \in H_0^1(\Omega; \mathbb{R}^d)$  be the function defined as in (2.37) with  $q$  replaced by  $q - \hat{q}$ , and let  $\hat{\mathbf{w}} = \frac{1}{2}(\hat{q}x, \hat{q}y)$ . Then, it is easily checked that the function  $\mathbf{w} = \bar{\mathbf{w}} - |\Omega|^{-1} \int_{\Omega} \bar{\mathbf{w}}$  with  $\bar{\mathbf{w}} = \mathbf{w}_0 + \hat{\mathbf{w}}$  belongs to  $\mathbb{V}$  and satisfies

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

## 2.5 $W_p^1$ -regularity of reaction-diffusion equation

It will be useful later in Lemma 5.20 to know whether  $L^\infty$ -coefficients  $(A, c)$  allow for enhanced regularity beyond  $H_0^1(\Omega)$  for solutions  $u$  of (2.7). We can reformulate this question as an extension of the Lax-Milgram theory that states that the solution operator  $L^{-1}$  of (2.5) is an isomorphism between  $H^{-1}(\Omega)$  and  $H_0^1(\Omega)$ ; see Corollary 2.10 (Lax-Milgram).

This issue is well understood for the Laplace operator, i.e.  $A = I$  and  $c = 0$ . It is known that for Lipschitz domains  $\Omega \subset \mathbb{R}^d$ , there exists  $p_0 = p_0(\Omega) > 2$  such that

$$\|\nabla u\|_{L^p(\Omega)} \leq K \|f\|_{W_p^{-1}(\Omega)} \quad \forall p \in [2, p_0], \quad (2.38)$$

where  $K$  depends on  $p$  (see for example [Jerison and Kenig 1995]); in particular,  $p_0 > 4$  for  $d = 2$  and  $p_0 > 3$  for  $d = 3$ . Hereafter,  $W_p^{-1}(\Omega)$  denotes the dual space of  $\dot{W}_q^1(\Omega)$  – functions in  $W_q^1(\Omega)$  with zero trace and  $q = \frac{p}{p-1}$ . For  $A \in L^\infty(\Omega, \mathbb{R}^{d \times d})$  and  $c = 0$ , estimate (2.38) was first derived by [Meyers 1963] as a perturbation result for the Laplacian; see also [Brenner and Scott 2008]. We now present a simple proof for  $p > 2$  following [Bonito, DeVore and Nochetto 2013b]. Let

$$\theta(p) := \frac{1/2 - 1/p}{1/2 - 1/p_0} \quad \forall p \in [2, p_0], \quad (2.39)$$

and note that  $\theta(p)$  increases strictly from 0 at  $p = 2$  to 1 at  $p = p_0$ . Let  $K_0$  be the constant  $K$  in (2.38) for  $p = p_0$  and, for any  $t \in (0, 1)$ , define

$$p_*(t) := \max\{p \in [2, p_0] : K_0^{\theta(p)}(1-t) \leq 1\}. \quad (2.40)$$

**Lemma 2.13 ( $W_p^1$ -regularity)** *Let  $A \in L^\infty(\Omega, \mathbb{R}^{d \times d})$  satisfy (2.6) with  $\alpha_1 \leq \alpha_2$ ,  $c \in L^\infty(\Omega)$  be non-negative, and let  $\Omega$  be Lipschitz. If  $f \in W_p^{-1}(\Omega)$  for some  $p \in [2, \min\{p_*(\frac{\alpha_1}{\alpha_2}), \frac{2d}{d-2}\})$ , then the solution  $u \in H_0^1(\Omega)$  of (2.7) satisfies*

$$\|\nabla u\|_{L^p(\Omega)} \leq C(p) \left(1 + c_{\mathcal{B}}^{-1} C(\Omega) \|c\|_{L^\infty(\Omega)}\right) \|f\|_{W_p^{-1}(\Omega)} \quad (2.41)$$

with constant

$$C(p) = \frac{1}{\alpha_2} \frac{K_0^{\theta(p)}}{1 - K_0^{\theta(p)} \left(1 - \frac{\alpha_1}{\alpha_2}\right)}$$

and  $C(\Omega) = C_S C_P$  where  $C_S$  is the constant in Lemma 2.1 (Sobolev embedding) and  $C_P$  is the constant in Lemma 2.2 (first Poincaré inequality).

*Proof.* We first consider the principal part of the operator  $L$  in (2.5), namely we take  $c = 0$ . In fact, let the operator  $S : \dot{W}_p^1(\Omega) \rightarrow W_p^{-1}(\Omega)$  be defined by  $Sv := -\operatorname{div} \left( \frac{1}{\alpha_2} \mathbf{A} \nabla v \right)$ . In order to prove (2.41) for  $S$ , we resort to a perturbation argument for the Laplace operator  $Tv := -\Delta v$ .

The first task is to bound  $K$  in (2.38) in terms of  $K_0$  and  $p$ . To this end, we recall that the operator  $T : H_0^1(\Omega) \rightarrow H^{-1}(\Omega)$  is an isomorphism with norm  $\|T^{-1}\|_2 = 1$ . Moreover,  $T : \dot{W}_{p_0}^1(\Omega) \rightarrow W_{p_0}^{-1}(\Omega)$  is also an isomorphism with norm  $\|T^{-1}\|_{p_0} = K_0$  according to (2.38) for  $p = p_0$ , provided we adopt the norm  $\|\nabla \cdot\|_{L^p(\Omega)}$  in  $\dot{W}_p^1(\Omega)$ . By the real method of interpolation, we know that  $\dot{W}_p^1(\Omega) = [H_0^1(\Omega), \dot{W}_{p_0}^1(\Omega)]_{\theta(p), p}$  is the interpolation space between  $H_0^1(\Omega)$  and  $\dot{W}_{p_0}^1(\Omega)$  with parameter  $\theta(p)$  given by (2.39). Hence, operator interpolation theory implies that  $T : \dot{W}_p^1(\Omega) \rightarrow W_p^{-1}(\Omega)$  is an isomorphism with

$$K = \|T^{-1}\|_p = K_0^{\theta(p)}.$$

We regard  $S$  as a perturbation of  $T$ , define the operator  $Q := T - S : \dot{W}_p^1(\Omega) \rightarrow W_p^{-1}(\Omega)$ , and observe that  $\|Q\|_p \leq 1 - \frac{\alpha_1}{\alpha_2}$  because

$$\langle Qv, w \rangle = \int_{\Omega} \left( I - \frac{1}{\alpha_2} \mathbf{A} \right) \nabla v \nabla w \leq \left( 1 - \frac{\alpha_1}{\alpha_2} \right) \|\nabla v\|_{L^p(\Omega)} \|\nabla w\|_{L^q(\Omega)} \quad w \in \dot{W}_q^1(\Omega).$$

Therefore, the operator  $T^{-1}Q : \dot{W}_p^1(\Omega) \rightarrow \dot{W}_p^1(\Omega)$  satisfies

$$\|T^{-1}Q\|_p \leq \|T^{-1}\|_p \|Q\|_p \leq K_0^{\theta(p)} \left( 1 - \frac{\alpha_1}{\alpha_2} \right)$$

as well as  $\|T^{-1}Q\|_p < 1$  for any  $p \in [2, p_*(\alpha_1/\alpha_2))$  in view of definition (2.40) of  $p_*(t)$ . We conclude by the Neumann Theorem that the operator  $S = T(I - T^{-1}Q) : \dot{W}_p^1(\Omega) \rightarrow W_p^{-1}(\Omega)$  is invertible and its norm is bounded by

$$\|S^{-1}\|_p \leq \|T^{-1}\|_p \|I - T^{-1}Q\|_p \leq \frac{\|T^{-1}\|_p}{1 - \|T^{-1}Q\|_p} \leq \frac{K_0^{\theta(p)}}{1 - K_0^{\theta(p)} \left( 1 - \alpha_1/\alpha_2 \right)}.$$

This yields the asserted estimate for  $S = -\operatorname{div} \left( \frac{1}{\alpha_2} \mathbf{A} \nabla v \right)$ .

Finally, we consider the operator  $L$  in (2.5) with  $c \neq 0$ . If  $u \in H_0^1(\Omega)$  is the solution of (2.7) given by Corollary 2.10 (Lax-Milgram), rewrite (2.7) as follows

$$Su = -\operatorname{div} \left( \frac{1}{\alpha_2} \mathbf{A} \nabla u \right) = \frac{1}{\alpha_2} (f - cu) = \frac{1}{\alpha_2} g,$$

and apply the preceding estimate for  $S$  to infer that

$$\|S^{-1}g\|_{\dot{W}_p^1(\Omega)} = \|\nabla u\|_{L^p(\Omega)} \leq C(p)\|g\|_{W_p^{-1}(\Omega)} \leq C(p)(\|f\|_{W_p^{-1}(\Omega)} + \|cu\|_{W_p^{-1}(\Omega)}).$$

It remains to estimate the last term on the right-hand side. Using Cauchy-Schwarz in conjunction with Lemma 2.2 (first Poincaré inequality), i.e.,  $\|u\|_{L^2(\Omega)} \leq C_P\|\nabla u\|_{L^2(\Omega)}$ , and Lemma 2.1 (Sobolev embeddings), i.e.,  $\|w\|_{L^2(\Omega)} \leq C\|w\|_{\dot{W}_q^1(\Omega)} \leq C_S\|\nabla w\|_{L^q(\Omega)}$  for  $q = p/(p-1) \geq 2d/(d+2)$ , we see that

$$\langle cu, w \rangle \leq C_P C_S \|c\|_{L^\infty(\Omega)} \|\nabla u\|_{L^2(\Omega)} \|\nabla w\|_{L^q(\Omega)} \leq \frac{C_P C_S}{c_{\mathcal{B}}} \|c\|_{L^\infty(\Omega)} \|f\|_{H^{-1}(\Omega)} \|\nabla w\|_{L^q(\Omega)}$$

because of the stability estimate (2.27) with constant  $\alpha = c_{\mathcal{B}}$ . Since  $\|f\|_{H^{-1}(\Omega)} \leq |\Omega|^{\frac{p-2}{2p}} \|f\|_{W_p^{-1}(\Omega)}$ , the asserted estimate (2.41) for  $c \neq 0$  follows immediately.  $\square$

### 3 A Priori Approximation Theory

We devote this section to discussing basic concepts about piecewise polynomial approximation in Sobolev spaces over graded meshes in any dimension  $d$ . We start by introducing the Petrov-Galerkin method in an abstract setting (Sect. 3.1); this motivates our interest in approximation results in Sobolev spaces. Hence, we briefly discuss the construction of finite element spaces in Sect. 3.2, along with polynomial interpolation of functions in Sobolev spaces in Sect. 3.3. This provides local estimates adequate for the comparison of quasi-uniform and graded meshes for  $d > 1$ . We exploit them in developing the so-called error equidistribution principle in Sect. 3.4 and the construction of suitably graded meshes via a greedy algorithm in Sect. 3.6. We conclude that graded meshes can deliver optimal interpolation rates for certain classes of singular functions, and thus supersede quasi-uniform refinement.

In the second part of the section, we explore the geometric aspects of mesh refinement for conforming meshes in Sect. 3.5 and nonconforming meshes in Sect. 3.7, but postpone a full and rather technical discussion until Sect. 8. We include a statement about complexity of the refinement procedure, which turns out to be instrumental later and will be proved in Sect. 8.

#### 3.1 The Galerkin method: best approximation

In order to render the variational problem (2.23) computable, we let  $\mathbb{V}_N \subset \mathbb{V}$  and  $\mathbb{W}_N \subset \mathbb{W}$  be subspaces with the same dimension  $N < \infty$  and consider the Petrov-Galerkin approximation:

$$u_N \in \mathbb{V}_N : \quad \mathcal{B}[u_N, w] = \langle f, w \rangle \quad \forall w \in \mathbb{W}_N. \quad (3.1)$$

If  $\mathbb{V}_N = \mathbb{W}_N$  this is called Galerkin approximation. Since (3.1) is a square algebraic system, existence and uniqueness of  $u_N \in \mathbb{V}_N$  is equivalent to the kernel of the corresponding linear discrete operator to be trivial. This leads to the following

equivalent conditions for unique solvability [Nečas 1962, Babuška 1971], see also [Nochetto *et al.* 2009, Proposition 1].

**Lemma 3.1 (discrete inf-sup condition)** *The following statements are equivalent:*

1. for every  $0 \neq v \in \mathbb{V}_N$  there exists  $w \in \mathbb{W}_N$  such that  $\mathcal{B}[v, w] \neq 0$ ;
2. for every  $0 \neq w \in \mathbb{W}_N$  there exists  $v \in \mathbb{V}_N$  such that  $\mathcal{B}[v, w] \neq 0$ ;
3. the following discrete inf-sup condition holds with a constant  $\beta_N > 0$

$$\inf_{v \in \mathbb{V}_N} \sup_{w \in \mathbb{W}_N} \frac{\mathcal{B}[v, w]}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}} = \inf_{w \in \mathbb{W}_N} \sup_{v \in \mathbb{V}_N} \frac{\mathcal{B}[v, w]}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}} = \beta_N; \quad (3.2)$$

4.  $\inf_{v \in \mathbb{V}_N} \sup_{w \in \mathbb{W}_N} \frac{\mathcal{B}[v, w]}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}} > 0$ ;
5.  $\inf_{w \in \mathbb{W}_N} \sup_{v \in \mathbb{V}_N} \frac{\mathcal{B}[v, w]}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}} > 0$ .

This is a discrete version of Theorem 2.8 (Nečas) and leads to the stability bound

$$\|u_N\|_{\mathbb{V}} \leq \frac{1}{\beta_N} \|f\|_{\mathbb{W}^*}. \quad (3.3)$$

Therefore,  $\beta_N^{-1}$  acts as a stability constant for (3.1) and is thus desirable that it is uniformly bounded below away from zero. This is always the case for *coercive* problems because (2.28) is inherited within the subspaces  $\mathbb{V}_N = \mathbb{W}_N \subset \mathbb{V}$ , whence  $\beta_N \geq \alpha > 0$ . In contrast, a uniform lower bound for *saddle point* problems

$$\beta_N \geq \beta > 0, \quad (3.4)$$

requires compatibility between the subspaces  $\mathbb{V}_N$  and  $\mathbb{W}_N$  [Boffi, Brezzi and Fortin 2013].

If we now subtract (3.1) from (2.23), we obtain *Galerkin orthogonality*

$$\mathcal{B}[u - u_N, w] = 0 \quad \forall w \in \mathbb{W}_N. \quad (3.5)$$

This relation is instrumental to derive the following best approximation property as well to develop a posteriori error estimates in Section 4.

**Proposition 3.2 (quasi-best-approximation property)** *Let  $\mathcal{B}: \mathbb{V} \times \mathbb{W} \rightarrow \mathbb{R}$  be continuous and satisfy (3.2). Then the error  $u - u_N$  satisfies the bound*

$$\|u - u_N\|_{\mathbb{V}} \leq \frac{\|\mathcal{B}\|}{\beta_N} \min_{v \in \mathbb{V}_N} \|u - v\|_{\mathbb{V}}. \quad (3.6)$$

*Proof.* We give a simplified proof, which follows Babuška [Babuška 1971, Babuška and Aziz 1972] and yields the constant  $1 + \frac{\|\mathcal{B}\|}{\beta_N}$ . The asserted constant is due to Xu and Zikatanov [Xu and Zikatanov 2003].

Combining (3.2), (3.5), and the continuity of  $\mathcal{B}$ , we derive for all  $v \in \mathbb{V}_N$

$$\beta_N \|u_N - v\|_{\mathbb{V}} \leq \sup_{w \in \mathbb{W}_N} \frac{\mathcal{B}[u_N - v, w]}{\|w\|_{\mathbb{W}}} = \sup_{w \in \mathbb{W}_N} \frac{\mathcal{B}[u - v, w]}{\|w\|_{\mathbb{W}}} \leq \|\mathcal{B}\| \|u - v\|_{\mathbb{V}},$$



whence

$$\|u_N - v\|_{\mathbb{V}} \leq \frac{\|\mathcal{B}\|}{\beta_N} \|u - v\|_{\mathbb{V}}.$$

Using the triangle inequality yields

$$\|u - u_N\|_{\mathbb{V}} \leq \|u - v\|_{\mathbb{V}} + \|v - u_N\|_{\mathbb{V}} \leq \left(1 + \frac{\|\mathcal{B}\|}{\beta_N}\right) \|u - v\|_{\mathbb{V}}$$

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

**Corollary 3.3 (quasi-monotonicity)** *Let  $\mathcal{B}: \mathbb{V} \times \mathbb{W} \rightarrow \mathbb{R}$  be continuous and satisfy (3.2). If  $\mathbb{V}_M$  is a subspace of  $\mathbb{V}_N$ , then for all  $v \in \mathbb{V}_M$*

$$\|u - u_N\|_{\mathbb{V}} \leq \frac{\|\mathcal{B}\|}{\beta_N} \|u - v\|_{\mathbb{V}}. \quad (3.7)$$

Moreover, if  $\mathbb{V} = \mathbb{W}$  and  $\mathcal{B}$  is symmetric and coercive with constants  $c_{\mathcal{B}} \leq C_{\mathcal{B}}$ , then for all  $v \in \mathbb{V}_M$

$$\|u - u_N\|_{\Omega} \leq \|u - v\|_{\Omega}, \quad \|u - u_N\|_{\mathbb{V}} \leq C_{\text{cea}} \|u - v\|_{\mathbb{V}}, \quad (3.8)$$

where  $C_{\text{cea}} := \sqrt{\frac{C_{\mathcal{B}}}{c_{\mathcal{B}}}}$ .

*Proof.* Inequality (3.7) is a consequence of the previous bound (3.6) upon taking  $v \in \mathbb{V}_M$  instead of  $\mathbb{V}_N$ . To show the left inequality in (3.8), we combine (2.28) and (3.5)

$$\|u - u_N\|_{\Omega}^2 = \mathcal{B}[u - u_N, u - v] \leq \|u - u_N\|_{\Omega} \|u - v\|_{\Omega}, \quad \forall v \in \mathbb{V}_M.$$

This together with the norm equivalence (2.30) gives the remaining inequality.  $\square$

The significance of (3.6) is that we need to construct discrete spaces  $\mathbb{V}_N$  with good approximation properties. We introduce next piecewise polynomial approximation, which gives rise to the finite element method.

### 3.2 Finite element spaces

In this section we focus on the construction of the discrete spaces  $\mathbb{V}_N$  and  $\mathbb{W}_N$ . We consider the bilinear forms  $\mathcal{B}$  introduced in Section 2.3 with emphasis on the diffusion-reaction case (2.8). We assume that  $\Omega$  is a bounded polyhedral domain  $\Omega \subset \mathbb{R}^d$  and is partitioned into a conforming or non-conforming mesh  $\mathcal{T}$  made of simplices  $T$ , which are assumed to be closed with non-overlapping interiors; thus,

$$\overline{\Omega} = \bigcup \{T : T \in \mathcal{T}\}.$$

The reference element is denoted

$$T_d := \{x = (x_1, \dots, x_d) \in \mathbb{R}^d \mid 0 \leq x_i \leq 1, i = 1, \dots, d, \sum_{i=1}^d x_i \leq 1\}.$$

We will discuss the construction of conforming meshes in Section 3.5 by the bisection method and that of non-conforming meshes (constrained to the fulfillment of an *admissibility condition*) in Section 3.7, both for  $d = 2$ . We will embark on a thorough discussion in Section 8. We assume for the moment that  $\mathcal{T}$  is an element of a (possibly infinite) class  $\mathbb{T}$  of conforming *shape regular* meshes. To define this geometric concept, we denote by  $\bar{h}_T$  the diameter of  $T \in \mathcal{T}$ , by  $\underline{h}_T$  the diameter of the largest ball contained in  $T$ , and impose the restriction

$$\sigma := \sup_{\mathcal{T} \in \mathbb{T}} \sup_{T \in \mathcal{T}} \frac{\bar{h}_T}{\underline{h}_T} < \infty. \quad (3.9)$$

The constant  $\sigma$  is refereed as the shape regularity constant of  $\mathbb{T}$ .

Given a shape regular mesh  $\mathcal{T} \in \mathbb{T}$ , we define the finite element space of *discontinuous* piecewise polynomials of total degree up to  $n \geq 1$

$$\mathbb{S}_{\mathcal{T}}^{n,-1} := \{v \in L^2(\Omega) \mid v|_T \in \mathbb{P}_n(T) \ \forall T \in \mathcal{T}\};$$

and its globally *continuous* counterpart

$$\mathbb{S}_{\mathcal{T}}^{n,0} := \mathbb{S}_{\mathcal{T}}^{n,-1} \cap C^0(\bar{\Omega}).$$

Note that  $\mathbb{S}_{\mathcal{T}}^{n,0} \subset H^1(\Omega)$  which makes it adequate for (2.7)-(2.8). We refer to Braess [Braess 2007], Brenner-Scott [Brenner and Scott 2008], Ciarlet [Ciarlet 2002] and Siebert [Siebert 2012] for a discussion on the local construction of this space (i.e. Lagrange elements of degree  $n \geq 1$ ) along with its properties. We denote by

$$\mathbb{V}_{\mathcal{T}} := \mathbb{S}_{\mathcal{T}}^{n,0} \cap H_0^1(\Omega) \quad (3.10)$$

the subspace of finite element functions which vanish on  $\partial\Omega$ . Note that we do not explicitly refer to the polynomial degree, which will be clear in each context.

We focus on the conforming piecewise linear case  $n = 1$  (Courant elements), but most results extend to non-conforming meshes or  $n > 1$ . In this vein, global continuity can be simply enforced by imposing continuity at the set  $\mathcal{V}$  of vertices  $z$  of  $\mathcal{T}$ , the so-called *nodal values*. However, the following local construction leads to global continuity. If  $T$  is a generic simplex of  $\mathcal{T}$ , namely the convex hull of  $\{z_i\}_{i=0}^d$ , then we associate to each vertex  $z_i$  a *barycentric coordinate*  $\lambda_i^T$ , which is the linear function in  $T$  with nodal value 1 at  $z_i$  and 0 at the other vertices of  $T$ . Upon pasting together the barycentric coordinates  $\lambda_z^T$  of all simplices  $T$  containing the vertex  $z \in \mathcal{V}$ , we obtain a continuous piecewise linear function  $\phi_z \in \mathbb{S}_{\mathcal{T}}^{1,0}$  as depicted in Figure 3.1 for  $d = 2$ :

The set  $\{\phi_z\}_{z \in \mathcal{V}}$  of all such functions is the nodal basis of  $\mathbb{S}_{\mathcal{T}}^{1,0}$ , or Courant basis. We denote by  $\omega_z := \text{supp}(\phi_z)$  the support of  $\phi_z$ , from now on called *star* associated to  $z$ , and by  $\gamma_z$  the interior skeleton of  $\omega_z$ , namely the union of all the faces containing  $z$ .

In view of the definition of  $\phi_z$ , we have the following unique representation of

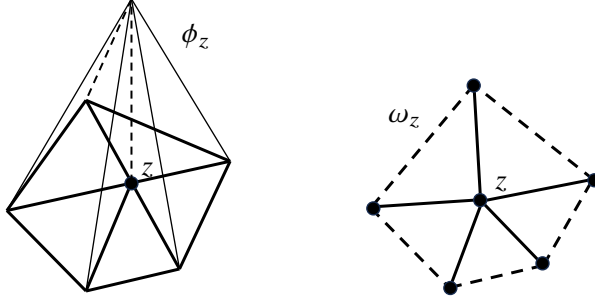


Figure 3.1. (Left) Piecewise linear basis function  $\phi_z$  corresponding to interior node  $z$ ; (Right) Support  $\omega_z$  of  $\phi_z$  and skeleton  $\gamma_z$  (in solid line)

any function  $v \in \mathbb{S}^{1,0}(\mathcal{T})$

$$v(x) = \sum_{z \in \mathcal{V}} v(z) \phi_z(x).$$

The functions  $\phi_z$  are non-negative and satisfy the *partition of unity* property

$$\sum_{z \in \mathcal{V}} \phi_z(x) = 1 \quad \forall x \in \Omega. \quad (3.11)$$

If we further impose  $v(z) = 0$  for all  $z \in \partial\Omega \cap \mathcal{V}$ , then  $v \in H_0^1(\Omega)$ .

For each simplex  $T \in \mathcal{T}$ , generated by vertices  $\{z_i\}_{i=0}^d$ , the *dual functions*  $\{\lambda_i^*\}_{i=0}^d \subset \mathbb{P}_1(T)$  to the barycentric coordinates  $\{\lambda_i\}_{i=0}^d$  satisfy the bi-orthogonality relation  $\int_T \lambda_i^* \lambda_j = \delta_{ij}$ , and are given by

$$\lambda_i^* = \frac{(1+d)^2}{|T|} \lambda_i - \frac{1+d}{|T|} \sum_{j \neq i} \lambda_j \quad \forall 0 \leq i \leq d.$$

The *Courant dual basis*  $\phi_z^* \in \mathbb{S}^{1,-1}(\mathcal{T})$  are the functions over  $\mathcal{T}$  given by

$$\phi_z^* = \frac{1}{V_z} \sum_{T \ni z} (\lambda_z^T)^* \chi_T \quad \forall z \in \mathcal{V},$$

where  $V_z \in \mathbb{N}$  is the valence of  $z$  (number of elements of  $\mathcal{T}$  containing  $z$ ) and  $\chi_T$  is the characteristic function of  $T$ . These functions have the same support  $\omega_z$  as the nodal basis  $\phi_z$  and satisfy the global bi-orthogonality relation

$$\int_{\Omega} \phi_z^* \phi_y = \delta_{zy} \quad \forall z, y \in \mathcal{V}.$$

Finally, we denote by  $\mathcal{N}$  the Lagrange nodes of order  $n$  of a mesh  $\mathcal{T}$ , and by  $\psi_z \in \mathbb{S}_{\mathcal{T}}^{n,0}$  the corresponding Lagrange basis of  $\mathbb{S}_{\mathcal{T}}^{n,0}$ ; hence  $\mathbb{S}_{\mathcal{T}}^{n,0} = \text{span}\{\psi_z\}_{z \in \mathcal{N}}$ .

### 3.3 Polynomial interpolation in Sobolev spaces

We wish to use the space  $\mathbb{V}_{\mathcal{T}}$  defined in (3.10) as the discrete space  $\mathbb{V}_N$  in the Galerkin method (3.1). If the bilinear form  $\mathcal{B}$  satisfies an inf-sup condition with constant  $\beta_{\mathcal{T}} > 0$ , we find a discrete solution  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}$  which satisfies the error bound (3.6), i.e.

$$\|u - u_{\mathcal{T}}\|_{\mathbb{V}} \leq \frac{\|\mathcal{B}\|}{\beta_{\mathcal{T}}} \min_{v \in \mathbb{V}_{\mathcal{T}}} \|u - v\|_{\mathbb{V}}.$$

In turn, the minimum on the right-hand side can be bounded from above by the quantity  $\|u - v\|_{\mathbb{V}}$  for any chosen  $v \in \mathbb{V}_{\mathcal{T}}$ . This motivates the search for quasi-best approximations of  $u$  in the norm of  $\mathbb{V}$ . One classical tool to generate approximations to any given function is *interpolation*. Interpolation in  $\mathbb{V}_{\mathcal{T}}$  is discussed next.

If  $v \in C^0(\bar{\Omega})$  we define the *Lagrange interpolant*  $I_{\mathcal{T}}v$  of  $v$  as follows:

$$I_{\mathcal{T}}v(x) = \sum_{z \in \mathcal{N}} v(z) \psi_z(x),$$

and note that  $I_{\mathcal{T}}v = v$  for all  $v \in \mathbb{S}_{\mathcal{T}}^{n,0}$  (i.e.  $I_{\mathcal{T}}$  is invariant in  $\mathbb{S}_{\mathcal{T}}^{n,0}$ ). For functions without point values, such as functions in  $H^1(\Omega)$  for  $d > 1$ , we need to determine nodal values by averaging. For any conforming mesh  $\mathcal{T} \in \mathbb{T}$ , the averaging process extends beyond nodes and so gives rise to the discrete neighborhood

$$\omega_{\mathcal{T}}(T) := \bigcup_{\substack{T' \in \mathcal{T} \\ T' \cap T \neq \emptyset}} T'$$

for each element  $T \in \mathcal{T}$  along with the *local quasi-uniformity* properties

$$\max_{T \in \mathcal{T}} \#\omega_{\mathcal{T}}(T) \leq C(\sigma), \quad \max_{T' \subset \omega_{\mathcal{T}}(T)} \frac{|T|}{|T'|} \leq C(\sigma),$$

where  $\sigma$  is the shape regularity coefficient defined in (3.9). We will often write  $\omega_T$  if there is no confusion about the underlying mesh  $\mathcal{T}$ . We shall also need the a smaller subset, namely the set of elements sharing a face with a given element  $T$ :

$$\tilde{\omega}_T := \tilde{\omega}_{\mathcal{T}}(T) := \bigcup_{\substack{T' \in \mathcal{T} \\ T' \cap T \in \mathcal{F}}} T', \quad (3.12)$$

where  $\mathcal{F}$  is the set of all  $(d-1)$ -dimensional faces of the mesh  $\mathcal{T}$ .

We introduce now one such operator  $I_{\mathcal{T}}$  due to Scott-Zhang [Brenner and Scott 2008, Scott and Zhang 1990], from now on called a *quasi-interpolation operator*. We focus on polynomial degree  $n = 1$ , but the construction is valid for any  $n$ ; see [Brenner and Scott 2008, Scott and Zhang 1990] for details. We recall that  $\{\phi_z\}_{z \in \mathcal{V}}$  is the global Lagrange basis of  $\mathbb{S}_{\mathcal{T}}^{1,0}$ ,  $\{\phi_z^*\}_{z \in \mathcal{V}}$  is the global dual basis,

and  $\text{supp } \phi_z^* = \text{supp } \phi_z$  for all  $z \in \mathcal{V}$ . We thus define  $I_{\mathcal{T}} : L^1(\Omega) \rightarrow \mathbb{S}_{\mathcal{T}}^{1,0}$  to be

$$I_{\mathcal{T}}v := \sum_{z \in \mathcal{V}} \langle v, \phi_z^* \rangle \phi_z, \quad (3.13)$$

If  $0 \leq s \leq 2$  (integer) is a regularity index and  $1 \leq p \leq \infty$  is an integrability index, then we would like to prove the *quasi-local error estimate*

$$|v - I_{\mathcal{T}}v|_{W_q^t(T)} \lesssim h_T^{\text{sob}(W_p^s) - \text{sob}(W_q^t)} |v|_{W_p^s(\omega_T)} \quad (3.14)$$

for all  $T \in \mathcal{T}$ , provided  $0 \leq t \leq s$ ,  $1 \leq q \leq \infty$  are such that  $\text{sob}(W_p^s) > \text{sob}(W_q^t)$ .

We first recall that  $I_{\mathcal{T}}$  is invariant in  $\mathbb{S}_{\mathcal{T}}^{1,0}$ , namely,

$$I_{\mathcal{T}}w = w \quad \forall w \in \mathbb{S}_{\mathcal{T}}^{1,0}.$$

Since the averaging process giving rise to the values of  $I_{\mathcal{T}}v$  for each element  $T \in \mathcal{T}$  takes place in the neighborhood  $\omega_T$ , we also deduce the local invariance

$$I_{\mathcal{T}}w|_T = w \quad \forall w \in \mathbb{P}_1(\omega_T)$$

as well as the local stability estimate for any  $1 \leq q \leq \infty$

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

We may thus write

$$v - I_{\mathcal{T}}v|_T = (v - w) - I_{\mathcal{T}}(v - w)|_T \quad \forall T \in \mathcal{T},$$

where  $w \in \mathbb{P}_{s-1}$  is arbitrary ( $w = 0$  if  $s = 0$ ). It suffices now to prove (3.14) in the reference element  $\widehat{T}$  and scale back and forth to  $T$ ; the definition (2.1) of Sobolev number accounts precisely for this scaling. We keep the notation  $T$  for  $\widehat{T}$ , apply the inverse estimate for linear polynomials  $|I_{\mathcal{T}}v|_{W_q^t(T)} \lesssim \|I_{\mathcal{T}}v\|_{L^q(T)}$  to  $v - w$  instead of  $v$ , and use the above local stability estimate, to infer that

$$|v - I_{\mathcal{T}}v|_{W_q^t(T)} \lesssim \|v - w\|_{W_q^t(\omega_T)} \lesssim \|v - w\|_{W_p^s(\omega_T)}.$$

The last inequality is a consequence of the inclusion  $W_p^s(\omega_T) \subset W_q^t(\omega_T)$  because  $\text{sob}(W_p^s) > \text{sob}(W_q^t)$  and  $t \leq s$ . Estimate (3.14) now follows from the Bramble-Hilbert lemma [Brenner and Scott 2008, Lemma 4.3.8], [Ciarlet 2002, Theorem 3.1.1], [Dupont and Scott 1980] or Proposition 6.34 below:

$$\inf_{w \in \mathbb{P}_{s-1}(\omega_T)} \|v - w\|_{W_p^s(\omega_T)} \lesssim |v|_{W_p^s(\omega_T)}. \quad (3.15)$$

This proves (3.14) for  $n = 1$ . The construction of  $I_{\mathcal{T}}$  and ensuing estimate (3.14) are still valid for any  $n > 1$  [Brenner and Scott 2008, Scott and Zhang 1990].

**Proposition 3.4 (quasi-interpolant without boundary values)** *Let  $s, t$  be regularity indices with  $0 \leq t \leq s \leq n + 1$ , and  $1 \leq p, q \leq \infty$  be integrability indices so that  $\text{sob}(W_p^s) > \text{sob}(W_q^t)$ . Then there exists a quasi-interpolation operator*

$I_{\mathcal{T}} : L^1(\Omega) \rightarrow \mathbb{S}_{\mathcal{T}}^{n,0}$ , which is invariant in  $\mathbb{S}_{\mathcal{T}}^{n,0}$  and satisfies

$$|v - I_{\mathcal{T}}v|_{W_q^t(T)} \lesssim h_T^{\text{sob}(W_p^s) - \text{sob}(W_q^t)} |v|_{W_p^s(\omega_T)} \quad \forall T \in \mathcal{T}. \quad (3.16)$$

The hidden constant in (3.14) depends on the shape coefficient of  $\mathcal{T}_0$  and  $d$ .

To impose a vanishing trace on  $I_{\mathcal{T}}v$  we may suitably modify the averaging process for boundary nodes. We thus define a set of dual functions with respect to an  $L^2$ -scalar product over  $(d-1)$ -subsimplices contained on  $\partial\Omega$ ; see again [Brenner and Scott 2008, Scott and Zhang 1990] for details. This retains the invariance property of  $I_{\mathcal{T}}$  on  $\mathbb{S}_{\mathcal{T}}^{n,0}(\mathcal{T})$  and guarantees that  $I_{\mathcal{T}}v$  has a zero trace if  $v \in W_1^1(\Omega)$  does. Hence, the above argument applies and (3.16) follows provided  $s \geq 1$ .

**Proposition 3.5 (quasi-interpolant with boundary values)** *Let  $s, t, p, q$  be as in Proposition 3.4. There exists a quasi-interpolation operator  $I_{\mathcal{T}} : W_1^1(\Omega) \rightarrow \mathbb{S}_{\mathcal{T}}^{n,0}$  invariant in  $\mathbb{S}_{\mathcal{T}}^{n,0}$  which satisfies (3.16) for  $s \geq 1$  and preserves the boundary values of  $v$  provided they are piecewise polynomial of degree  $\leq n$ . In particular, if  $v \in W_1^1(\Omega)$  has a vanishing trace on  $\partial\Omega$ , then so does  $I_{\mathcal{T}}v$ .*

**Remark 3.6 (fractional regularity)** We observe that (3.16) does not require the regularity indices  $t$  and  $s$  to be integer. The proof follows along the same lines but replaces the polynomial degree  $n$  by the greatest integer smaller than  $s$ ; the generalization of (3.15) can be taken from [Dupont and Scott 1980].

**Remark 3.7 (local error estimate for Lagrange interpolant)** Let the regularity index  $s$  and integrability index  $1 \leq p \leq \infty$  satisfy  $s - d/p > 0$ . This implies that  $\text{sob}(W_p^s) > \text{sob}(L^\infty)$ , whence  $W_p^s(\Omega) \subset C(\overline{\Omega})$  and the Lagrange interpolation operator  $I_{\mathcal{T}} : W_p^s(\Omega) \rightarrow \mathbb{S}_{\mathcal{T}}^{n,0}$  is well defined and satisfies the *local error estimate*

$$|v - I_{\mathcal{T}}v|_{W_q^t(T)} \lesssim h_T^{\text{sob}(W_p^s) - \text{sob}(W_q^t)} |v|_{W_p^s(T)}, \quad (3.17)$$

provided  $0 \leq t \leq s$ ,  $1 \leq q \leq \infty$  are such that  $\text{sob}(W_p^s) > \text{sob}(W_q^t)$ . We point out that  $\omega_T$  in (3.14) is now replaced by  $T$  in (3.17). We also remark that if  $v$  vanishes on  $\partial\Omega$  so does  $I_{\mathcal{T}}v$ . The proof of (3.17) proceeds along the same lines as that of Proposition 3.4 except that the nodal evaluation does not extend beyond the element  $T \in \mathcal{T}$  and the inverse and stability estimates over the reference element are replaced by

$$|I_{\mathcal{T}}v|_{W_q^t(\widehat{T})} \lesssim \|I_{\mathcal{T}}v\|_{L^q(\widehat{T})} \lesssim \|v\|_{L^\infty(\widehat{T})} \lesssim \|v\|_{W_p^s(\widehat{T})}.$$

The following global interpolation error estimate builds on Proposition 3.4 and relates to Fig. 2.1 ( DeVore diagram).

**Theorem 3.8 (global interpolation error estimate)** *Let  $1 \leq s \leq n+1$ ,  $t = 0, 1$ ,  $t < s$ , and  $1 \leq p \leq q$  satisfy  $r := \text{sob}(W_p^s) - \text{sob}(W_q^t) > 0$ . If  $v \in W_p^s(\Omega)$ , then*

$$|v - I_{\mathcal{T}}v|_{W_q^t(\Omega)} \lesssim \left( \sum_{T \in \mathcal{T}} h_T^{rp} |v|_{W_p^s(\omega_T)}^p \right)^{\frac{1}{p}}. \quad (3.18)$$

*Proof.* Use Proposition 3.4 along with the elementary property of series  $\sum_n a_n \leq (\sum_n a_n^{p/q})^{q/p}$  for  $0 < p/q \leq 1$ .  $\square$

*Continuous vs discontinuous approximation of gradients.* The preceding discussion might induce us to believe that when dealing with Sobolev functions  $v \in W_p^1(\Omega)$  without pointvalues, namely  $1 \leq p \leq d$ , global continuity of the quasi-interpolant  $I_{\mathcal{T}}v$  might degrade the approximation quality relative to discontinuous approximations. The following instrumental result shows that this is not the case [Veeseer 2016, Theorem 2]. It hinges on a new geometric concept: we say that a star  $\omega_z$  is  $(d-1)$ -faced connected if for any element  $T \subset \omega_z$  and  $(d-1)$ -face  $F \subset \omega_z$  containing  $z$  there is a sequence  $\{T_i\}_{i=1}^m$  such that

- any  $T_i$  is an element of  $\omega_z$  for  $0 \leq i \leq m$ ;
- any intersection  $T_i \cap T_{i+1}$  is a  $(d-1)$ -face of  $\omega_z$  for  $i = 0 \leq i \leq m-1$ ;
- $T_0$  contains  $F$  and  $T_m = T$ .

Notice that a star  $\omega_z$  is  $(d-1)$ -faced connected if the set  $\omega_z \cap \Omega$  is connected.

**Proposition 3.9 (approximation of gradients)** *Let  $v \in W_p^1(\Omega)$  for  $1 \leq p \leq d$ . Let  $\mathcal{T}$  be a conforming mesh such that its stars are  $(d-1)$ -face-connected. Then there exists a constant  $C(\sigma)$  depending on the shape regularity coefficient  $\sigma$  of (3.9), the dimension  $d$  and the polynomial degree  $n \geq 1$ , such that*

$$1 \leq \frac{\min_{w \in \mathbb{S}_{\mathcal{T}}^{n,0}} \|\nabla(v-w)\|_{L^p(\Omega)}}{\min_{w \in \mathbb{S}_{\mathcal{T}}^{n,-1}} \|\nabla(v-w)\|_{L^p(\Omega;\mathcal{T})}} \leq C(\sigma), \quad (3.19)$$

where  $\|\nabla w\|_{L^p(\Omega;\mathcal{T})}$  stands for the broken norm over  $\mathcal{T}$ .

The left inequality in (3.19) is obvious because  $\mathbb{S}_{\mathcal{T}}^{n,0} \subset \mathbb{S}_{\mathcal{T}}^{n,-1}$ . In contrast, the right inequality is delicate and relies on examining the quasi-interpolant (3.13) [Veeseer 2016, Theorems 1 and 22]. An important consequence of (3.19) is the following localized version of (3.18):

**Proposition 3.10 (localized quasi-interpolation estimate)** *Let  $1 < s \leq n+1$ ,  $1 \leq p \leq d$  and  $r = \text{sob}(W_p^s) - \text{sob}(W_q^1) > 0$ . If  $v \in W_q^1(\Omega)$ , then*

$$\|\nabla(v - I_{\mathcal{T}}v)\|_{L^q(\Omega)} \lesssim \left( \sum_{T \in \mathcal{T}} h_T^{pr} |v|_{W_p^s(T)}^p \right)^{\frac{1}{p}}, \quad (3.20)$$

*Proof.* Since  $v - I_{\mathcal{T}}v = (v-w) - I_{\mathcal{T}}(v-w)$  for any  $w \in \mathbb{S}_{\mathcal{T}}^{n,0}$ , combine Proposition 3.9 with Proposition 6.34 (Bramble-Hilbert for Sobolev spaces) to write

$$\|\nabla(v - I_{\mathcal{T}}v)\|_{L^q(\Omega)} \lesssim \min_{w \in \mathbb{S}_{\mathcal{T}}^{n,-1}} \|\nabla(v-w)\|_{L^q(\Omega)} \lesssim \left( \sum_{T \in \mathcal{T}} h_T^{pr} |v|_{W_p^s(T)}^p \right)^{\frac{1}{p}},$$

This concludes the proof.  $\square$

The crucial difference between (3.20) and (3.18) is that the function  $v \in W_q^1(\Omega)$

does not have to belong globally to  $W_p^s(\Omega)$  but rather locally, namely  $v \in W_p^s(T)$  for every  $T \in \mathcal{T}$ , to get optimal a priori error estimates. This property will find several applications later. A special case of (3.20) for  $p = q = 2$  and  $s = 2$  reads

$$|v - I_{\mathcal{T}}v|_{H^1(\Omega)}^2 \lesssim \sum_{T \in \mathcal{T}} h_T^2 |v|_{H^2(T)}^2,$$

for  $v \in H^2(\Omega; \mathcal{T}) := \{w \in H^1(\Omega) : w|_T \in H^2(T) \forall T \in \mathcal{T}\}$ .

*Quasi-Uniform Meshes.* We now apply Theorem 3.8 to *quasi-uniform* meshes, namely meshes  $\mathcal{T} \in \mathbb{T}$  for which all its elements are of comparable size  $h$ , whence

$$h \approx (\#\mathcal{T})^{-1/d} |\Omega|^{1/d} \approx (\#\mathcal{T})^{-1/d}.$$

**Corollary 3.11 (quasi-uniform meshes)** *Let  $1 < s \leq n + 1$  and  $v \in H^s(\Omega)$ . If  $\mathcal{T} \in \mathbb{T}$  is quasi-uniform, then*

$$\|\nabla(v - I_{\mathcal{T}}v)\|_{L^2(\Omega)} \lesssim |v|_{H^s(\Omega)} (\#\mathcal{T})^{-(s-1)/d}. \quad (3.21)$$

**Remark 3.12 (optimal rate)** If  $s = n + 1$ , and so  $v$  has the maximal regularity  $v \in H^{n+1}(\Omega)$ , then we obtain the optimal convergence rate in a linear Sobolev scale

$$\|\nabla(v - I_{\mathcal{T}}v)\|_{L^2(\Omega)} \lesssim |v|_{H^{n+1}(\Omega)} (\#\mathcal{T})^{-n/d}. \quad (3.22)$$

The order  $-n/d$  is just dictated by the polynomial degree  $n$  and cannot be improved upon assuming either higher regularity that  $H^{n+1}(\Omega)$  or a graded mesh  $\mathcal{T}$ . The presence of  $d$  in the exponent is referred to as *curse of dimensionality*.

*Example (corner singularity in 2d).* To explore the effect of a geometric singularity on (3.21), we let  $\Omega = (-1, 1)^2 \setminus [0, 1]^2$  be a L-shaped domain and  $v \in H^1(\Omega)$  be

$$v(r, \theta) = r^{\frac{2}{3}} \sin(2\theta/3) - r^2/4.$$

This function  $v \in H^1(\Omega)$  exhibits the typical corner singularity of the solution of  $-\Delta v = f$  with suitable Dirichlet boundary condition:  $v \in H^s(\Omega)$  for  $s < 5/3$ . Table 3.1 displays the best approximation error for polynomial degree  $n = 1, 2, 3$  and a sequence of *uniform* refinements in the seminorm  $|\cdot|_{H^1(\Omega)} = \|\nabla \cdot\|_{L^2(\Omega)}$ . This gives a *lower* bound for the interpolation error in (3.21).

Even though  $s$  is fractional, the error estimate (3.21) is still valid as stated in Remark 3.6. In fact, for uniform refinement, (3.21) can be derived by space interpolation between  $H^1(\Omega)$  and  $H^{n+1}(\Omega)$ . The asymptotic rate  $(\#\mathcal{T})^{-1/3}$  reported in Table 3.1 is consistent with (3.21) and independent of the polynomial degree  $n$ ; this shows that (3.21) is sharp. It is also suboptimal as compared with the optimal rate  $(\#\mathcal{T})^{-n/2}$  of Remark 3.12.

The question arises whether the rate  $(\#\mathcal{T})^{-1/3} \approx h^{2/3}$  in Table 3.1 is just a consequence of uniform refinement or unavoidable. It is important to realize that  $v \notin H^s(\Omega)$  for  $s \geq 5/3$  and thus (3.21) is not applicable. However, the problem is not that second-order derivatives of  $v$  do not exist but rather that they are not



$h$	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 3.1. Rate of convergence  $s$  in term of uniform mesh-size  $h$ . We observe an asymptotic error decay of about  $h^{2/3}$  (i.e.  $s = 2/3$ ), or equivalently  $(\#\mathcal{T})^{-1/3}$ , irrespective of the polynomial degree  $n$ . This provides a lower bound for  $\|v - I_{\mathcal{T}}v\|_{L^2(\Omega)}$  and thus shows that (3.21) is sharp.

square-integrable. In particular, it is true that  $v \in W_p^2(\Omega)$  if  $1 \leq p < 3/2$ . We therefore may apply Theorem 3.8 with, e.g.,  $n = 1$ ,  $s = 2$ , and  $p \in [1, 3/2)$  and then ask whether the structure of (3.18) can be exploited, e.g., by compensating the local singular behavior of  $v$  with the local mesh-size  $h$ . This enterprise naturally leads to *graded* meshes adapted to  $v$ .

#### 3.4 Principle of error equidistribution.

We investigate the relation between local interpolation error and regularity for the design of optimal graded meshes adapted to a given function  $v \in H^1(\Omega)$  for  $d = 2$ . We recall that  $W_1^2(\Omega)$  is in the same nonlinear Sobolev scale of  $H^1(\Omega)$ , namely  $\text{sob}(W_1^2) = \text{sob}(H^1)$ , but  $W_1^2(\Omega) \subset C^0(\overline{\Omega})$  [Brenner and Scott 2008, Lemma 4.3.4], and the *Lagrange interpolant*  $I_{\mathcal{T}}v$  is well defined and satisfies

$$\|\nabla(v - I_{\mathcal{T}}v)\|_{L^2(T)} \leq C|v|_{W_1^2(T)} =: e_{\mathcal{T}}(v, T) \quad \forall T \in \mathcal{T}. \quad (3.23)$$

We formulate a discrete minimization problem on the surrogate quantity  $\mathbf{e} := (e_{\mathcal{T}}(v, T))_{T \in \mathcal{T}} \in \mathbb{R}^N$  with  $N = \#\mathcal{T}$ : minimize the square of the total  $H^1$ -error  $E_{\mathcal{T}}(v)$

$$E_{\mathcal{T}}(v)^2 := \sum_{T \in \mathcal{T}} e_{\mathcal{T}}(v, T)^2$$

subject to the constraint

$$\sum_{T \in \mathcal{T}} e_{\mathcal{T}}(v, T) = C|v|_{W_1^2(\Omega)}.$$

We idealize the problem upon allowing  $e_{\mathcal{T}}(v, T)$  to attain any nonnegative real value despite the fact that shape regularity of  $\mathcal{T}$  entails geometric restrictions between adjacent elements. We next form the Lagrangian

$$\mathcal{L}[\mathbf{e}, \lambda] := \sum_{T \in \mathcal{T}} e_{\mathcal{T}}(v, T)^2 - \lambda \left( \sum_{T \in \mathcal{T}} e_{\mathcal{T}}(v, T) - C|v|_{W_1^2(\Omega)} \right),$$

with Lagrange multiplier  $\lambda \in \mathbb{R}$ . We thus realize that the optimality condition reads

$$e_{\mathcal{T}}(v, T) = \frac{\lambda}{2} \quad \forall T \in \mathcal{T}$$

or that  $e_{\mathcal{T}}(v, T)$  is constant over  $\mathcal{T}$ . We rewrite this insightful conclusion as follows:

$$\text{A graded mesh is quasi-optimal if the local error is equidistributed.} \quad (3.24)$$

This calculation yields

$$E_{\mathcal{T}}(v)^2 = \frac{\lambda^2}{4} N, \quad C|v|_{W_1^2(\Omega)} = \frac{\lambda}{2} N,$$

whence

$$E_{\mathcal{T}}(v) = C|v|_{W_1^2(\Omega)} N^{-1/2} \quad (3.25)$$

is the optimal decay rate but with regularity  $v \in W_1^2(\Omega)$  rather than  $H^2(\Omega)$ ; this is the second instance of nonlinear approximation, namely mesh design tailored to the specific function  $v$  at hand; the first one was in Section 1. The principle of error equidistribution (3.24) was originally derived by Babuška and Rheinboldt [Babuška and Rheinboldt 1978] for  $d = 1$ , and extended to  $d = 2$  by Nochetto and Veiser [Nochetto and Veiser 2012, Section 1.6], using an idealized continuous minimization problem involving a meshsize function. The current formulation is closer to applications and does not require a positive power of  $h_T$  in (3.23).

**Remark 3.13 (point singularities)** Corner singularities [Grisvard 1985] as well as singularities due to intersecting interfaces [Kellogg 1974/75] are of the form

$$v(x) \approx r(x)^\gamma, \quad 0 < \gamma < 1, \quad (3.26)$$

for  $d = 2$ . This implies  $v \in W_1^2(\Omega)$  for all  $\gamma$  and the decay rate (3.25) provided  $\mathcal{T}$  equidistributes the  $H^1$ -error. Babuška et al [Babuška, Kellogg and Pitkäranta 1979] and Grisvard [Grisvard 1985] designed such meshes for corner singularities using weighted  $H^2$ -regularity. The preceding approach is more powerful in that it does not require any characterization of the singularities, rather than  $v \in W_1^2(\Omega)$ , and applies as well to line discontinuities for  $d = 2$ . We will come back to this point in Section 6.

We consider now an important abstract variant of the discrete minimization process leading to (3.24) that will be instrumental to understand the success of greedy algorithms later. Suppose that  $0 < q, p \leq \infty$ ,  $v \in L^q(\Omega)$ ,  $X_p^t(\Omega)$  is an abstract regularity space with  $t = \frac{1}{p} - \frac{1}{q} > 0$ , and  $E_{\mathcal{T}}(v)_q$  and  $e_{\mathcal{T}}(v, T)_q$  are global and local  $L^q$ -interpolation error indicators of  $v$  that satisfy the following two abstract properties:

- *Summability in  $\ell^q$* : There exists a constant  $C_1 > 0$  such that

$$E_{\mathcal{T}}(v)_q^q \leq C_1^q \sum_{T \in \mathcal{T}} e_{\mathcal{T}}(v, T)_q^q; \quad (3.27)$$

- *Summability in  $\ell^p$* : There exists a constant  $C_2 > 0$  such that

$$\sum_{T \in \mathcal{T}} e_{\mathcal{T}}(v, T)_q^p = C_2^p |v|_{X_p^t(\Omega)}^p. \quad (3.28)$$

We intend to find conditions on a mesh  $\mathcal{T}$  that minimize the global  $L^q$ -error  $E_{\mathcal{T}}(v)_q$  of  $v$  subject to the constraint (3.28). We again propose a Lagrangian

$$\mathcal{L}[\mathbf{e}, \lambda] := \sum_{T \in \mathcal{T}} e_{\mathcal{T}}(v, T)_q^q - \lambda \left( \sum_{T \in \mathcal{T}} e_{\mathcal{T}}(v, T)_q^p - C_2^p |v|_{X_p^t(\Omega)}^p \right).$$

The optimality condition for  $\mathbf{e}$  reads

$$e_{\mathcal{T}}(v, T)_q = \left( \lambda \frac{p}{q} \right)^{\frac{1}{q-p}} \quad \forall T \in \mathcal{T},$$

which is a third instance of error equidistribution and is thus consistent with (3.24).

We now resort (3.27) and (3.28) to arrive at

$$\sum_{T \in \mathcal{T}} e_{\mathcal{T}}(v, T)_q^q = N \left( \lambda \frac{p}{q} \right)^{\frac{q}{q-p}}, \quad C_2^p |v|_{X_p^t(\Omega)}^p = N \left( \lambda \frac{p}{q} \right)^{\frac{p}{q-p}},$$

whence

$$E_{\mathcal{T}}(v)_q \leq C_1 C_2 |v|_{X_p^t(\Omega)} N^{\frac{1}{q} - \frac{1}{p}}. \quad (3.29)$$

We see that the decay rate in (3.29) is  $-t = \frac{1}{q} - \frac{1}{p} < 0$  and is just dictated by the different summabilities of (3.27) and (3.28). In the applications of (3.29) below,  $t = \frac{s}{d}$  will be proportional to a differentiability index  $s$  and the condition

$$0 = t - \frac{1}{p} + \frac{1}{q} = \frac{s}{d} - \frac{1}{p} + \frac{1}{q}$$

will correspond to the spaces  $L^q(\Omega)$  and  $X_p^s(\Omega)$  being on the same nonlinear Sobolev scale. This minimization process is an idealization that does not account for mesh regularity of  $\mathcal{T}$ , which in turn entail some geometric constraints in the construction of  $\mathcal{T}$ . A key question is whether estimates of the form (3.29) can be achieved under practical but weaker conditions than (3.24). In Section 3.5 we will study the bisection method, a flexible technique for conforming mesh refinement with optimal complexity. In Section 3.6 we will present and analyze GREEDY, a practical algorithm that implements these ideas and constructs quasi-optimal conforming bisection meshes under the slightly stronger assumption

$$s - \frac{d}{p} + \frac{d}{q} > 0. \quad (3.30)$$

Moreover, in Section 3.7 we will extend this analysis to non-conforming meshes.

We realize from (3.29) that, in order to maximize the error decay rate, we would like to have  $p$  as small as possible, even  $0 < p < 1$ . The range of  $q, p$  does not matter in the argument above and, despite the fact that  $q \geq 1$  in all applications

below, the range of  $p$  is only limited by that of  $s$ , which in turn depends on the polynomial degree  $n \geq 1$  in that  $0 < s \leq n + 1$ .

We now return to the special case (3.23), namely  $q = 2$ ,  $p = 1$  and  $\nabla v \in L^2(\Omega)$ . As already shown in (3.23), in the nonlinear Sobolev scale

$$\text{sob}(W_1^2) - \text{sob}(H^1) = \left(2 - \frac{2}{1}\right) - \left(1 - \frac{2}{2}\right) = 0$$

we expect the best error decay

$$\|\nabla(v - I_{\mathcal{T}}v)\|_{L^2(\Omega)} \lesssim |v|_{W_1^2(\Omega)}(\#\mathcal{T})^{-\frac{1}{2}},$$

whereas the linear Sobolev scale yields the reduced order

$$\|\nabla(v - I_{\mathcal{T}}v)\|_{L^2(\Omega)} \lesssim |v|_{H^s(\Omega)}(\#\mathcal{T})^{-(s-1)/2}$$

for  $s < 1 + \gamma < 2$  and  $v$  satisfying (3.26), where  $I_{\mathcal{T}}$  is the Lagrange interpolation operator. The nonlinear Sobolev scale entails a trade of differentiability with integrability: we gain up to differentiability  $s = 2$  at the expense of lower integrability  $p = 1$  for polynomial degree  $n = 1$ . This trade-off is at the heart of the optimal estimate (3.25) and is represented in the so-called *DeVore diagram* in Fig. 2.1.

If the polynomial degree is  $n \geq 2$ , then the largest differentiability index is  $s = n + 1$ , which for  $d = 2$  leads to integrability index  $p < 1$ :

$$\left(s - \frac{2}{p}\right) - \left(1 - \frac{2}{2}\right) = 0 \quad \Rightarrow \quad p = \frac{2}{n+1} < 1. \quad (3.31)$$

To measure regularity of  $v$ , the corresponding Sobolev space must be replaced by the Besov space  $B_{p,p}^{n+1}(\Omega)$  or the Lipschitz space  $Lip_p^{n+1}(\Omega)$ . We will introduce and study these spaces in Section 6.8.

### 3.5 Conforming meshes: the bisection method

In order to approximate functions in  $W_p^k(\Omega)$  by piecewise polynomials, we decompose  $\Omega$  into simplices. We briefly discuss the *bisection* method, an elegant and versatile technique for subdividing  $\Omega$  in any dimension into a conforming mesh. We also discuss briefly nonconforming meshes in §3.7. We present complete proofs, especially of the complexity of bisection, later in §8.

We focus on  $d = 2$  and follow [Binev et al. 2004], but the results carry over to any dimension  $d > 2$  [Stevenson 2008]. We refer to [Nochetto et al. 2009] for a rather complete discussion for  $d \geq 2$ .

Let  $\mathcal{T}$  denote a *mesh* (triangulation or grid) made of simplices  $T$ , and let  $\mathcal{T}$  be *conforming* (edge-to-edge). Each element is labeled, namely it has an edge  $E(T)$  assigned for refinement (and an opposite vertex  $v(T)$  for  $d = 2$ ); see Figure 3.2.

The bisection method consists of a suitable *labeling* of the initial mesh  $\mathcal{T}_0$  and a rule to assign the refinement edge to the two children. For  $d = 2$  we consider the *newest vertex bisection* as depicted in Figure 3.2. For  $d > 2$  the situation is more complicated and one needs the concepts of type and vertex order [Nochetto et al.

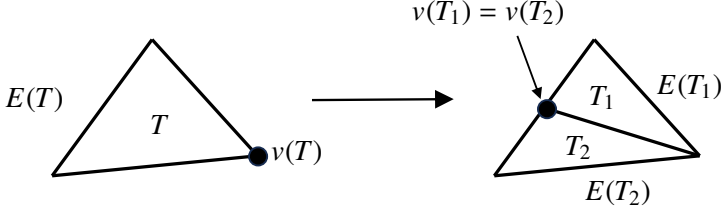


Figure 3.2. Triangle  $T \in \mathcal{T}$  with vertex  $v(T)$  and opposite refinement edge  $E(T)$ . The bisection rule for  $d = 2$  consists of connecting  $v(T)$  with the midpoint of  $E(T)$ , thereby giving rise to children  $T_1, T_2$  with common vertex  $v(T_1) = v(T_2)$ , the newly created vertex, and opposite refinement edges  $E(T_1), E(T_2)$ .

2009, Stevenson 2008]. More precisely, we identify a simplex  $T$  with the set of its *ordered vertices* and its *type*  $t$  by

$$T = \{z_0, z_1, \dots, z_d\}_t,$$

with  $t \in \{0, \dots, d-1\}$ . 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 face  $S = \{\bar{z}, z_1, \dots, z_{d-1}\}$  shared by the two children  $T_1, T_2$  of  $T$ . The *bisection rule* dictates the following vertex order and type for  $T_1, T_2$

$$\begin{aligned} T_1 &:= \{z_0, \bar{z}, z_1, \dots, z_t, \underbrace{z_{t+1}, \dots, z_{d-1}}_{\rightarrow}\}_{(t+1) \bmod d}, \\ T_2 &:= \{z_d, \bar{z}, \underbrace{z_1, \dots, z_t}_{\rightarrow}, \underbrace{z_{d-1}, \dots, z_{t+1}}_{\leftarrow}\}_{(t+1) \bmod d}, \end{aligned} \quad (3.32)$$

with the convention that arrows point in the direction of increasing indices and  $\{z_1, \dots, z_0\} = \emptyset$ ,  $\{z_d, \dots, z_{d-1}\} = \emptyset$ . For instance, in 3d the children of  $T = \{z_0, z_1, z_2, z_3\}_t$  are (see Fig. 3.3)

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

Note that the vertex labeling of  $T_1$  is type-independent, whereas that of  $T_2$  is the same for type  $t = 1$  and  $t = 2$ . 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. 3.3. The type of  $T$  then dictates which component of the triple is used to label the vertex.

Bisection creates a *unique* master forest  $\mathbb{F}$  of binary trees with infinite depth, where each node is a simplex (triangle in 2d), its two successors are the two

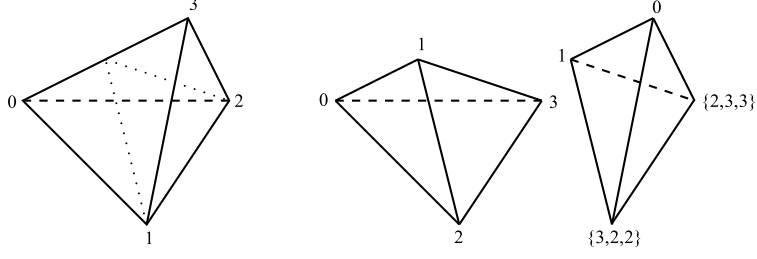


Figure 3.3. Refinement of a single tetrahedron  $T$  of type  $t$ . The child  $T_1$  in the middle has the same node ordering regardless of type. In contrast, for the child  $T_2$  on the right a triple is appended to two nodes. The local vertex index is given for these nodes by the  $t$ -th component of the triple.

children created by bisection, and the roots of the binary trees are the elements of the initial conforming partition  $\mathcal{T}_0$ . It is important to realize that, no matter how an element arises in the subdivision process, its associated newest vertex is unique and only depends on the labeling of  $\mathcal{T}_0$ : so the edge  $E(T)$  assigned for refinement (and the opposite vertex  $v(T)$  for  $d = 2$ ) are independent of the order of the subdivision process for all  $T \in \mathbb{F}$ ; see Lemma 8.1 in Sect. 8. Therefore,  $\mathbb{F}$  is unique.

A finite subset  $\mathcal{F} \subset \mathbb{F}$  is called a *forest* if  $\mathcal{T}_0 \subset \mathcal{F}$  and the nodes of  $\mathcal{F}$  satisfy

- all nodes of  $\mathcal{F} \setminus \mathcal{T}_0$  have a predecessor;
- all nodes in  $\mathcal{F}$  have either two successors or none.

Any node  $T \in \mathcal{F}$  is thus uniquely connected with a node  $T_0$  of the initial triangulation  $\mathcal{T}_0$ , i.e.  $T$  belongs to the infinite tree  $\mathbb{F}(T_0)$  emanating from  $T_0$ . Furthermore, any forest may have *interior nodes*, i.e. nodes with successors, as well as *leaf nodes*, i.e. nodes without successors. The set of leaves corresponds to a mesh (or triangulation, grid, partition)  $\mathcal{T} = \mathcal{T}(\mathcal{F})$  of  $\mathcal{T}_0$  which may not be conforming or edge-to-edge.

We thus introduce the set  $\mathbb{T}$  of all *conforming* refinements of  $\mathcal{T}_0$ :

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

If  $\mathcal{T}_* = \mathcal{T}(\mathcal{F}_*) \in \mathbb{T}$  is a conforming refinement of  $\mathcal{T} = \mathcal{T}(\mathcal{F}) \in \mathbb{T}$ , we write  $\mathcal{T}_* \geq \mathcal{T}$  and understand this inequality in the sense of trees, namely  $\mathcal{F} \subset \mathcal{F}_*$ .

*Example.* Consider  $\mathcal{T}_0 = \{T_i\}_{i=1}^4$  and the longest edge to be the refinement edge. Figure 3.4 displays a sequence of conforming meshes  $\mathcal{T}_k \in \mathbb{T}$  created by bisection. Each element  $T_i$  of  $\mathcal{T}_0$  is a root of a finite tree emanating from  $T_i$ , which together form the forest  $\mathcal{F}_2$  corresponding to mesh  $\mathcal{T}_2 = \mathcal{T}(\mathcal{F}_2)$ . Figure 3.5 displays  $\mathcal{F}_2$ , whose leaf nodes are the elements of  $\mathcal{T}_2$ .

*Properties of Bisection.* We now discuss several crucial geometric properties of bisection. We start by recalling the concept of shape regularity. For any  $T \in \mathcal{T}$ , we define

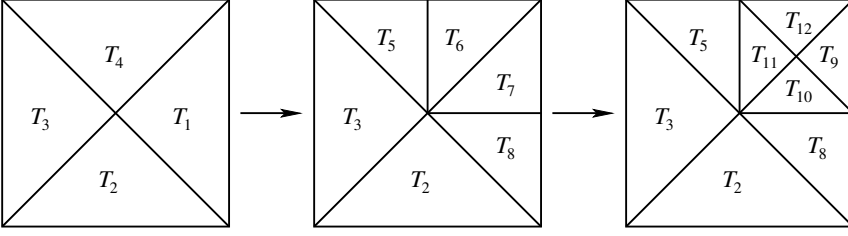
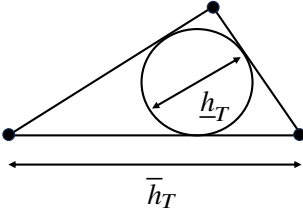


Figure 3.4. Sequence of bisection meshes  $\{\mathcal{T}_k\}_{k=0}^2$  starting from the initial mesh  $\mathcal{T}_0 = \{T_i\}_{i=1}^4$  with longest edges labeled for bisection. Mesh  $\mathcal{T}_1$  is created from  $\mathcal{T}_0$  upon bisecting  $T_1$  and  $T_4$ , whereas mesh  $\mathcal{T}_2$  arises from  $\mathcal{T}_1$  upon refining  $T_6$  and  $T_7$ . The bisection rule is described in Figure 3.2.



Figure 3.5. Forest  $\mathcal{F}_2$  corresponding to the grid sequence  $\{\mathcal{T}_k\}_{k=0}^2$  of Figure 3.4. The roots of  $\mathcal{F}_2$  form the initial mesh  $\mathcal{T}_0$  and the leaves of  $\mathcal{F}_2$  constitute the conforming bisection mesh  $\mathcal{T}_2$ . Moreover, each level of  $\mathcal{F}_2$  corresponds to all elements with generation equal to the level.



$$\bar{h}_T := \text{diam}(T)$$

$$h_T := |T|^{1/d}$$

$$\underline{h}_T := 2 \sup\{r > 0 \mid B(x, r) \subset T \text{ for } x \in T\}.$$

Then

$$\underline{h}_T \leq h_T \leq \bar{h}_T \leq \sigma \underline{h}_T \quad \forall T \in \mathcal{T},$$

where  $\sigma > 1$  is the shape regularity constant of (3.9). The next lemma guarantees that bisection keeps  $\sigma$  bounded.

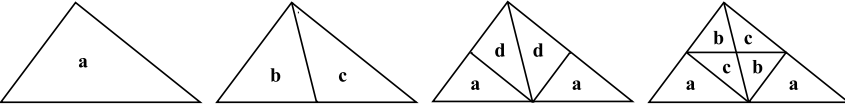


Figure 3.6. Bisection produces at most 4 similarity classes for any triangle.

**Lemma 3.14 (shape regularity)** *The partitions  $\mathcal{T}$  generated by newest vertex bi-*

section satisfy a uniform minimal angle condition, or equivalently  $\sigma$  is uniformly bounded, only depending on the initial partition  $\mathcal{T}_0$ .

*Proof.* Each  $T \in \mathcal{T}_0$  gives rise to a fixed number of similarity classes, namely 4 for  $d = 2$  according to Figure 3.6. This, combined with the fact that  $\#\mathcal{T}_0$  is finite, yields the assertion.  $\square$

We define the *generation (or level)*  $g(T)$  of an element  $T \in \mathcal{T}$  as the number of bisections needed to create  $T$  from its ancestor  $T_0 \in \mathcal{T}_0$ . Since bisection splits an element into two children with equal measure, we realize that

$$h_T = 2^{-g(T)/2} h_{T_0} \quad \forall T \in \mathcal{T}. \quad (3.33)$$

Referring to Figure 3.5 we observe that the leaf nodes  $T_9, T_{10}, T_{11}, T_{12}$  have generation 2, whereas  $T_5, T_8$  have generation 1 and  $T_2, T_3$  have generation 0.

The following geometric property is a simple consequence of (3.33).

**Lemma 3.15 (element size vs generation)** *There exist constants  $0 < D_1 < D_2$ , only depending on  $\mathcal{T}_0$ , such that*

$$D_1 2^{-g(T)/2} \leq h_T < \bar{h}_T \leq D_2 2^{-g(T)/2} \quad \forall T \in \mathcal{T}. \quad (3.34)$$

*Labeling and Bisection Rule.* Whether the recursive application of bisection does not lead to inconsistencies depends on a suitable initial labeling of edges and a bisection rule. For  $d = 2$  they are simple to state [Binev et al. 2004]. Given  $T \in \mathcal{T}$  with generation  $g(T) = i$ , we assign the label  $(i+1, i+1, i)$  to  $T$  with  $i$  corresponding to the refinement edge  $E(T)$ . The following rule dictates how the labeling changes with refinement: the side  $i$  is bisected and both new sides as well as the bisector are labeled  $i+2$  whereas the remaining labels do not change. To guarantee that the label of an edge is independent of the elements sharing this edge, we need a special labeling for  $\mathcal{T}_0$ , due [Mitchell 1989, Theorem 2.9] and [Binev et al. 2004, Lemma 2.1] for  $d = 2$ :

$$\begin{aligned} &\text{edges of } \mathcal{T}_0 \text{ have labels 0 or 1 and all elements } T \in \mathcal{T} \text{ have} \\ &\text{exactly two edges with label 1 and one with label 0.} \end{aligned} \quad (3.35)$$

There is a variant for  $d > 2$  due to [Stevenson 2008, Section 4]. It is not obvious that labeling (3.35) exists, but if it does then all elements of  $\mathcal{T}_0$  can be split into pairs of compatibly divisible elements. We refer to Figure 3.7 for an example of initial labeling of  $\mathcal{T}_0$  satisfying (3.35) and the way it evolves for two successive refinements  $\mathcal{T}_2 \geq \mathcal{T}_1 \geq \mathcal{T}_0$  corresponding to Figure 3.4.

To guarantee (3.35) we can proceed as follows: given a coarse mesh of elements  $T$  we can bisect twice each  $T$  and label the four grandchildren, as indicated in Figure 3.8 for the resulting mesh  $\mathcal{T}_0$  to satisfy the initial labeling [Binev et al. 2004].

For  $d \geq 3$  a general strategy of initial labeling is due to [Stevenson 2008, Section 4 - Condition (b)], who in turn improves upon [Maubach 1995] and [Traxler 1997] and shows how to impose it upon further refining each element of  $\mathcal{T}_0$ . We refer to the



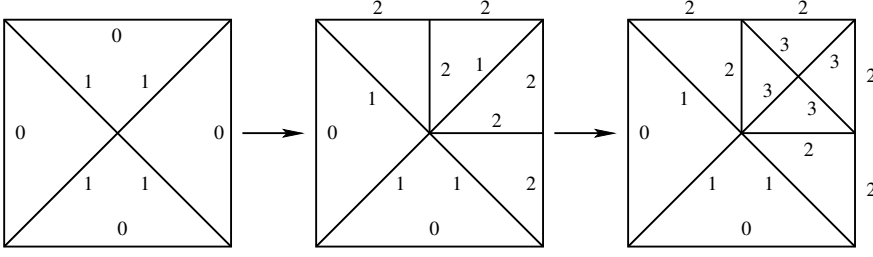


Figure 3.7. Initial labeling and its evolution for the sequence of conforming refinements  $\mathcal{T}_0 \leq \mathcal{T}_1 \leq \mathcal{T}_2$  of Figure 3.4.

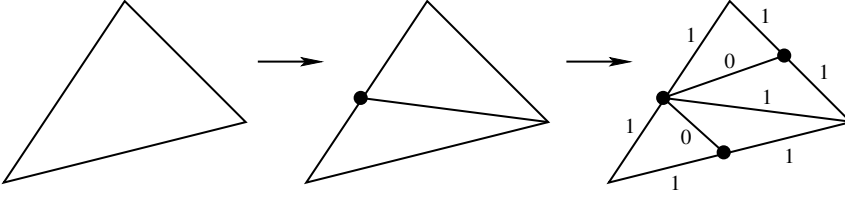


Figure 3.8. Bisecting each triangle of  $\mathcal{T}_0$  twice and labeling edges in such a way that all boundary edges have label 1 yields an initial mesh satisfying (3.35).

survey [Nochetto *et al.* 2009] for a discussion of this condition: a key consequence is

$$\text{every uniform refinement of } \mathcal{T}_0 \text{ gives a conforming bisection mesh.} \quad (3.36)$$

Condition (3.35) is still valid, and a construction by successive bisections similar, but much trickier, than the one described for  $d = 2$  can be performed to fulfil it; yet for  $d = 3$  the number of elements increases by an order of magnitude, which indicates that (3.35) is a severe restriction in practice. Finding alternative, more practical, conditions is an important problem.

*Initialization of arbitrary triangulations.* A novel initialization procedure that can be applied to *any* conforming triangulation  $\mathcal{T}_0$  has been recently proposed by [Diening, Gehring and Storn 2023]; hereafter we present a short account of it.

The key concept is that of *coloring* the vertices of  $\mathcal{T}_0$ . A colored initial triangulation in  $\mathbb{R}^d$  is a pair  $(\mathcal{T}_0, c)$ , where  $c : \mathcal{V}_{\mathcal{T}_0} \rightarrow \{0, \dots, d\}$  is such that the colors of all vertices of each  $T \in \mathcal{T}_0$  are distinct. The color map  $c$  allows one to sort the vertices of each initial element  $T = \{z_0, \dots, z_d\}_t \in \mathcal{T}_0$  so that

$$c(z_j) = j \quad j \in \{0, \dots, d\}.$$

To refine a marked  $T \in \mathcal{T}_0$ , one applies the Maubach bisection rule leading to (3.32), and possibly adds a recursive closure, which is proven to terminate, to guarantee the conformity of the final triangulation. The coloring property is conserved in this process, and the conclusion of Theorem 3.16 below holds true, starting from any initially colored triangulation  $\mathcal{T}_0$ .

Unfortunately, not every initial triangulation can be colored. For instance, consider in dimension  $d = 2$  a patch of triangles sharing a common vertex. If color 0 is assigned to such inner vertex, then the outer vertices must take successively colors 1 and 2, but if the number of triangles in the patch is odd there will be a vertex which is not colorable.

To overcome this obstruction, [Diening *et al.* 2023] propose to use more colors, and introduce the concept of *generalized coloring*: a pair  $(\mathcal{T}_0, c)$  is a  $(N+1)$ -colored triangulation if there exists an integer  $N \geq d$  and a mapping  $c : \mathcal{V}_{\mathcal{T}_0} \rightarrow \{0, \dots, N\}$  such that the colors of all vertices of each  $T \in \mathcal{T}_0$  are distinct. Any initial  $\mathcal{T}_0$  can be colored in this generalized sense: indeed, after the initialization  $c(z) = +\infty$  for all  $z \in \mathcal{V}_{\mathcal{T}_0}$ , one defines

$$c(z) := \min (\mathbb{N}_0 \setminus \{c(w) : [z, w] \text{ is an edge of } \mathcal{T}_0\}), \quad z \in \mathcal{V}_{\mathcal{T}_0},$$

as the smallest color not already attained by a neighboring vertex. Then, one sets  $N := \max \{c(z) : z \in \mathcal{V}_{\mathcal{T}_0}\}$ , and notes that  $N$  is bounded by the maximal number of edges connected to a vertex of  $\mathcal{T}_0$ .

A generalized  $(N+1)$ -colored triangulation  $(\mathcal{T}_0, c)$  in  $\mathbb{R}^d$  can be seen as a collection of  $d$ -simplices contained in a virtual, colored triangulation  $\mathcal{T}_0^+$  in  $\mathbb{R}^N$ . It suffices to add  $N - d$  virtual nodes to each simplex in  $\mathcal{T}_0$ , so that it becomes a  $N$ -simplex, and attribute to these nodes the remaining  $N - d$  colors. Note that these virtual simplices are only connected via their  $d$ -subsimplices belonging to  $\mathcal{T}_0$ . In the example mentioned above of a patch of triangles sharing a vertex, a  $(3+1)$ -colored triangulation is defined as follows: a tetrahedron is built on top of each triangle; the previously uncolorable vertex takes the new color 3, whereas color 2 is attributed to the new vertices of the two tetrahedra sharing that vertex; the new vertex of any other tetrahedron takes the color 3.

With the new triangulation  $\mathcal{T}_0^+$  at hand, one could apply the Mauback bisection rule to it, which as a by-product would refine the initial triangulation  $\mathcal{T}_0$ . However, [Diening *et al.* 2023] suggests a short-cut that refines directly  $\mathcal{T}_0$ , by invoking an algorithm that bisects a  $k$ -simplex in dimension  $m > k$ . A further round of recursive refinements may be needed to guarantee conformity. Diening, Gehring, and Storn prove that the recursion terminates. In addition, for any  $(N+1)$ -colored initial triangulation, the conclusion of Theorem 3.16 below holds true also in this case, with a constant  $D$  satisfying  $D \lesssim N^d$ .

*The procedure REFINE.* Given  $\mathcal{T} \in \mathbb{T}$  and a selected subset  $\mathcal{M} \subset \mathcal{T}$  (the set of *marked elements*), the procedure

$$[\mathcal{T}_*] = \text{REFINE}(\mathcal{T}, \mathcal{M})$$

creates a new conforming refinement  $\mathcal{T}_*$  of  $\mathcal{T}$  by bisecting all elements of  $\mathcal{M}$  at least once and perhaps additional elements to keep conformity.

Conformity is a constraint in the refinement procedure that prevents it from being completely local. The propagation of refinement beyond the set of marked elements  $\mathcal{M}$  is a rather delicate matter, which we discuss later in Sect. 8. For instance, we

show that a naive estimate of the form

$$\#\mathcal{T}_* - \#\mathcal{T} \leq D \#\mathcal{M}$$

is *not* valid with an absolute constant  $D$  independent of the refinement level. This can be repaired upon considering the cumulative effect for a sequence of conforming bisection meshes  $\{\mathcal{T}_k\}_{k=0}^\infty$ . This is expressed in the following crucial complexity result due to [Binev *et al.* 2004] for  $d = 2$  and [Stevenson 2008] for  $d > 2$ . We present a complete proof later in Sect. 8.

**Theorem 3.16 (complexity of REFINE)** *If  $\mathcal{T}_0$  satisfies the initial labeling (3.35) for  $d = 2$ , or that in [Stevenson 2008, Section 4] for  $d > 2$ , then there exists a constant  $D > 0$  only depending on  $\mathcal{T}_0$  and  $d$  such that for all  $k \geq 1$*

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \leq D \sum_{j=0}^{k-1} \#\mathcal{M}_j.$$

If elements  $T \in \mathcal{M}$  are to be bisected  $b \geq 1$  times, then the procedure REFINE can be applied recursively, and Theorem 3.16 remains valid with  $D$  also depending on  $b$ .

*Mesh Overlay.* For the subsequent discussion it will be convenient to merge (or superpose) two conforming meshes  $\mathcal{T}_1, \mathcal{T}_2 \in \mathbb{T}$ , thereby giving rise to the so-called *overlay*  $\mathcal{T}_1 \oplus \mathcal{T}_2$ . This operation corresponds to the union in the sense of trees [Cascón, Kreuzer, Nochetto and Siebert 2008, Stevenson 2007]. We next bound the cardinality of  $\mathcal{T}_1 \oplus \mathcal{T}_2$  in terms of that of  $\mathcal{T}_1$  and  $\mathcal{T}_2$ .

**Lemma 3.17 (mesh overlay)** *Let  $\mathcal{T}_1, \mathcal{T}_2 \in \mathbb{T}$ . The overlay  $\mathcal{T} = \mathcal{T}_1 \oplus \mathcal{T}_2 \in \mathbb{T}$  is conforming and*

$$\#\mathcal{T} \leq \#\mathcal{T}_1 + \#\mathcal{T}_2 - \#\mathcal{T}_0. \quad (3.37)$$

For a proof we refer to [Cascón *et al.* 2008, Lemma 3.7] and to Proposition 8.15 below for a more general situation.

### 3.6 Constructive approximation

We now construct graded bisection meshes  $\mathcal{T}$  for  $n = 1, d = 2$  that achieve the optimal decay rate  $(\#\mathcal{T})^{-1/2}$  of (3.25) under the global regularity assumption

$$v \in W_p^2(\Omega), \quad p > 1. \quad (3.38)$$

Therefore,  $W_p^2(\Omega)$  is strictly above the Sobolev line for the space  $H^1(\Omega)$ :  $\text{sob}(W_p^2) = 2 - \frac{2}{p} > 0 = \text{sob}(H^1)$ . Note that  $s = 1, p > 1$  and  $q = 2$  obey the restriction (3.30) for  $\nabla v \in L^2(\Omega)$ . In particular,  $W_p^2(\Omega)$  is compactly embedded into  $H^1(\Omega)$  according to Lemma 2.1 (Sobolev embedding).

Following [Binev, Dahmen, DeVore and Petrushev 2002] and [Gaspoz and Morin 2014], we use a greedy algorithm that is based on the knowledge of the element errors and on bisection. The algorithm hinges on (3.24): if  $\delta > 0$  is a given

tolerance, the element error is equidistributed and within tolerance  $e_{\mathcal{T}}(v, T) \approx \delta$ , and the global error decays with maximum rate  $(\#\mathcal{T})^{-1/2}$ , then

$$\delta^2 \#\mathcal{T} \approx \sum_{T \in \mathcal{T}} e_{\mathcal{T}}(v, T)^2 = |v - I_{\mathcal{T}}v|_{H^1(\Omega)}^2 \lesssim (\#\mathcal{T})^{-1}$$

whence  $\#\mathcal{T} \lesssim \delta^{-1}$ ; here  $I_{\mathcal{T}}$  stands for the Lagrange interpolation operator. With this in mind, we impose  $e_T(v) \leq \delta$  as a threshold to stop refining and expect  $\#\mathcal{T} \lesssim \delta^{-1}$ . The following algorithm implements this idea.

**Algorithm 3.18 (greedy algorithm)** Given a tolerance  $\delta > 0$  and a conforming mesh  $\mathcal{T}_0$ , GREEDY finds a conforming refinement  $\mathcal{T} \geq \mathcal{T}_0$  of  $\mathcal{T}_0$  by bisection such that  $e_{\mathcal{T}}(v, T) \leq \delta$  for all  $T \in \mathcal{T}$ : let  $\mathcal{T} = \mathcal{T}_0$  and

```

 $[\mathcal{T}] = \text{GREEDY}(\mathcal{T}, \delta, v)$ 
  while  $\mathcal{M} := \{T \in \mathcal{T} \mid e_{\mathcal{T}}(v, T) > \delta\} \neq \emptyset$ 
     $\mathcal{T} := \text{REFINE}(\mathcal{T}, \mathcal{M})$ 
  return  $\mathcal{T}$ 

```

Since  $W_p^2(\Omega) \subset C^0(\overline{\Omega})$ , because  $p > 1$ , we can use the Lagrange interpolant and local estimate (3.17) with  $r = \text{sob}(W_p^2) - \text{sob}(H^1) = 2 - 2/p > 0$ . We deduce

$$e_{\mathcal{T}}(v, T) \lesssim h_T^r \|D^2 v\|_{L^p(T)}. \quad (3.39)$$

We assess the quality of the resulting mesh in a slightly more general setting, following [Bonito, Cascón, Mekchay, Morin and Nochetto 2016, Proposition 1 and Corollary 1], needed later in Sections 6 and 7 for solution and data approximation for any polynomial degree.

*An abstract greedy algorithm.* We consider a generic (possibly vector-valued) function  $v \in L^q(\Omega, \mathbb{R}^M)$ , with  $M \geq 1$  and  $1 \leq q \leq \infty$  and denote by  $e_{\mathcal{T}}(v, T) = \|v - \Pi_{\mathcal{T}}v\|_{L^q(T)}$  the abstract  $L^q$ -local error for  $T \in \mathcal{T}$  used in the GREEDY procedure and by  $E_{\mathcal{T}}(v) = \|v - \Pi_{\mathcal{T}}v\|_{L^q(\Omega)}$  the global  $L^q$ -interpolation error by either continuous or discontinuous piecewise polynomials (the definition of  $\Pi_{\mathcal{T}}v$  is irrelevant now). We formulate the following assumptions:

- *Summability in  $\ell^q$ :* The errors  $\{e_{\mathcal{T}}(v, T)\}_{T \in \mathcal{T}}$  satisfy

$$\|v\|_{L^q(\Omega)}^q \lesssim \sum_{T \in \mathcal{T}} e_{\mathcal{T}}(v, T)^q. \quad (3.40)$$

Rather than (3.38) we assume that  $v$  belongs to an abstract space  $X_p^s(\Omega; \mathcal{T}_0)$  of functions with differentiability index  $s \in (0, n]$  and integrability index  $p \in (0, \infty]$  piecewise over  $\mathcal{T}_0$  with two crucial properties:

- *Local error estimate:* For  $r = s - \frac{d}{p} + \frac{d}{q} > 0$  and all  $T \in \mathcal{T}$

$$e_{\mathcal{T}}(v, T) \lesssim h_T^r |v|_{X_p^s(T)} \quad (3.41)$$

- *Norm subadditivity*: For  $p < \infty$ , and obvious modification for  $p = \infty$ ,

$$\sum_{T \in \mathcal{T}} |v|_{X_p^s(T)}^p \lesssim |v|_{X_p^s(\Omega; \mathcal{T}_0)}^p. \quad (3.42)$$

The space  $X_p^s(\Omega; \mathcal{T}_0)$  will later be either a Sobolev space  $W_p^s(\Omega; \mathcal{T}_0)$ , a Besov space  $B_{p,p}^s(\Omega; \mathcal{T}_0)$  or a Lipschitz space  $Lip_p^s(\Omega; \mathcal{T}_0)$ , with piecewise regularity over  $\mathcal{T}_0$ ; the latter two will allow  $0 < p < 1$ . For the moment we do not need to be specific and just rely on the two properties above.

**Proposition 3.19 (abstract greedy error)** *Let  $\mathcal{T}_0$  be an initial subdivision of  $\Omega \subset \mathbb{R}^d$  satisfying the initial labeling property (3.35) for  $d = 2$ , or its variant for  $d > 2$ . Let  $M \geq 1, 0 < q, p \leq \infty$  and  $s - \frac{d}{p} + \frac{d}{q} > 0$ . Let  $v \in L^q(\Omega, \mathbb{R}^M)$  satisfy (3.40), (3.41) and (3.42). Then  $\text{GREEDY}(\mathcal{T}_0, \delta, v)$  terminates in a finite number of iterations with local errors verifying  $e_{\mathcal{T}}(v, T) \leq \delta$  for all  $T \in \mathcal{T}$ , and there is a constant  $C = C(p, q, s, d, \Omega, \mathcal{T}_0)$  such that the output  $\mathcal{T} \in \mathbb{T}$  satisfies*

$$\|v - \Pi_{\mathcal{T}} v\|_{L^q(\Omega)} \leq C |v|_{X_p^s(\Omega; \mathcal{T}_0)} (\#\mathcal{T} - \#\mathcal{T}_0)^{-\frac{s}{d}}. \quad (3.43)$$

*Proof.* We proceed in several steps.

**[1] Termination.** Since  $h_T$  decreases monotonically to 0 with bisection, so does  $e_{\mathcal{T}}(v, T)$  in view of (3.41). Consequently,  $\text{GREEDY}$  terminates in finite number  $k \geq 1$  of iterations. Upon termination, the local errors satisfy  $e_{\mathcal{T}}(v, T) \leq \delta$  for all  $T \in \mathcal{T}$  by construction, whence (3.40) implies

$$\|v - \Pi_{\mathcal{T}} v\|_{L^q(\Omega)} \lesssim \delta (\#\mathcal{T})^{\frac{1}{q}}.$$

**[2] Counting.** Let  $\mathcal{M} = \mathcal{M}_0 \cup \dots \cup \mathcal{M}_{k-1}$  be the set of marked elements. We organize the elements in  $\mathcal{M}$  by size in a way that allows for a counting argument. Let  $\mathcal{P}_j$  be the set of elements  $T$  of  $\mathcal{M}$  with size

$$2^{-(j+1)} \leq |T| < 2^{-j} \quad \Rightarrow \quad 2^{-\frac{j+1}{d}} \leq h_T < 2^{-\frac{j}{d}},$$

because  $h_T = |T|^{1/d}$  for shape regular meshes  $\mathcal{T} \in \mathbb{T}$ .

We first observe that all  $T$ 's in  $\mathcal{P}_j$  are *disjoint*. This is because if  $T_1, T_2 \in \mathcal{P}_j$  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 which works in any dimension  $d \geq 1$ ; see Section 8. Hence,

$$|T_1| \leq \frac{1}{2} |T_2|$$

contradicting the definition of  $\mathcal{P}_j$ . This implies the first bound

$$2^{-(j+1)} \#\mathcal{P}_j \leq |\Omega| \quad \Rightarrow \quad \#\mathcal{P}_j \leq |\Omega| 2^{j+1}. \quad (3.44)$$

In light of (3.41), we have for  $T \in \mathcal{P}_j$

$$\delta \leq e_{\mathcal{T}}(v, T) \lesssim 2^{-\frac{j}{d}} |v|_{X_p^s(T)}.$$

Therefore, accumulating these quantities in  $\ell^p$  and invoking (3.42) yields

$$\delta^p \#\mathcal{P}_j \lesssim 2^{-\frac{irp}{d}} \sum_{T \in \mathcal{P}_j} |v|_{X_p^s(T)}^p \lesssim 2^{-\frac{irp}{d}} |v|_{X_p^s(\Omega; \mathcal{T}_0)}^p$$

and gives rise to the second bound

$$\#\mathcal{P}_j \lesssim \delta^{-p} 2^{-\frac{irp}{d}} |v|_{X_p^s(\Omega; \mathcal{T}_0)}^p. \quad (3.45)$$

3 *Cardinality.* The two bounds for  $\#\mathcal{P}$  in (3.44) and (3.45) are complementary. The first one 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

$$2^{j_0+1} |\Omega| \approx \delta^{-p} 2^{-\frac{j_0 r p}{d}} |v|_{X_p^s(\Omega; \mathcal{T}_0)}^p \Rightarrow 2^{j_0} \approx \left( |\Omega|^{-1} \delta^{-p} |v|_{X_p^s(\Omega; \mathcal{T}_0)}^p \right)^{\frac{d}{d+rp}}.$$

We now compute

$$\#\mathcal{M} = \sum_j \#\mathcal{P}_j \lesssim \sum_{j \leq j_0} 2^j |\Omega| + \delta^{-p} |v|_{X_p^s(\Omega)}^p \sum_{j > j_0} (2^{-\frac{rp}{2}})^j.$$

Since

$$\sum_{j \leq j_0} 2^j \approx 2^{j_0}, \quad \sum_{j > j_0} (2^{-\frac{rp}{2}})^j \lesssim 2^{-\frac{rp j_0}{d}},$$

we can write

$$\#\mathcal{M} \lesssim |\Omega|^{1-\frac{d}{d+rp}} (\delta^{-1} |v|_{X_p^s(\Omega; \mathcal{T}_0)})^{\frac{dp}{d+rp}}.$$

We finally apply Theorem 3.16 (complexity of REFINES) to arrive at

$$\#\mathcal{T} - \#\mathcal{T}_0 \lesssim \#\mathcal{M} \lesssim |\Omega|^{\frac{rp}{d+rp}} (\delta^{-1} |v|_{X_p^s(\Omega; \mathcal{T}_0)})^{\frac{dp}{d+rp}},$$

or equivalently

$$\delta \lesssim |\Omega|^{\frac{r}{d}} |v|_{X_p^s(\Omega; \mathcal{T}_0)} (\#\mathcal{T} - \#\mathcal{T}_0)^{-\frac{d+rp}{d}}.$$

4 *Total error.* Since  $\frac{d+rp}{dp} = \frac{s}{d} + \frac{1}{q}$  we deduce from Step 1

$$\|v - \Pi_{\mathcal{T}} v\|_{L^q(\Omega)} \lesssim \delta (\#\mathcal{T})^{\frac{1}{q}} \lesssim |\Omega|^{\frac{r}{d}} |v|_{X_p^s(\Omega; \mathcal{T}_0)} (\#\mathcal{T} - \#\mathcal{T}_0)^{-\frac{s}{d}},$$

which is the desired estimate.  $\square$

The output mesh  $\mathcal{T}$  of GREEDY( $\mathcal{T}_0, \delta, v$ ) starting from  $\mathcal{T}_0$  satisfies  $\#\mathcal{T} \geq c_0 \#\mathcal{T}_0$  for some  $c_0 > 1$ , whence  $\#\mathcal{T} - \#\mathcal{T}_0 \geq (1 - \frac{1}{c_0}) \#\mathcal{T}$  and (3.43) yields

$$\|v - \Pi_{\mathcal{T}} v\|_{L^q(\Omega)} \lesssim C |v|_{X_p^s(\Omega; \mathcal{T}_0)} (\#\mathcal{T})^{-\frac{s}{d}}, \quad (3.46)$$

where  $C$  depends on  $c_0$ . It will be convenient in many applications of GREEDY, to be discussed later in Sections 6 and 7, that the starting mesh be a conforming

refinement of  $\mathcal{T}_0$  to enhance its efficiency. We will prove in Section 7.1 that (3.46) remains valid.

It is instructive to realize that GREEDY is a practical algorithm that hinges on the different summabilities of (3.40) and (3.42), and delivers a global  $L^q$ -error consistent with (3.29) of Section 3.4. Moreover, the outcome graded grid  $\mathcal{T}$  is quasi-optimal but may not equidistribute the error, not even approximately.

We are now in a position to show that GREEDY constructs optimal graded meshes for the interpolation error in  $H^1(\Omega)$  alluded to at the beginning of this section. To this end, we let  $I_{\mathcal{T}}$  be the Lagrange interpolation operator for  $d = 2$ .

**Corollary 3.20 (optimal  $H^1$ -convergence rate)** *If  $v \in H^1(\Omega) \cap W_p^2(\Omega)$  for  $1 < p \leq 2$  and  $d = 2$ , then GREEDY yields graded bisection meshes  $\mathcal{T}$  so that*

$$|v - I_{\mathcal{T}}v|_{H^1(\Omega)} \lesssim |\Omega|^{1-1/p} \|D^2v\|_{L^p(\Omega)} (\#\mathcal{T})^{-1/2}.$$

*Proof.* We invoke Proposition 3.19 (abstract greedy error) and equation (3.46) for  $\nabla v \in L^2(\Omega, \mathbb{R}^2)$  with  $\Pi_{\mathcal{T}}\nabla w = \nabla I_{\mathcal{T}}v$  and  $s = 1$ ,  $q = 2$ ,  $p > 1$  whence  $s - \frac{d}{p} + \frac{d}{q} > 0$ .  $\square$

**Remark 3.21 (piecewise  $W_p^2$ -smoothness)** Since (3.39) is completely local for  $d = 2$ , we see from (3.42) that it suffices for  $v \in H^1(\Omega)$  to be piecewise in  $W_p^2$  over the initial partition  $\mathcal{T}_0$ , namely  $W_p^2(\Omega; \mathcal{T}_0)$ . It turns out that this statement is valid for any dimension  $d \geq 2$  in view of Proposition 3.9 (approximation of gradients). We will revisit this issue in Section 6.8.

**Remark 3.22 (case  $p < 1$ )** We consider now polynomial degree  $n \geq 1$ . The integrability  $p$  corresponding to differentiability  $n+1$  results from equating Sobolev numbers:

$$n + 1 - \frac{d}{p} = \text{sob}(H^1) = 1 - \frac{d}{2} \quad \Rightarrow \quad p = \frac{2d}{2n + d}.$$

Depending on  $d \geq 2$  and  $n \geq 1$ , this may lead to  $0 < p < 1$ , in which case  $W_p^{n+1}(\Omega)$  is to be replaced by the Besov space  $B_{p,p}^s(\Omega)$  for  $s < n+1$  or the Lipschitz space  $Lip_p^{n+1}(\Omega)$  [ DeVore 1998 ]. We will discuss this matter in Section 6.8 and make the abstract greedy setting precise.

**Remark 3.23 (isotropic vs anisotropic elements)** Since geometric singularities are of the form (3.26) for  $d = 2$ , Corollary 3.20 (optimal  $H^1$ -convergence rate) shows that isotropic graded meshes are able to deliver optimal convergence rates for  $d = 2$ . Unfortunately, this is no longer the case for  $d > 2$  and anisotropic meshes are necessary for optimal meshes. This topic is delicate in several respects. Deriving reliable and efficient a posteriori error estimators is largely open for anisotropic meshes; this is the subject of Section 4 for isotropic meshes. Even having such estimators, building a theory of adaptivity is open; this is the subject of Sections 5, 6 and 10 for isotropic meshes. Finally, constructing anisotropic meshes based on a posteriori information alone and that easily allow for refinement and coarsening is

problematic. For these reasons we do not dwell on anisotropic refinement in this survey.

### 3.7 Nonconforming meshes

More general subdivisions of  $\Omega$  than those in §3.5 are used in practice. If the elements of  $\mathcal{T}_0$  are quadrilaterals for  $d = 2$ , or their multidimensional variant for  $d > 2$ , then it is natural to allow for improper or *hanging nodes* for the resulting refinements  $\mathcal{T}$  to be graded; see Figure 3.9 (a). On the other hand, if  $\mathcal{T}_0$  is made of triangles for  $d = 2$ , or simplices for  $d > 2$ , then red refinement without green completion also gives rise to graded meshes with hanging nodes; see Figure 3.9 (b). In both cases, the presence of hanging nodes is inevitable to enforce mesh grading. Finally, bisection may produce meshes with hanging nodes, as depicted in Figure 3.9 (c), if the completion process is incomplete. All three refinements maintain shape regularity, but for both practice and theory, they cannot be arbitrary: we need to restrict the level of nonconformity. We discuss this next, starting with the case of polynomial degree  $n = 1$  [Bonito and Nochetto 2010, Beirão da Veiga *et al.* 2023].

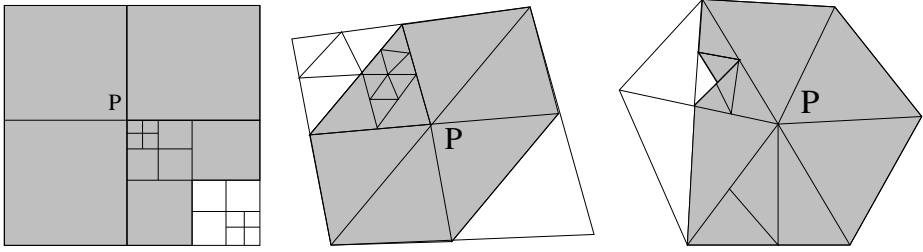


Figure 3.9. Nonconforming meshes made of quadrilaterals (a), triangles with red refinement (b), and triangles with bisection (c). The shaded regions depict the domain of influence of a proper or conforming node  $P$ .

We say that a node  $P$  of  $\mathcal{T}$  is a *proper* (or *conforming*) node if it is a vertex of all elements containing  $P$ ; otherwise, we say that  $P$  is an *improper* (nonconforming or hanging) node. The set  $\mathcal{N}$  of all nodes of  $\mathcal{T}$  is thus partitioned into the set  $\mathcal{P}$  of proper nodes, and the set  $\mathcal{H} = \mathcal{N} \setminus \mathcal{P}$  of hanging nodes.

A useful notion in dealing with hanging nodes is the *global index* of a node, introduced in [Beirão da Veiga *et al.* 2023]: it measures the number of nonconforming refinements needed to generate a hanging node from proper nodes. To define it, for any  $x \in \mathcal{H}$  which has been generated by the bisection of an edge  $[x', x'']$ , let us set  $\mathcal{B}(x) = \{x', x''\}$ .

**Definition 3.24 (global index of a node)** The global index  $\lambda(x)$  of a node  $x \in \mathcal{N}$  is defined recursively as follows:

- if  $x \in \mathcal{P}$ , set  $\lambda(x) = 0$ ;



- if  $x \in \mathcal{H}$  and  $\mathcal{B}(x) = \{x', x''\}$ , set  $\lambda(x) = \max(\lambda(x'), \lambda(x'')) + 1$ .

The set of all nodes of  $\mathcal{T}$  is thus partitioned according to the value of the global index: for any integer  $l \geq 0$ , we set  $\mathcal{H}_l = \{x \in \mathcal{N} : \lambda(x) = l\}$ . Note that  $\mathcal{H}_0 = \mathcal{P}$ . An example of distribution of global indices for  $d = 2$  is shown in Figure 3.10.

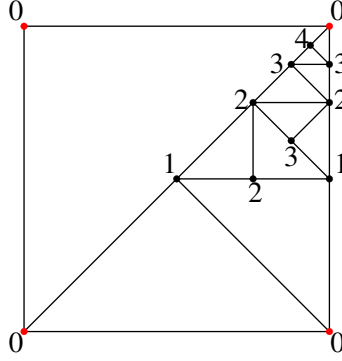


Figure 3.10. Example of distributions of proper nodes (red) and hanging nodes (black), with associated global indices  $\lambda$ .

We define the global index of the triangulation  $\mathcal{T}$  by  $\lambda(\mathcal{T}) := \max_{x \in \mathcal{N}} \lambda(x)$ . The level of nonconformity of the triangulations we are dealing with is controlled by the following condition of admissibility.

**Definition 3.25 ( $\Lambda$ -admissibility)** Let  $\Lambda \geq 0$  be an integer. A refinement  $\mathcal{T}$  of  $\mathcal{T}_0$  is  $\Lambda$ -admissible if

$$\lambda(\mathcal{T}) \leq \Lambda. \quad (3.47)$$

If  $\lambda(\mathcal{T}) \geq 1$ , then  $\mathcal{T}$  is nonconforming, but otherwise  $\mathcal{T} \in \mathbb{T}$  is conforming if  $\lambda(\mathcal{T}) = 0$ . The collection of all  $\Lambda$ -admissible partitions is denoted by  $\mathbb{T}^\Lambda$ .

$\Lambda$ -admissibility has the following basic implications.

**Proposition 3.26 (properties of  $\Lambda$ -admissible partitions)** *Let  $T$  be any element of a  $\Lambda$ -admissible partition  $\mathcal{T}$ .*

- (i) *If  $e \subset \partial T$  is an edge of  $T$ , then  $e$  may contain at most  $2^\Lambda - 1$  hanging nodes.*
- (ii) *If  $e \subset \partial T$  is an edge of some other element  $T'$ , then  $h_{T'} \simeq h_T$ , where the hidden constants only depend on the shape of the initial triangulation  $\mathcal{T}_0$  and possibly on  $\Lambda$ .*

*Proof.* (i) stems from the fact that the edge may contain at most  $2^{k-1}$  hanging nodes of level  $k$  for  $1 \leq k \leq \Lambda$ . To prove (ii) we observe that the length ratio  $\frac{|\bar{e}|}{|e|}$ , where  $\bar{e}$  is the edge of  $T$  containing  $e$ , is at most  $2^\Lambda$ , and we conclude invoking the shape regularity of the partition.  $\square$

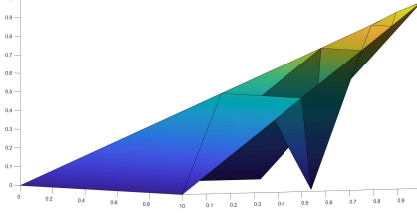


Figure 3.11. Plot of the basis function  $\phi_P$  on the nonconforming triangulation shown in Figure 3.10, for  $P$  equal to the upper right corner of the domain, after using bisection to convert the lowest hanging nodes with global index 3 into a proper node.

In the space  $\mathbb{V}_{\mathcal{T}}$  of continuous piecewise-linear maps over  $\mathcal{T}$ , functions are uniquely defined by their values at the proper nodes of  $\mathcal{T}$ . So it is natural to introduce the canonical continuous piecewise-linear basis functions  $\phi_P$  associated with any proper node  $P$ . They satisfy

$$v = \sum_{P \in \mathcal{P}} v(P) \phi_P \quad \forall v \in \mathbb{V}_{\mathcal{T}}, \quad (3.48)$$

and are defined by the conditions:  $\phi_P \in \mathbb{V}_{\mathcal{T}}$  and

- $\phi_P(z) = 1$  if  $z = P$ ,  $\phi_P(z) = 0$  if  $z \in \mathcal{P} \setminus \{P\}$ .

The values of  $\phi_P$  at the hanging nodes, hence everywhere in the domain, can be reconstructed by linear interpolation as follows: assuming that  $\phi_P$  has been defined at all nodes of global index  $< l$ , if  $z \in \mathcal{H}_l$  and  $\mathcal{B}(z) = \{z', z''\}$ , then

$$\phi_P(z) = \frac{1}{2} (\phi_P(z') + \phi_P(z'')).$$

An example of basis function  $\phi_P$  on a nonconforming triangulation is provided in Figure 3.11.

The *domain of influence* of a proper node  $P$  is the set

$$\omega_{\mathcal{T}}(P) = \text{supp}(\phi_P),$$

highlighted in grey in Figure 3.9; this notion was introduced in [Babuška and Miller 1987] in the context of  $K$ -meshes; see also [Bonito and Nochetto 2010]. To identify elements  $T \in \mathcal{T}$  contained in  $\omega_{\mathcal{T}}(P)$ , we introduce for any node  $x \in \mathcal{N}$  the set  $\mathcal{P}(x)$  of the *proper nodes influencing*  $x$ , which is defined recursively as follows:

- initialize  $\mathcal{P}(x) = \{x\}$ ;
- while  $\mathcal{P}(x) \cap \mathcal{H} \neq \emptyset$ , if  $y \in \mathcal{P}(x) \cap \mathcal{H}$  replace  $\mathcal{P}(x)$  by  $(\mathcal{P}(x) \setminus \{y\}) \cup \mathcal{B}(y)$ .

Then,  $T \subseteq \omega_{\mathcal{T}}(P)$  if and only if  $P$  influences some vertex of  $T$ , i.e., there exists a vertex  $v$  of  $T$  such that  $P \in \mathcal{P}(v)$ .

One of the consequences of the  $\Lambda$ -admissibility assumption of  $\mathcal{T}$  is the following result, which says that all elements  $T$  contained in  $\omega_{\mathcal{T}}(P)$  have comparable size.

**Proposition 3.27 (size of the domain of influence)** *There exists a positive constant  $C = C(\mathcal{T}_0, \Lambda)$ , only depending on the shape of the initial triangulation  $\mathcal{T}_0$  and possibly on  $\Lambda$ , such that for any  $P \in \mathcal{P}$*

$$\text{diam } \omega_{\mathcal{T}}(P) \leq C h_T \quad \forall T \in \mathcal{T}, T \subseteq \omega_{\mathcal{T}}(P).$$

*Proof.* Elements in  $\omega_{\mathcal{T}}(P)$  having  $P$  as a vertex share in pairs an edge or a portion of an edge, hence – as noted above –  $\Lambda$ -admissibility implies the existence of a characteristic size, say  $h_P$ , which is comparable to the diameter of each of them. On the other hand, any  $T \subset \omega_{\mathcal{T}}(P)$  not containing  $P$  has at least one vertex  $v_T \in \mathcal{H}$  such that  $P \in \mathcal{P}(v_T)$ . Thus, there exists a sequence  $\{y_k : 0 \leq k \leq K\}$  of vertices satisfying  $y_0 = v_T$ ,  $y_K = P$ , and  $y_{k+1} \in \mathcal{B}(y_k)$  for  $0 \leq k < K$ ; since  $\lambda(y_k) \geq \lambda(y_{k+1}) + 1$ , necessarily  $K \leq \Lambda$ . Correspondingly, we can find a chain of at most  $K$  elements, starting at  $T$  and ending at an element containing  $P$ , which share in pairs an edge or a portion of an edge. We deduce that  $h_T \simeq h_P$ , and  $\text{dist}(T, P) \simeq h_T$ , where the hidden constants may depend only on the shape of the initial triangulation and on  $\Lambda$ . The conclusion easily follows from these results.  $\square$

We now turn to the case of polynomial degree  $n > 1$ ; we refer to [Canuto and Fazzino 2023] for more details. The concept of hanging node is no longer solely related to the geometry of the mesh, but also to the distribution of degrees of freedom along the edges of the elements. For instance, consider a full edge  $e$  shared by two triangles  $T$  and  $T'$ , and bisect  $T'$  to create two new elements  $T_1$  and  $T_2$  having  $e$  as a vertex. If we use quadratic Lagrangian elements, the midpoint  $x$  of  $e$  carries a degree of freedom for the three elements that share it, so we do not consider it as a hanging node; on the other hand, the nodes at distance  $\frac{1}{4}|e|$  and  $\frac{3}{4}|e|$  from an endpoint of  $e$  are hanging nodes (despite they are vertices of no triangle) since they do not carry a degree of freedom for the element  $T$ . If we move to cubic Lagrangian elements, then  $x$  becomes a hanging node, together with the nodes at distance  $\frac{1}{6}|e|$  and  $\frac{5}{6}|e|$  from an endpoint of  $e$ , whereas the nodes at distance  $\frac{1}{3}|e|$  and  $\frac{2}{3}|e|$  are not hanging nodes, since they carry a degree of freedom for each triangle they belong to.

In general, for a partition  $\mathcal{T}$  made of classical affine Lagrangian or Hermitian elements, the hanging nodes are defined as follows.

- Given an element  $T \in \mathcal{T}$ , the set  $\mathcal{P}_T$  of the *proper nodes of  $T$*  is made of all images of the reference  $n$ -lattice via the affine transformation. The set  $\mathcal{H}_T$  of the *hanging nodes of  $T$*  collects the points of  $\partial T$  that are not proper nodes of  $T$ , but are proper nodes of some other contiguous element  $T'$ . The set of all nodes of  $T$  is  $\mathcal{N}_T := \mathcal{P}_T \cup \mathcal{H}_T$ .
- At the global level, if  $\mathcal{N} = \bigcup_{T \in \mathcal{T}} \mathcal{N}_T$  is the set of all nodes of  $\mathcal{T}$ , the set  $\mathcal{P} \subseteq \mathcal{N}$  of the *proper nodes of  $\mathcal{T}$*  contains those nodes that are proper nodes for all elements they belong to. The complementary set  $\mathcal{H} := \mathcal{N} \setminus \mathcal{P}$  is the set of the *hanging nodes of  $\mathcal{T}$* .

In other words, a hanging node of  $\mathcal{T}$  is a point that carries a degree of freedom

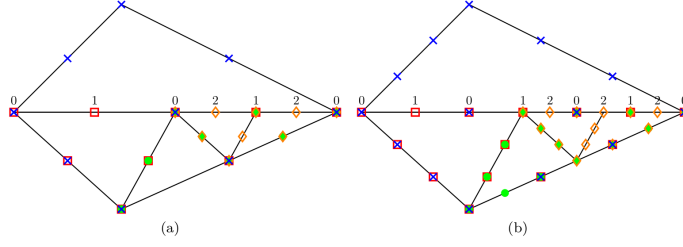


Figure 3.12. Triangulation after the three refinements in the case  $n = 2$  (a) and in the case  $n = 3$  (b). Blue crosses represent the original degrees of freedom on the initial conforming mesh. Red squares, green circles and orange triangles are used for the degrees of freedom of the first, second and third refinement, respectively. All nodes are proper, except those on the horizontal line, whose global index is reported.

for some, but not all elements it belong to. With this definition of proper nodes, representation (3.48) of continuous piecewise-linear maps extends to  $n > 1$ .

The *global index*  $\lambda(x)$  of a node  $x \in \mathcal{N}$  is precisely defined as in Definition 3.24. The set  $\mathcal{B}(x) \subset \mathcal{N}$  collects the endpoints of an interval  $[x', x'']$ , contained in the skeleton of  $\mathcal{T}$ , that has been bisected when  $x$  has been created, and contains no other node inside.

Figure 3.12 provides two examples, for  $n = 2$  and  $n = 3$ , of distributions of hanging nodes and corresponding global indices, created by successive bisections starting from an initial conforming partition.

The concept of  $\Lambda$ -admissibility, given in Definition 3.25, remains unchanged for  $n > 1$ . The statements in Proposition 3.26, too, extend to the higher order case; the maximum number of hanging nodes on an edge being now  $O(n2^\Lambda)$ . Consequently, the conclusion of Proposition 3.27 remains valid when  $n > 1$  as well: there is a constant  $C = C(\mathcal{T}_0, \Lambda)$  such that

$$\text{diam } \omega_{\mathcal{T}}(P) \leq C h_T \quad \forall T \in \mathcal{T}, T \subseteq \omega_{\mathcal{T}}(P). \quad (3.49)$$

**Remark 3.28 (quadrilateral and hexahedral partitions)** It is readily seen that the definitions of global index of a node and  $\Lambda$ -admissible partition extend seamlessly to shape-regular meshes made of quadrilaterals refined by a quadtree strategy ( $d = 2$ ) (see Fig. 3.9 (a) for an example), or by hexahedra refined by an octree strategy ( $d = 3$ ). The same holds for heterogeneous partitions made of a combination of simplices and hexahedra. All results reported above are valid for such partitions. We refer to [Bonito and Nochetto 2010] for details.

*$\Lambda$ -admissible meshes under refinement.* Given a  $\Lambda$ -admissible grid  $\mathcal{T}$ , a subset  $\mathcal{M}$  of elements marked for refinement, and a desired number  $b \geq 1$  of subdivisions to be performed in each marked element, the procedure

$$\mathcal{T}_* = \text{REFINE}(\mathcal{T}, \mathcal{M}, \Lambda)$$

creates a minimal  $\Lambda$ -admissible mesh  $\mathcal{T}_* \geq \mathcal{T}$  such that all the elements of  $\mathcal{M}$  are subdivided at least  $b$  times. In order for  $\mathcal{T}_*$  to be  $\Lambda$ -admissible, perhaps other elements not in  $\mathcal{M}$  must be partitioned. Despite the fact that admissibility is a constraint on the refinement procedure weaker than conformity, it cannot avoid the propagation of refinements beyond  $\mathcal{M}$ . The complexity of REFINE is again an issue which we discuss in §8.2: we show that Theorem 3.16 extends to this case.

**Theorem 3.29 (complexity of REFINE for  $\Lambda$ -admissible meshes)** *Let  $\mathcal{T}_0$  be an arbitrary conforming partition of  $\Omega$ , except for the bisection algorithm in which case  $\mathcal{T}_0$  satisfies the labeling (3.35) for  $d = 2$  or its higher dimensional counterpart [Stevenson 2008]. Then the estimate*

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \leq D \sum_{j=0}^{k-1} \#\mathcal{M}_j \quad \forall k \geq 1$$

*holds with a constant  $D$  depending on  $\mathcal{T}_0$ ,  $d$ ,  $n$  and  $\Lambda$ .*

The following result about uniform refinements of  $\Lambda$ -admissible partitions will be used in the sequel. *The uniform refinement  $\mathcal{T}_*$  of a partition  $\mathcal{T} \in \mathbb{T}^\Lambda$  is the partition obtained by bisecting  $d$  times each element of  $\mathcal{T}$ .* This implies, in particular, that each edge of  $\mathcal{T}$  is bisected once.

**Proposition 3.30 ( $\Lambda$ -admissibility of uniform refinements)** *If  $\mathcal{T} \in \mathbb{T}^\Lambda$  is a  $\Lambda$ -admissible partition and  $\mathcal{T}_*$  is its uniform refinement, then  $\mathcal{T}_*$  is  $\Lambda$ -admissible.*

*Proof.* A simple recursion argument on the global index of the hanging nodes of  $\mathcal{T}$  shows that after refinement each such node either becomes a proper node or its global index is reduced by 1. At the same time, new nodes are created by the refinement, whose global index is at most 1 plus the maximal global index of the pre-existing nodes. In both cases, the maximal global index of  $\mathcal{T}_*$  cannot exceed  $\Lambda$ .  $\square$

A simple consequence is the following result, which is useful to control the meshsize between consecutive refinements.

**Corollary 3.31 (bound on the refinements)** *REFINE with  $b = 1$  never refines an element of a  $\Lambda$ -admissible partition  $\mathcal{T}$  more than  $d$  times.*

*Proof.* REFINE gives the smallest  $\Lambda$ -admissible mesh  $\mathcal{T}_*$  such that all the marked elements of  $\mathcal{T}$  have been refined. Since the uniform refinement of  $\mathcal{T}$  remains  $\Lambda$ -admissible, the minimality of  $\mathcal{T}_*$  implies that no element of the marked set can be refined more than  $d$  times.  $\square$

We conclude by emphasizing that the polynomial interpolation and adaptive

approximation theories of Sects. 3.3 and 3.6 extend to nonconforming meshes with fixed level of incompatibility as well.

#### 4 A Posteriori Error Analysis

Numerical solutions to a boundary value problem serve to approximate its unknown exact solution. In such a context, it is of interest

- to quantify the error of the numerical solution,  $m$
- to gain information for *adapting* the discretization to the exact solution

in a computationally accessible manner. These are the two goals of an *a posteriori error analysis*, where the adjective ‘a posteriori’ hints to the fact that the numerical solution itself can be involved. To achieve the two goals, the a posteriori analysis individuates so-called *error estimators* that, ideally, are computable, split into local contributions called *indicators*, and bound the error from above and below.

This section exemplifies such an analysis, considering the numerical solution of the boundary value problem (2.5), i.e.

$$-\operatorname{div}(A\nabla u) + cu = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega,$$

with Lagrange elements of arbitrary fixed order  $n \geq 1$ . Throughout this section, we adopt the notations and assumptions of previous sections for this model setting. In particular, the *exact solution*  $u \in H_0^1(\Omega)$  solves the variational problem (2.7) and, given a simplicial conforming mesh  $\mathcal{T} \in \mathbb{T}$  of  $\Omega$  and finite element space

$$\mathbb{V}_{\mathcal{T}} := \{v \in \mathbb{S}_{\mathcal{T}}^{n,0} \mid v|_{\Omega} = 0\} \subset H_0^1(\Omega),$$

the *Galerkin approximation* solves

$$u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}} : \quad \mathcal{B}[u_{\mathcal{T}}, w] = \langle f, w \rangle \quad \forall w \in \mathbb{V}_{\mathcal{T}}, \quad (4.1)$$

with the bilinear form  $\mathcal{B}$  from (2.8).

We stress that the analysis will be conducted under the regularity assumptions

$$A \in L^\infty(\Omega; \mathbb{R}^{d \times d}), \quad c \in L^\infty(\Omega), \quad f \in H^{-1}(\Omega) = H_0^1(\Omega)^*, \quad (4.2)$$

used in Sect. 2.4 to establish existence and uniqueness of the exact solution. This fact distinguishes the approach below, which builds on [Kreuzer and Veeser 2019], from most other ones requiring additional regularity; cf., e.g., [Verfürth 2013]. Notably, this difference not only allows for covering more examples but is also related to strengthening the relationship between error and estimator to a true equivalence on any admissible mesh  $\mathcal{T} \in \mathbb{T}$ .

It is useful to recall two differences between the forcing  $f$  and the coefficients  $(A, c)$ . First, while the exact solution  $u$  depends linearly on the forcing  $f$ , it depends nonlinearly on the diffusion tensor  $A$  and on the reaction coefficient  $c$ . To state the second difference, let  $u \in H_0^1(\Omega)$  and note that the assumptions (4.2) on  $(A, c)$  imply the “missing” one  $f \in H^{-1}(\Omega)$ . On the other hand, the assumptions on  $(A, f)$

imply only  $cu \in H^{-1}(\Omega)$  while the assumptions on  $(c, f)$  imply only  $-\operatorname{div}(A\nabla u) \in H^{-1}(\Omega)$ . These conditions are weaker than the “missing” one  $u \in H_0^1(\Omega)$ , and are due to the multiplicative role of  $(A, c)$  in the differential equation.

In order to elucidate the new twists allowing for (4.2), this section is organized as follows. We start with steps of the a posteriori analysis that are common to the ‘classical’ and the new approach. We then illustrate the classical approach with the *standard residual estimator*, and afterwards develop the new approach resulting in a modification of the standard residual estimator, called *modified residual estimator*. Finally, we conclude by adapting the new approach to other techniques of a posteriori error estimation and boundary conditions.

In what follows, the notation may or may not indicate the dependencies of a given quantity. We shall balance readability and the importance of the dependence in the given context. For example, in (4.1), the discrete solution depends not only on the mesh  $\mathcal{T}$  but also on the data  $\Omega, A, c$ , and  $f$  in problem (2.5). We write  $\mathcal{T}$  explicitly because of its more prominent role in the a posteriori analysis. Let  $\mathcal{F} := \mathcal{F}_{\mathcal{T}}$  denote the set of all interior  $(d - 1)$ -dimensional faces of  $\mathcal{T}$ . The letter  $C$  will be used for a generic constant, with possibly different values at each occurrence. If not stated otherwise, it may depend on the shape regularity coefficient  $\sigma$  from (3.9), the dimension  $d$ , and the polynomial degree  $n$  in  $\mathbb{V}_{\mathcal{T}}$ .

#### 4.1 Error, residual and localization of residual norm

This section starts the a posteriori analysis by establishing that a suitable norm of the so-called *residual*

- is equivalent to the error  $\|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)}$  and
- admits a *localization* in the sense that it splits into suitable local contributions depending on accessible quantities, i.e. on data  $\mathcal{D} = (A, c, f)$  and the discrete solution  $u_{\mathcal{T}}$ .

We do not consider computability yet — this important aspect will be addressed in the following sections.

Replacing in the weak form (2.7) the exact solution  $u$  by its approximation  $u_{\mathcal{T}}$ , we define the *residual*  $R_{\mathcal{T}} \in H^{-1}(\Omega)$ :

$$\langle R_{\mathcal{T}}, w \rangle = \langle f, w \rangle - \mathcal{B}[u_{\mathcal{T}}, w] \quad \forall w \in H_0^1(\Omega).$$

We thus have a quantity that depends only on data  $\mathcal{D}$  and the approximate solution  $u_{\mathcal{T}}$  and relates to the error function  $u - u_{\mathcal{T}}$  as follows:

$$\langle R_{\mathcal{T}}, w \rangle = \mathcal{B}[u - u_{\mathcal{T}}, w] \quad \forall w \in H_0^1(\Omega). \quad (4.3)$$

Continuity and coercivity of the bilinear form  $\mathcal{B}$  then provide a quantitative relationship between error and residual.

**Lemma 4.1 (error and residual)** *The error of the approximation  $u_{\mathcal{T}}$  is equivalent*

to the residual norm. More precisely,

$$\frac{1}{\|\mathcal{B}\|} \|R_{\mathcal{T}}\|_{H^{-1}(\Omega)} \leq \|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)} \leq \frac{1}{\alpha} \|R_{\mathcal{T}}\|_{H^{-1}(\Omega)},$$

where  $\|\mathcal{B}\| \geq \alpha > 0$  are, respectively, the continuity and coercivity constants of the bilinear form  $\mathcal{B}$ .

*Proof.* The error-residual relationship (4.3) yields the lower bound,

$$\begin{aligned} \|R_{\mathcal{T}}\|_{H^{-1}(\Omega)} &= \sup_{w \in H_0^1(\Omega)} \frac{\langle R_{\mathcal{T}}, w \rangle}{\|\nabla w\|_{L^2(\Omega)}} = \sup_{w \in H_0^1(\Omega)} \frac{\mathcal{B}[u - u_{\mathcal{T}}, w]}{\|\nabla w\|_{L^2(\Omega)}} \\ &\leq \|\mathcal{B}\| \|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)}, \end{aligned}$$

while the choice  $w = u - u_{\mathcal{T}}$  therein gives

$$\begin{aligned} \alpha \|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)}^2 &\leq \mathcal{B}[u - u_{\mathcal{T}}, u - u_{\mathcal{T}}] = \langle R_{\mathcal{T}}, u - u_{\mathcal{T}} \rangle \\ &\leq \|R_{\mathcal{T}}\|_{H^{-1}(\Omega)} \|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)} \end{aligned}$$

and thus the upper bound.  $\square$

**Remark 4.2 (role of forcing vs role of coefficients)** In addition to the two differences between right-hand side  $f$  and coefficients  $(A, c)$  mentioned in the introduction of this section, a third one implicitly arises in the proof of Lemma 4.1: the coefficients defining the bilinear form  $\mathcal{B}$  are fixed, while the right-hand side  $f$  is replaced by the residual  $R_{\mathcal{T}}$  in (4.3), which varies with the mesh  $\mathcal{T}$ .

**Remark 4.3 (local lower estimate for the error)** The proof of Lemma 4.1 shows that the lower bound of the error hinges on the continuity of the bilinear form  $\mathcal{B}$ . Since the evaluation of  $\mathcal{B}$  involves only local operators, one might expect that there are also local lower bounds. This however depends on the interplay of the underlying differential operator and the choice of the test space norm. Indeed, in the case of the Poisson problem, i.e.  $A = I$ ,  $c = 0$ , and the test space norm  $\|\nabla \cdot\|_{L^2(\Omega)}$ , one easily sees that

$$\|R_{\mathcal{T}}\|_{H^{-1}(\omega)} \leq \|\nabla(u - u_{\mathcal{T}})\|_{L^2(\omega)}$$

for any subdomain  $\omega \subset \Omega$ . This local lower bound however does not carry over to the general case with  $c \neq 0$  as the error function itself is bounded by its gradient only through the global inequality in Lemma 2.2 (first Poincaré inequality). On the other side, endowing the test space  $H_0^1(\Omega)$  with the full  $H^1$ -norm  $\|\cdot\|_{H^1(\Omega)}$  yields

$$\|R_{\mathcal{T}}\|_{(H^1(\omega))^*} \leq \max \{\alpha_1, \|c\|_{L^\infty(\Omega)}\} \|u - u_{\mathcal{T}}\|_{H^1(\omega)}$$

for any subdomain  $\omega \subset \Omega$ .

In line with [Carstensen *et al.* 2014], we shall not invoke local lower bounds to derive convergence and rate optimality for the error of AFEM, although they might appear useful or even crucial in other settings.



**Remark 4.4 (constants in error-residual relationship)** Given the norm measuring the error, that is the norm of the trial space, the choice of the test space norm is important for the ensuing constants in the error-residual relationship; see, e.g., [Verfürth 2013, Sects. 4.3 and 4.6]. To avoid related additional technicalities and difficulties, the test space is endowed with the straightforward norm  $\|\nabla \cdot\|_{L^2(\Omega)}$ .

Lemma 4.1 establishes the first goal that was set out at the beginning of this section. We now turn to the second one, that is, we split the residual norm  $\|R_{\mathcal{T}}\|_{H^{-1}(\Omega)}$  into local contributions. Note that the nature of the dual norm  $\|\cdot\|_{H^{-1}(\Omega)}$  makes this task less obvious than for integral norms as in the error  $\|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)}$ .

We start by recalling that the definition of the Galerkin approximation  $u_{\mathcal{T}}$  implies that its residual is orthogonal to the discrete trial space  $\mathbb{V}_{\mathcal{T}} = \mathbb{S}_{\mathcal{T}}^{n,0} \cap H_0^1(\Omega)$ :

$$\langle R_{\mathcal{T}}, w \rangle = 0 \quad \forall w \in \mathbb{V}_{\mathcal{T}}.$$

Denote by  $\mathcal{V}$  the set of vertices of  $\mathcal{T}$  and by  $\phi_z \in \mathbb{S}_{\mathcal{T}}^{1,0}$  the hat function with  $\phi_z(y) = \delta_{yz}$  for all vertices  $y \in \mathcal{V}$ . In what follows, the *partial* orthogonality

$$\langle R_{\mathcal{T}}, \phi_z \rangle = 0 \quad \forall z \in \mathcal{V} \quad (4.4)$$

will be crucial to split the nonlocal norm of the residual into local contributions. The latter ones will be formulated in terms of the supports of the hat functions and, thus, we associate to each vertex  $z \in \mathcal{V}$  the following subset and submesh:

$$\omega_z := \text{supp } \phi_z = \bigcup_{T \in \mathcal{T}_z} T \quad \text{with} \quad \mathcal{T}_z := \{T \in \mathcal{T} \mid T \ni z\}. \quad (4.5)$$

These subsets, called *stars*, form a *subdomain covering* of  $\Omega$ , viz. each interior  $\omega_z$  is a domain and  $\bar{\Omega} = \bigcup_{z \in \mathcal{V}} \bar{\omega}_z$ . The *overlapping index*  $\text{esssup}_{x \in \Omega} \#\{z \in \mathcal{V} \mid \omega_z \ni x\}$  of this covering is bounded by  $(d + 1)$ .

**Lemma 4.5 (localization of  $H^{-1}$ -norm)** *Let  $\ell \in H^{-1}(\Omega)$  be an arbitrary linear functional on  $H_0^1(\Omega)$ .*

(i) *If  $\langle \ell, \phi_z \rangle = 0$  for all interior vertices  $z \in \mathcal{V} \cap \Omega$ , then*

$$\|\ell\|_{H^{-1}(\Omega)}^2 \leq (d + 1) C_{\text{loc}}^2 \sum_{z \in \mathcal{V}} \|\ell\|_{H^{-1}(\omega_z)}^2$$

*where  $C_{\text{loc}}$  depends only on the shape regularity coefficient  $\sigma$  from (3.9) and  $d$ .*

(ii) *For any subdomain covering  $(\omega_i)_{i \in I}$  of  $\Omega$  with finite overlapping index  $C_{\text{ovrl}} := \text{esssup}_{x \in \Omega} \#\{i \in I \mid \omega_i \ni x\}$ , we have*

$$\sum_{i \in I} \|\ell\|_{H^{-1}(\omega_i)}^2 \leq C_{\text{ovrl}} \|\ell\|_{H^{-1}(\Omega)}^2.$$

[Blechta, Málek and Vohralík 2020, Theorem 3.5] generalizes Lemma 4.5 to the  $W_p^{-1}$ -norm,  $1 < p < \infty$ . Lemma 4.66 below provides an alternative localization with different local norms.

*Proof.* [1] We start by showing statement (i). Thanks to the orthogonality of  $\ell$ , we may write

$$\langle \ell, w \rangle = \langle \ell, w - \sum_{z \in \mathcal{V}} c_z \phi_z \rangle,$$

where any  $c_z$  is an arbitrary constant if  $z \in \mathcal{V} \cap \Omega$  is an interior vertex and 0 if  $z \in \mathcal{V} \cap \partial\Omega$  is a boundary vertex. Using the partition of unity  $\sum_{z \in \mathcal{V}} \phi_z = 1$  on  $\Omega$ , we split the new test function

$$w - \sum_{z \in \mathcal{V}} c_z \phi_z = \sum_{z \in \mathcal{V}} (w - c_z) \phi_z,$$

into local contributions  $(w - c_z) \phi_z \in H_0^1(\omega_z)$ ,  $z \in \mathcal{V}$ . The constant  $c_z$  allows us to counter the gradient generated by the cut-off with  $\phi_z$ . Indeed, the product rule,  $0 \leq \phi_z \leq 1$ , and  $|\nabla \phi_z| \leq C(d)\sigma h_T^{-1}$  on an element  $T \in \mathcal{T}$  lead to

$$\begin{aligned} \|\nabla((w - c_z) \phi_z)\|_{L^2(\omega_z)} &\leq \|\phi_z \nabla(w - c_z)\|_{L^2(\omega_z)} + \|(w - c_z) \nabla \phi_z\|_{L^2(\omega_z)} \\ &\leq \|\phi_z\|_{L^\infty(\omega_z)} \|\nabla w\|_{L^2(\omega_z)} + \|\nabla \phi_z\|_{L^\infty(\omega_z)} \|w - c_z\|_{L^2(\omega_z)} \\ &\leq \|\nabla w\|_{L^2(\omega_z)} + C(d)\sigma \left( \max_{T \subset \omega_z} h_T^{-1} \right) \|w - c_z\|_{L^2(\omega_z)}. \end{aligned}$$

If we choose  $c_z = \int_{\omega_z} w$  for interior vertices  $z \in \mathcal{V} \cap \Omega$ , then Lemma 2.3 (second Poincaré inequality) on reference stars implies

$$\|w - c_z\|_{L^2(\omega_z)} \lesssim \text{diam } \omega_z \|\nabla w\|_{L^2(\omega_z)}.$$

The same inequality follows for boundary vertices  $z \in \mathcal{V} \cap \partial\Omega$  thanks to the fact that  $w$  vanishes at least on one face of  $\partial\omega_z \cap \partial\Omega$ . Combing this with  $\text{diam } \omega_z \lesssim h_T$  for  $T \subset \omega_z$ , we thus obtain, for all local contributions, the stability bound

$$\|\nabla((w - c_z) \phi_z)\|_{L^2(\omega_z)} \leq C_{\text{loc}} \|\nabla w\|_{L^2(\omega_z)}, \quad (4.6)$$

where the constant  $C_{\text{loc}}$  depends only on  $d$  and  $\sigma$ . Hence,

$$\langle \ell, w \rangle = \langle \ell, w - \sum_{z \in \mathcal{V}} c_z \phi_z \rangle = \sum_{z \in \mathcal{V}} \langle \ell, (w - c_z) \phi_z \rangle$$

gives

$$\begin{aligned} |\langle \ell, w \rangle| &\leq C_{\text{loc}} \sum_{z \in \mathcal{V}} \|\ell\|_{H^{-1}(\omega_z)} \|\nabla w\|_{L^2(\omega_z)} \\ &\leq \sqrt{d+1} C_{\text{loc}} \left( \sum_{z \in \mathcal{V}} \|\ell\|_{H^{-1}(\omega_z)}^2 \right)^{1/2} \|\nabla w\|_{L^2(\Omega)} \end{aligned}$$

and (i) is proven.

[2] We verify statement (ii). For each index  $i \in I$ , define  $v_i \in H_0^1(\omega_i) \subset H_0^1(\Omega)$

by

$$\int_{\omega_i} \nabla v_i \cdot \nabla w = \langle \ell, w \rangle, \quad \forall w \in H_0^1(\omega_i).$$

We obtain  $\langle \ell, v_i \rangle = \|\nabla v_i\|_{L^2(\omega_i)}^2 = \|\ell\|_{H^{-1}(\omega_z)}^2$  by arguments similar to the ones in the proof of Lemma 4.1 (error and residual). The sum  $v := \sum_{i \in I} v_i$  is in  $H_0^1(\Omega)$  with

$$\begin{aligned} \|\nabla v\|_{L^2(\Omega)}^2 &\leq \int_{\Omega} \left| \sum_{i \in I_x} \nabla v_i(x) \right|^2 dx \leq \int_{\Omega} \#I_x \sum_{i \in I_x} |\nabla v_i(x)|^2 dx \\ &\leq C_{\text{ovrl}} \sum_{i \in I} \|\nabla v_i\|_{L^2(\omega_i)}^2 = C_{\text{ovrl}} \sum_{i \in I} \|\ell\|_{H^{-1}(\omega_z)}^2, \end{aligned}$$

where we denote the set of active indices in  $x \in \Omega$  by  $I_x := \{i \in I \mid \omega_i \ni x\}$ . Inserting this in

$$\sum_{i \in I} \|\ell\|_{H^{-1}(\omega_z)}^2 = \sum_{i \in I} \langle \ell, v_i \rangle = \langle \ell, v \rangle \leq \|\ell\|_{H^{-1}(\omega_z)} \|\nabla v\|_{L^2(\Omega)}$$

establishes the desired inequality.  $\square$

Thanks to the partial orthogonality (4.4) and the properties of the star covering  $\omega_z, z \in \mathcal{V}$ , we readily obtain the following statement.

**Corollary 4.6 (star localization of residual norm)** *The  $H^{-1}$ -norm of the residual can be split into local contributions on stars:*

$$\frac{1}{d+1} \sum_{z \in \mathcal{V}} \|R_{\mathcal{T}}\|_{H^{-1}(\omega_z)}^2 \leq \|R_{\mathcal{T}}\|_{H^{-1}(\Omega)}^2 \leq (d+1)C_{\text{loc}} \sum_{z \in \mathcal{V}} \|R_{\mathcal{T}}\|_{H^{-1}(\omega_z)}^2,$$

where  $C_{\text{loc}}$  depends only on  $d$  and the shape regularity coefficient  $\sigma$ .

The upper bound of the global residual norm in Corollary 4.6 employs the stars  $\omega_z, z \in \mathcal{V}$ , as local domains. The next remark assesses this choice by discussing conceivable alternatives in terms of elements and domains of the type

$$\omega_F := \bigcup_{T \in \mathcal{T}_F} T \quad \text{with} \quad \mathcal{T}_F := \{T \in \mathcal{T} \mid T \supset F\}, \quad (4.7)$$

where  $F \in \mathcal{F}$  is an interior face of  $\mathcal{T}$ .

**Remark 4.7 (star localization is minimal for  $d \geq 2$ )** The use of stars in the localization of the global residual norm is a sort of minimal choice, except for the special case  $d = 1$  where elements can be used:

- If  $d = 1$ , point values are defined for functions in  $H^1(\Omega)$ . This allows an upper bound with elements instead of stars as local domains. In fact, choosing  $c_z = w(z)$  for all interval endpoints, the function  $\sum_{z \in \mathcal{V}} c_z \phi_z$  amounts to the Lagrange interpolant  $I_{\mathcal{T}} w \in \mathbb{S}_{\mathcal{T}}^{1,0} \cap H_0^1(\Omega)$ , and we have  $(w - I_{\mathcal{T}} w)|_I \in H_0^1(I)$

with  $\|\nabla(w - I_{\mathcal{T}}w)\|_{L^2(I)} \lesssim \|\nabla w\|_{L^2(I)}$  for any interval  $I$  of the mesh  $\mathcal{T}$ . Arguing as in the proof of Lemma 4.5 (i) then gives

$$\|R\|_{H^{-1}(\Omega)}^2 \lesssim \sum_{I \in \mathcal{T}} \|R\|_{H^{-1}(I)}^2.$$

- An upper bound where the stars are replaced by elements  $T \in \mathcal{T}$  cannot hold in general because it does not account for face-supported residual contributions. For example, consider our setting with

$$d \geq 2, \quad A = I, \quad c = 0, \quad \langle f, w \rangle = \int_F q w, \quad w \in H_0^1(\Omega)$$

where  $F \in \mathcal{F}$  and  $q \neq 0$  is  $L^2$ -orthogonal to  $\mathbb{P}_n(F)$ .

Then we have  $u \neq 0 = u_{\mathcal{T}}$  and therefore  $\|R_{\mathcal{T}}\|_{H^{-1}(\Omega)} > 0$  but  $\|R_{\mathcal{T}}\|_{H^{-1}(T)} = 0$  for any  $T \in \mathcal{T}$ .

- An upper bound with pairs  $\omega_F, F \in \mathcal{F}$ , instead of stars cannot hold in general. This can be shown by considering our setting with

$$d = 2, \quad A = I, \quad c = 0, \quad \langle f_{\varepsilon}, w \rangle = \frac{1}{\pi \varepsilon^2} \int_{B_{\varepsilon}(z)} w, \quad w \in H_0^1(\Omega)$$

where  $z \in \mathcal{V}$  is a vertex of a suitable triangulation  $\mathcal{T}$

for  $\varepsilon \searrow 0$ . The limiting right-hand side is the Dirac measure in  $z$ , which, formally, is not seen by any  $\|\cdot\|_{H^{-1}(\omega_F)}, F \in \mathcal{F}$ . We thus have  $\|R_{\mathcal{T}}\|_{H^{-1}(\Omega)} \rightarrow \infty$  but  $\sum_{F \in \mathcal{F}} \|R_{\mathcal{T}}\|_{H^{-1}(\omega_F)} \lesssim 1$ ; cf. [Tantardini, Veeder and Verfürth n.d.].

The bisection method for mesh refinement is element-oriented. It is therefore advantageous to dispose of an element-indexed reformulation of the localization in Corollary 4.6. For that purpose, we recall the notion of *patches*

$$\omega_T := \bigcup_{\substack{T' \in \mathcal{T} \\ T' \cap T \neq \emptyset}} T' \quad (4.8)$$

and may use the following equivalence.

**Lemma 4.8 (localization re-indexing)** *For any functional  $\ell \in H^{-1}(\Omega)$ , we have*

$$\sum_{z \in \mathcal{V}} \|\ell\|_{H^{-1}(\omega_z)}^2 \approx \sum_{T \in \mathcal{T}} \|\ell\|_{H^{-1}(\omega_T)}^2,$$

where the hidden constants depend on  $d$  and the shape regularity coefficient  $\sigma$ .

Note that, in contrast to the localization itself, its re-indexing does not require any orthogonality like  $\langle \ell, \phi_z \rangle = 0$  for all  $z \in \mathcal{V} \cap \Omega$ .

*Proof.* For any vertex  $z \in \mathcal{V}$ , there is an element  $T \in \mathcal{T}$  containing  $z$ . Then the inclusion  $\omega_z \subset \omega_T$  yields the inequality  $\|\ell\|_{H^{-1}(\omega_z)} \leq \|\ell\|_{H^{-1}(\omega_T)}$ . Hence,

$$\sum_{z \in \mathcal{V}} \|\ell\|_{H^{-1}(\omega_z)}^2 \leq \sum_{T \in \mathcal{T}} \|\ell\|_{H^{-1}(\omega_T)}^2.$$

To show the converse inequality, let  $T \in \mathcal{T}$  be any element and  $w \in H_0^1(\omega_T)$ . Given any vertex  $z \in \mathcal{V} \cap \omega_T$ , Lemma 2.2 (first Poincaré inequality) on  $\omega_T$  implies the stability bound

$$\begin{aligned} \|\nabla(w\phi_z)\|_{L^2(\omega_z \cap \omega_T)} &\leq \|\phi_z \nabla w\|_{L^2(\omega_z \cap \omega_T)} + \|w \nabla \phi_z\|_{L^2(\omega_z \cap \omega_T)} \\ &\leq \|\nabla w\|_{L^2(\omega_z \cap \omega_T)} + C(d, \sigma) \max_{T \subset \omega_z \cap \omega_T} h_T^{-1} \|w\|_{L^2(\omega_T)} \\ &\lesssim \|\nabla w\|_{L^2(\omega_T)}. \end{aligned}$$

We thus derive

$$\begin{aligned} \langle \ell, w \rangle &= \sum_{z \in \mathcal{V} \cap \omega_T} \langle \ell, w\phi_z \rangle \leq \sum_{z \in \mathcal{V} \cap \omega_T} \|\ell\|_{H^{-1}(\omega_z \cap \omega_T)} \|\nabla(w\phi_z)\|_{L^2(\omega_z \cap \omega_T)} \\ &\lesssim \left( \sum_{z \in \mathcal{V} \cap \omega_T} \|\ell\|_{H^{-1}(\omega_z)} \right) \|\nabla w\|_{L^2(\omega_T)} \end{aligned}$$

and, since  $\#(\mathcal{V} \cap \omega_T)$  is bounded in terms of the shape regularity coefficient  $\sigma$ ,

$$\|\ell\|_{H^{-1}(\omega_T)}^2 \lesssim \sum_{z \in \mathcal{V} \cap \omega_T} \|\ell\|_{H^{-1}(\omega_z)}^2.$$

Summing over  $T \in \mathcal{T}$  and taking into account that  $\#\{T \in \mathcal{T} \mid \omega_T \ni z\}$  is again bounded in terms of  $\sigma$ , concludes the proof.  $\square$

## 4.2 Standard residual estimator and its flaws

Exploiting the results of Sect. 4.1, we derive an *a posteriori upper bound* of the error in terms of the standard residual estimator and discuss its flawed sharpness. This discussion will serve as the starting point for an improved *a posteriori* analysis in the following sections.

The *standard residual estimator* needs the additional regularity

$$f \in L^2(\Omega) \quad \text{and} \quad \mathbf{A} \in W_\infty^1(\Omega; \mathbb{R}^{d \times d}) \quad (4.9)$$

for the data in our model problem (2.5). Given the Galerkin approximation  $u_{\mathcal{T}}$  from (4.1), it may be defined as follows, cf., e.g., [Verfürth 2013]:

$$\mathcal{E}_{\mathcal{T}}^{\text{std}} := \mathcal{E}_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, \mathcal{D}) := \left( \sum_{T \in \mathcal{T}} \mathcal{E}_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, \mathcal{D}, T)^2 \right)^{\frac{1}{2}} \quad (4.10a)$$

with the *local indicators*

$$\mathcal{E}_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, \mathcal{D}, T)^2 := h_T \|j(u_{\mathcal{T}})\|_{L^2(\partial T \setminus \partial\Omega)}^2 + h_T^2 \|r(u_{\mathcal{T}})\|_{L^2(T)}^2, \quad (4.10b)$$

where

- the *scaling factor*  $h_T = |T|^{1/d}$  measures the size of the element  $T \in \mathcal{T}$ ,

- $j(v) = j_{\mathcal{T}}(v)$  is the *jump residual* given face-wise for  $v \in \mathbb{V}_{\mathcal{T}}$  by

$$\begin{aligned} j(v)|_F &:= ([[\mathbf{A}\nabla v]] \cdot \mathbf{n}_{T_1})|_F := ((\mathbf{A}\nabla v)|_{T_1} - (\mathbf{A}\nabla v)|_{T_2}) \cdot \mathbf{n}_{T_1} \\ &= (\mathbf{A}\nabla v)|_{T_1} \cdot \mathbf{n}_{T_1} + (\mathbf{A}\nabla v)|_{T_2} \cdot \mathbf{n}_{T_2} \end{aligned}$$

where  $F \in \mathcal{F}$ ,  $T_1, T_2 \in \mathcal{T}$  are such that  $F = T_1 \cap T_2$ ,  $\mathbf{n}_{T_i}$  denotes the outer normal of  $\partial T_i$ ,  $i = 1, 2$ , and

- $r(v) = r_{\mathcal{T}}(v)$  is the *element residual*, a function given for  $v \in \mathbb{V}_{\mathcal{T}}$  by

$$r(v)|_T := (f - cv + \operatorname{div}(\mathbf{A}\nabla v))|_T$$

on any element  $T \in \mathcal{T}$ .

Note that the definition itself already uses the extra regularity (4.9). For notational simplicity, we shall write  $j$  and  $r$  instead of  $j(u_{\mathcal{T}})$  and  $r(u_{\mathcal{T}})$  for the rest of this section. Also, for any interior face  $F \in \mathcal{F}$ , we have  $F = T_1 \cap T_2$  with  $T_1, T_2 \in \mathcal{T}$ . If  $\mathbf{n}_{T_i}$  denotes the outer normal of  $\partial T_i$ ,  $i = 1, 2$ , we set  $\mathbf{n}_F = \mathbf{n}_{T_1}$ . This particular choice of  $\mathbf{n}_F$  is irrelevant as it does not affect the following definition of normal jump of any vector-valued field  $\mathbf{g}$  with well defined trace on  $F$

$$[[\mathbf{g}]] \cdot \mathbf{n}_F := \mathbf{g}|_{T_1} \cdot \mathbf{n}_{T_1} + \mathbf{g}|_{T_2} \cdot \mathbf{n}_{T_2}.$$

**Theorem 4.9 (upper bound with standard residual estimator)** *Suppose the additional regularity (4.9) holds. Then the error is bounded by the standard residual estimator:*

$$\|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)} \lesssim \mathcal{E}_{\mathcal{T}}^{\text{std}},$$

where the hidden constant depends on the coefficients  $(\mathbf{A}, c)$ , the shape regularity coefficient  $\sigma$ , and  $d$ .

*Proof.* As Lemma 4.1 (error and residual) and Corollary 4.6 (star localization of residual norm) imply

$$\|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)}^2 \lesssim \|R_{\mathcal{T}}\|_{H^{-1}(\Omega)}^2 \lesssim \sum_{z \in \mathcal{V}} \|R_{\mathcal{T}}\|_{H^{-1}(\omega_z)}^2$$

and  $\#\{z \in \mathcal{V} \mid \omega_z \supset T\} = d + 1$ , it suffices to establish

$$\|R_{\mathcal{T}}\|_{H^{-1}(\omega_z)}^2 \lesssim \sum_{T \subset \omega_z} \mathcal{E}_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, f, T)^2 \quad (4.11)$$

for any vertex  $z \in \mathcal{V}$ . To this end, let  $w \in H_0^1(\omega_z)$ . The extra regularity  $f \in L^2(\Omega)$  and  $\mathbf{A} \in W_{\infty}^1(\Omega; \mathbb{R}^{d \times d})$  allows for piecewise integration by parts, which leads to the following  $L^2$ -representation of the residual:

$$\langle R_{\mathcal{T}}, w \rangle = \sum_{F \ni z} \int_F jw + \sum_{T \ni z} \int_T rw.$$

In order to bound the right-hand side suitably, we use the scaled trace theorem

$$\|w\|_{L^2(F)}^2 \leq \frac{|F|}{|T|} \|w\|_{L^2(T)}^2 + \frac{2}{d} \frac{|F| \operatorname{diam} T}{|T|} \|w\|_{L^2(T)} \|\nabla w\|_{L^2(T)} \quad (4.12)$$

for any face  $F \subset \partial T$ , see e.g. [Veeder and Verfürth 2009, Corollary 4.5], the inequality

$$\|w\|_{L^2(\omega_z)} \leq \operatorname{diam} \omega_z \|\nabla w\|_{L^2(\omega_z)}$$

from Lemma 2.2 (first Poincaré inequality), and the two geometric relationships

$$\operatorname{diam} \omega_z \lesssim h_T \text{ whenever } T \subset \omega_z, \quad |F| \operatorname{diam} T \lesssim |T| \text{ for } F \subset \partial T.$$

We thus obtain

$$|\langle R_{\mathcal{T}}, w \rangle| \lesssim \left( \sum_{T \ni z} h_T \|r\|_{L^2(T)} + h_T^{1/2} \sum_{F \ni z, F \subset T} \|j\|_{L^2(F)} \right) \|\nabla w\|_{L^2(\omega_z)}.$$

As the number of faces and elements in the star  $\omega_z$  is bounded in terms of the shape regularity coefficient  $\sigma$ , we arrive at the desired bound (4.11) and the proof is finished.  $\square$

**Remark 4.10 (alternative derivation of upper bound)** The upper bound for the standard residual estimator in Theorem 4.9 is often derived with a suitable interpolation operator, by-passing the localization of the  $H^{-1}$ -norm in Lemma 4.5. That approach is useful for the proof of Theorem 4.48 below and is presented therein. Here we opted for using the localization of the  $H^{-1}$ -norm in order to facilitate the comparison with the following subsections. The approach at hand is also convenient to keep the ensuing constants small; cf. [Veeder and Verfürth 2009].

An important question is the sharpness of the upper bound in Theorem 4.9. The so-called *a posteriori lower bounds* provide some answer by trying to bound the estimator in terms of the error. For many estimators, there arises however additional terms of *oscillatory* nature. The following remark justifies the presence of such terms for the case at hand.

**Remark 4.11 (nonasymptotic overestimation)** The lower bound

$$\mathcal{E}_{\mathcal{T}}^{\text{std}} \lesssim \|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)},$$

which would imply equivalence of error and estimator, *cannot hold* in general for the following reason.

Fix a mesh  $\mathcal{T}$  and a functional  $f \in H^{-1}(\Omega) \setminus L^2(\Omega)$  and consider a sequence  $(f_n)_n$  of functions in  $L^2(\Omega)$  with  $\lim_{n \rightarrow \infty} \|f - f_n\|_{H^{-1}(\Omega)} = 0$ . Then the sequences  $(u_n)_n$  and  $(u_{\mathcal{T},n})_n$  of exact and Galerkin solutions on a fixed mesh  $\mathcal{T}$  remain bounded. The error sequence  $(\|\nabla(u_n - u_{\mathcal{T},n})\|_{L^2(\Omega)})_n$  is therefore also bounded, while the standard residual estimator  $\mathcal{E}_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T},n}, f_n) \rightarrow \infty$  becomes unbounded. Note that, in the special case  $f = -\operatorname{div}(A \nabla v) + cv$  with  $v \in \mathbb{V}_{\mathcal{T}}$ , we even have for the error  $\lim_{n \rightarrow \infty} \|\nabla(u_n - u_{\mathcal{T},n})\|_{L^2(\Omega)} = 0$ .

In other words: in certain cases, the standard residual estimator bounds almost 0 by almost  $\infty$  and a lower bound has to involve an additional term that cannot be bounded by the error in general.

We shall define these additional terms with the help of the following local best approximations. Let  $K$  be an element or face of  $\mathcal{T}$  and  $m \in \mathbb{N}_0$  a polynomial degree. Given  $v \in L^2(K)$ , denote by  $\Pi_K v := \Pi_K^m v$  the best approximation in  $\mathbb{P}_m(K)$  with respect to the norm  $\|\cdot\|_{L^2(K)}$ . It is convenient to allow also for  $m = -1$  with  $\mathbb{P}_{-1}(K) = \{0\}$  and  $\Pi_K^{(-1)} v = 0$ . Writing  $\mathcal{D} = (A, c, f)$  for the data in problem (2.5), the  $(m_1, m_2)$ -oscillation for the standard residual estimator is then given by

$$\text{osc}_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, \mathcal{D})^2 := \sum_{T \in \mathcal{T}} \text{osc}_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, \mathcal{D}, T)^2, \quad (4.13a)$$

with the local indicators

$$\text{osc}_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, \mathcal{D}, T)^2 := h_T^2 \|r - \Pi_T^{m_2} r\|_{L^2(T)}^2 + h_T \sum_{F \subset \partial T \setminus \partial \Omega} \|j - \Pi_F^{m_1} j\|_{L^2(F)}^2. \quad (4.13b)$$

**Proposition 4.12 (partial lower bound)** *If  $f \in L^2(\Omega)$  and  $A \in W_{\infty}^1(\Omega; \mathbb{R}^{d \times d})$ , the standard residual estimator is bounded by error and oscillation:*

$$\mathcal{E}_{\mathcal{T}}^{\text{std}} \lesssim \|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)} + \text{osc}_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, \mathcal{D}),$$

where the hidden constant depends on  $d$ , the coefficients  $A$  and  $c$ , the shape regularity coefficient  $\sigma$  as well as the oscillation degrees  $(m_1, m_2)$ .

*Proof.* In light of Lemma 4.1 (error and residual) and Corollary 4.6 (localization of residual norm), we may establish the claimed bound by bounding each indicator with a corresponding local residual norm. To this end, we shall consider here only the case of the oscillation degrees  $(m_1, m_2) = (0, 0)$ . The general case can be verified along the same lines with additional technicalities and is treated in the proof of Lemma 4.28 below in a slightly different context.

[1] We start by bounding an arbitrary element residual  $h_T \|r\|_{L^2(T)}$ ,  $T \in \mathcal{T}$ , in terms of some local residual norm. To this end, we may try to invert the following consequence of Lemma 2.2 (first Poincaré inequality):

$$\|R_{\mathcal{T}}\|_{H^{-1}(T)} = \sup_{w \in H_0^1(T)} \frac{|\langle R_{\mathcal{T}}, w \rangle|}{\|\nabla w\|_{L^2(T)}} = \sup_{w \in H_0^1(T)} \frac{\int_T r w}{\|\nabla w\|_{L^2(T)}} \lesssim h_T \|r\|_{L^2(T)},$$

the residual norm of which avoids involving the jump residual. We thus actually ask for an equivalence of two different smoothness norms. Such an equivalence can hold only for special  $r$ , e.g., from a finite-dimensional space. Furthermore, writing  $\|r\|_{L^2(T)}^2 = \int_T r(\chi_T)$  suggests the choice  $w = r\chi_T$ , which is however not admissible for the residual  $R_{\mathcal{T}}$  as both  $r$  and the characteristic function  $\chi_T$  do not belong to  $H_0^1(\Omega)$ . We shall overcome these issues by replacing  $r$  with its mean



value  $\Pi_T^0 r$  and  $\chi_T$  with the element bubble

$$\phi_T := (d+1)^{(d+1)} \prod_{z \in \mathcal{V} \cap T} \phi_z. \quad (4.14)$$

Thanks to  $\int_T \phi_T = C_d |T|$  and the inverse estimate  $\|\nabla w\|_{L^2(T)} \lesssim h_T^{-1} \|w\|_{L^2(T)}$  for  $w = (\Pi_T^0 r) \phi_T \in H_0^1(T) \cap \mathbb{P}_{d+1}(T)$ , we derive

$$\begin{aligned} \|\Pi_T^0 r\|_{L^2(T)} &\lesssim \int_T (\Pi_T^0 r) w \leq \|(\Pi_T^0 r) \chi_T\|_{H^{-1}(T)} \|\nabla w\|_{L^2(T)} \\ &\lesssim h_T^{-1} \|(\Pi_T^0 r) \chi_T\|_{H^{-1}(T)} \|w\|_{L^2(T)} \leq h_T^{-1} \|(\Pi_T^0 r) \chi_T\|_{H^{-1}(T)} \|\Pi_T^0 r\|_{L^2(T)}, \end{aligned}$$

whence

$$h_T \|\Pi_T^0 r\|_{L^2(T)} \lesssim \|(\Pi_T^0 r) \chi_T\|_{H^{-1}(T)}. \quad (4.15)$$

This implies the desired partial lower bound for the element residual by a perturbation argument and the inequality  $\|r - \Pi_T^0 r\|_{H^{-1}(T)} \lesssim h_T \|r - \Pi_T^0 r\|_{L^2(T)}$ , which follows from another application of Lemma 2.2 (first Poincaré inequality):

$$\begin{aligned} h_T \|r\|_{L^2(T)} &\leq h_T \|\Pi_T^0 r\|_{L^2(T)} + h_T \|r - \Pi_T^0 r\|_{L^2(T)} \\ &\lesssim \|(\Pi_T^0 r) \chi_T\|_{H^{-1}(T)} + h_T \|r - \Pi_T^0 r\|_{L^2(T)} \\ &\lesssim \|r \chi_T\|_{H^{-1}(T)} + h_T \|r - \Pi_T^0 r\|_{L^2(T)} \\ &= \|R_{\mathcal{T}}\|_{H^{-1}(T)} + h_T \|r - \Pi_T^0 r\|_{L^2(T)}. \end{aligned} \quad (4.16)$$

[2] We bound an arbitrary jump residual  $\|j\|_{L^2(F)}$ ,  $F \in \mathcal{F}$ , in a similar manner. Note that here an interference of the element residual is unavoidable because the support of nontrivial test functions have nonempty interior. We thus may try to invert

$$\|R_{\mathcal{T}}\|_{H^{-1}(\omega_F)} = \sup_{w \in H_0^1(\omega_F)} \frac{\int_F j w + \int_{\omega_F} r w}{\|\nabla w\|_{L^2(\omega_F)}} \lesssim h_F^{\frac{1}{2}} \|j\|_{L^2(F)} + \sum_{T \subset \omega_F} h_T \|r\|_{L^2(T)},$$

where also the scaled trace theorem (4.12) is used. To this end, we write  $\delta_F$  for the Dirac measure of the face  $F$ ,

$$\phi_F := d^d \prod_{z \in \mathcal{V} \cap F} \phi_z \quad (4.17)$$

for the face bubble of  $F$  and choose the test function  $w = (\Pi_F^0 j) \phi_F \in H_0^1(\omega_F)$ .

Using in addition  $\|w\|_{L^2(\omega_F)} \lesssim h_F^{\frac{1}{2}} \|w\|_{L^2(F)}$  and (4.15), we deduce

$$\begin{aligned} \|\Pi_F^0 j\|_{L^2(F)}^2 &\lesssim \int_F (\Pi_F^0 j) w + \sum_{T \subset \omega_F} \int_T (\Pi_T^0 r) w - \sum_{T \subset \omega_F} \int_T (\Pi_T^0 r) w \\ &\leq \left\| (\Pi_F^0 j) \delta_F + \sum_{T \subset \omega_F} (\Pi_T^0 r) \chi_T \right\|_{H^{-1}(\omega_F)} \|\nabla w\|_{L^2(\omega_F)} \end{aligned}$$

$$\begin{aligned}
& + \sum_{T \subset \omega_F} \|\Pi_F^0 r\|_{L^2(T)} \|w\|_{L^2(T)} \\
& \lesssim \left\| (\Pi_F^0 j) \delta_F + \sum_{T \subset \omega_F} (\Pi_T^0 r) \chi_T \right\|_{H^{-1}(\omega_F)} h_F^{-\frac{1}{2}} \|\Pi_F^0 j\|_{L^2(F)},
\end{aligned}$$

whence

$$h_F^{\frac{1}{2}} \|\Pi_F^0 j\|_{L^2(F)}^2 \lesssim \left\| (\Pi_F^0 j) \delta_F + \sum_{T \subset \omega_F} (\Pi_T^0 r) \chi_T \right\|_{H^{-1}(\omega_F)}. \quad (4.18)$$

Passing to the proper jump residual  $j$ , we arrive at the partial lower bound for the jump residual:

$$\begin{aligned}
h_F^{\frac{1}{2}} \|\Pi_F^0 j\|_{L^2(F)} & \lesssim \|R_{\mathcal{T}}\|_{H^{-1}(\omega_F)} \\
& + h_F^{\frac{1}{2}} \|j - \Pi_F^0 j\|_{L^2(F)} + \sum_{T \subset \omega_F} h_T \|r - \Pi_T^0 r\|_{L^2(T)}. \quad (4.19)
\end{aligned}$$

3 We square the bounds (4.16) and (4.19) from the previous steps and sum them, respectively, over all elements and faces to conclude the claimed partial lower bound with the help of Lemma 4.5(ii) (localization of  $H^{-1}$ -norm) and Lemma 4.1 (error and residual).  $\square$

The significance of Proposition 4.12 (partial lower bound) strongly depends on the choice of the polynomial degrees  $(m_1, m_2)$  in the oscillation from (4.13). The following two remarks address this important aspect.

**Remark 4.13 (oscillation degrees - asymptotics)** It is desirable that, under refinement, the oscillation in Proposition 4.12 (partial lower bound) converges to 0 at least as fast as the error. The maximal convergence order of the error under uniform refinement is  $\|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)} = O(h^n)$  as  $h \rightarrow 0$ . In view of the scaling factors and derivative orders appearing in jump and element residual, we are thus led to require

$$m_1 \geq n - 1 \quad \text{and} \quad m_2 \geq n - 2.$$

One might hope that strict inequalities lead to higher order. Note however that, since  $\text{osc}^{\text{std}}$  involves in general both discrete solution  $u_{\mathcal{T}}$  and data  $\mathcal{D} = (\mathbf{A}, c, f)$ , this will not be guaranteed without additional assumptions. Furthermore, increasing  $m_1$  and  $m_2$  entails bigger hidden constants in the lower bounds (4.15) and (4.18), as these bounds cannot hold for arbitrary  $L^2$ -functions. Consequently, a potentially higher asymptotic speed of the oscillation  $\text{osc}^{\text{std}}$  comes with a bigger constant in front of it and, therefore, with diminished non-asymptotic significance.

**Remark 4.14 (oscillation degrees - data oscillation reduction)** In the particular case of the Poisson equation, i.e.  $\mathbf{A} = \mathbf{I}$  and  $c = 0$ , and linear elements, i.e.  $n = 1$ ,

the oscillation with the degrees  $(m_1, m_2) = (0, 0)$  reduces to the *data oscillation*

$$\text{osc}_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, \mathcal{D})^2 = \sum_{T \in \mathcal{T}} h_T^2 \|f - \Pi_T^0 f\|_{L^2(T)}^2;$$

it depends only on the data, here the right-hand side  $f$ . Note also that here the regularity of  $f$  is determined by the regularity of the exact solution  $u$ .

For elements with degree  $n \geq 2$ , the choices  $(m_1, m_2) = (n-1, n-2)$  ensure that for  $F \in \mathcal{F}$

$$\Pi_F^{n-1}([\nabla u_{\mathcal{T}}] |_F \cdot \mathbf{n}_F) = [\nabla u_{\mathcal{T}}] |_F \cdot \mathbf{n}_F \quad \text{and} \quad \Pi_T^{n-2}(\Delta u_{\mathcal{T}}|_T) = \Delta u_{\mathcal{T}}|_T \quad (4.20)$$

and so, again, oscillation reduces to data oscillation in  $f$ :

$$\text{osc}_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, \mathcal{D})^2 = \sum_{T \in \mathcal{T}} h_T^2 \|f - \Pi_T^{n-2} f\|_{L^2(T)}^2.$$

If we add a reaction term, viz. we consider  $\mathbf{A} = \mathbf{I}$  and  $c = 1$ , we can again obtain the reduction to data oscillation by increasing  $m_2$  to  $n$ .

For a more general operator with piecewise polynomial coefficients

$$\mathbf{A} \in (\mathbb{S}_{\mathcal{T}}^{n_A, -1})^{d \times d} \quad \text{and} \quad c \in \mathbb{S}_{\mathcal{T}}^{n_c, -1},$$

the choice

$$(m_1, m_2) = (n_A + n - 1, \max\{n_c + n, n_A + n - 2\}) \quad (4.21)$$

again reduces  $\text{osc}^{\text{std}}$  to data oscillation in  $f$ .

Finally, for a general operator without piecewise polynomial coefficients  $(\mathbf{A}, c)$ , a reduction to data oscillation with piecewise polynomial best approximations as before is not possible. The argument in Remark 4.13 suggests approximating the general coefficients with piecewise polynomial coefficients satisfying

$$n_A = n - 1 \quad \text{and} \quad n_c = n - 1.$$

As we shall see below in Sect. 4.8, the choice (4.21) with these values allows us to bound  $\text{osc}^{\text{std}}$  in terms of  $\|\nabla u_{\mathcal{T}}\|_{L^2(\Omega)}$ , which is controlled by stability, and data oscillation terms involving  $f$  and the coefficients  $\mathbf{A}$  and  $c$ . Note however that the nature of these data oscillation terms differs from the preceding reductions: e.g., the regularity of the coefficients  $\mathbf{A}$  and  $c$  is not determined by the exact solution  $u$ .

In the light of Remark 4.13 (oscillation degrees - asymptotics), one might hope that the overestimation described in Remark 4.11 (nonasymptotic overestimation) disappears under refinement. This can be ensured under suitable regularity assumption but is not guaranteed in general as the following remark reveals.

**Remark 4.15 (asymptotic overestimation)** Considering a variant of the standard residual estimator that allows for  $f \in H^{-1}(\Omega)$  and adaptive refinement, [Cohen, DeVore and Nochetto 2012, Sect. 6.4] give an example where the error converges asymptotically faster than the estimator; see also [Kreuzer and Veeser 2021, Lemma 21].

After having recognized the above flaws of the standard residual estimator, let us conclude with an observation that will be the departure point of an improved analysis.

**Corollary 4.16 (equivalence for discrete data)** *Suppose all data  $\mathcal{D} = (\mathbf{A}, f, c)$  of problem (2.5) are piecewise polynomial, i.e. there are  $n_A, n_c, n_f \in \mathbb{N}_0$  such that*

$$\mathbf{A} \in (\mathbb{S}_{\mathcal{T}}^{n_A, -1})^{d \times d}, \quad c \in \mathbb{S}_{\mathcal{T}}^{n_c, -1}, \quad \text{and} \quad f \in \mathbb{S}_{\mathcal{T}}^{n_f, -1}.$$

*Then error and standard residual estimator are equivalent:*

$$\|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)} \approx \mathcal{E}_{\mathcal{T}}^{\text{std}},$$

*where the hidden constants depend only on  $d$ , the coefficients  $\mathbf{A}$  and  $c$ , the shape regularity coefficient  $\sigma$ , and the degrees  $n_A, n_c$ , and  $n_f$ .*

*Proof.* The upper bound follows from Theorem 4.9 (upper bound with standard residual estimator), while Proposition 4.12 (partial lower bound) with

$$(m_1, m_2) = (n_A + n - 1, \max\{n_A + n - 2, n_c + n, n_f\}).$$

yields the lower bound.  $\square$

Motivated by the above discussion, one may define a variant of the standard residual estimator, characterized by a splitting into two different parts. More precisely, choosing  $(m_1, m_2)$  according to Remark 4.14 (oscillation degrees - data oscillation reduction), one may replace the local indicators in (4.10b) by

$$\mathcal{E}_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, f, T)^2 := \eta_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, T)^2 + \text{osc}_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, \mathcal{D}, T)^2, \quad (4.22a)$$

where the first part, the so-called *PDE indicator*, is given by

$$\eta_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, T)^2 := h_T^2 \|\Pi_T^{m_2} r\|_{L^2(T)}^2 + h_T \sum_{F \in \partial T \setminus \partial \Omega} \|\Pi_F^{m_1} j\|_{L^2(F)}^2, \quad (4.22b)$$

while the second part corresponds to the local oscillation from (4.13b); compare with [Verfürth 2013, Theorems 1.5 and 4.7]. In this way,

- the PDE indicators are computable (in terms of the Galerkin approximation  $u_{\mathcal{T}}$  and the local projections),
- the oscillation indicators typically have to be approximated by numerical quadrature,
- both types of indicators are, in general, not dominated by the error.

### 4.3 Discrete functionals and a posteriori error analysis

This section introduces the notion of *discrete functionals* and individuates properties in their approximation that are useful in a posteriori error analysis. The realization of these properties distinguishes the subsequent approach, which is adapted from [Kreuzer and Veaser 2021] and [Kreuzer et al. n.d.].

The notion of discrete functionals and its local counterparts are of interest for at least two reasons. The first one is that their  $H^{-1}$ -norm can be rather easily quantified, as we shall see in Corollary 4.30 below. This property is related to Corollary 4.16 (equivalence for discrete data), which can be read in the following way: the standard residual estimator is equivalent to the error whenever the residual is a discrete functional. The second reason lies in the observation that an important part of the residual, namely the application of the differential operator to a discrete function, is itself of discrete nature. This feature is partially captured by the following definition of discrete functionals with polynomial densities and is discussed in Remark 4.18.

**Definition 4.17 (discrete functionals and meshed subdomains)** Given  $m_1 \in \mathbb{N}_0$ ,  $m_2 \in \mathbb{N}_0 \cup \{-1\}$ , let  $\mathbb{F}_{\mathcal{T}} := \mathbb{F}(\mathcal{T}) := \mathbb{F}_{m_1, m_2}(\mathcal{T})$  denote the subspace

$$\left\{ \ell \in H^{-1}(\Omega) \mid \forall w \in H_0^1(\Omega) \langle \ell, w \rangle = \sum_{F \in \mathcal{F}} \int_F q_F w + \sum_{T \in \mathcal{T}} \int_T q_T w \right. \\ \left. \text{with fixed } q_F \in \mathbb{P}_{m_1}(F), q_T \in \mathbb{P}_{m_2}(T) \right\}$$

of *discrete functionals*, i.e., functionals that are given by piecewise polynomial densities over elements and interior faces. We call  $(m_1, m_2)$  the *degrees of the discrete functionals*.

A set  $\omega$  is a  $\mathcal{T}$ -*meshed subdomain* if it is a subdomain of  $\Omega$  and it is triangulated by a submesh  $\mathcal{T}_\omega \subset \mathcal{T}$ , i.e. we have  $\overline{\omega} = \cup_{T \in \mathcal{T}_\omega} T$ . A functional  $\ell \in H^{-1}(\Omega)$  is then *discrete in the meshed subdomain*  $\omega$  whenever  $\ell|_{H_0^1(\omega)} \in \mathbb{F}(\mathcal{T}_\omega)$ . Here the faces

$$\mathcal{F}_\omega := \{F \in \mathcal{F} \mid F \cap \omega \neq \emptyset, F \not\subset \partial\omega\}$$

involved in  $\mathbb{F}(\mathcal{T}_\omega)$  are interior to  $\omega$ ; e.g., the subspaces  $\mathbb{F}(\{T\})$ ,  $T \in \mathcal{T}$ , do not involve any faces. In accordance with (4.5), we use the abbreviations  $\mathcal{T}_z$  and  $\mathcal{F}_z$  for  $\mathcal{T}_{\omega_z}$  and  $\mathcal{F}_{\omega_z}$ .

Alternatively, the local space  $\mathbb{F}(\mathcal{T}_\omega)$  can be obtained from the global space  $\mathbb{F}(\mathcal{T})$  by restriction:

$$\mathbb{F}(\mathcal{T}_\omega) = \mathbb{F}(\mathcal{T})|_{H_0^1(\omega)} := \left\{ \ell|_{H_0^1(\omega)} \mid \ell \in \mathbb{F}(\mathcal{T}) \right\}. \quad (4.23)$$

**Remark 4.18 (differential operator and discrete functionals)** The image of the finite element space  $\mathbb{V}_{\mathcal{T}}$  under the linear differential operator  $-\operatorname{div}(A\nabla \cdot) + c \cdot$  is again a finite-dimensional space. For differential operators with piecewise polynomial coefficients  $A$  and  $c$ , the above notion captures this by the property that the application of such operators to discrete functions  $v \in \mathbb{V}_{\mathcal{T}}$  yields discrete functionals. Indeed, if

$$A \in (\mathbb{S}_{\mathcal{T}}^{n_A, -1})^{d \times d}, \quad \text{and} \quad c \in \mathbb{S}_{\mathcal{T}}^{n_c, -1}$$

with  $n_A, n_c \in \mathbb{N}_0$ , piecewise integration by parts gives the representation

$$\begin{aligned} & \int_{\Omega} A \nabla v \cdot \nabla w + c v w \\ &= \sum_{F \in \mathcal{F}} \int_F [[A \nabla v]] \cdot \mathbf{n}_F w + \sum_{T \in \mathcal{T}} \int_T (c v - \operatorname{div}(A \nabla v)) w, \end{aligned} \quad (4.24)$$

where, for any interior face  $F \in \mathcal{F}$  and any element  $T \in \mathcal{T}$ ,

$$[[A \nabla v]] \cdot \mathbf{n}_F \in \mathbb{P}_{m_1}(F), \quad c v - \operatorname{div}(A \nabla v) \in \mathbb{P}_{m_2}(T)$$

with  $m_1 = n_A + n - 1$  and  $m_2 = \max\{n_A + n - 2, n_c + n\}$ . Note however that not every functional in  $\mathbb{F}_{m_1, m_2}(\mathcal{T})$  can be written in the form of (4.24). In fact, as the representation of a discrete functional is made up of  $L^2$ -scalar products on domains that are mutually disjoint or of different dimension, we have

$$\dim \mathbb{F}_{m_1, m_2}(\mathcal{T}) = \#\mathcal{F} \dim \mathbb{P}_{m_1} + \#\mathcal{T} \dim \mathbb{P}_{m_2}, \quad (4.25)$$

which is strictly greater than  $\dim \mathbb{V}_{\mathcal{T}}$ . This enlargement, which is implicitly used in the proof of Lemma 4.12 (partial lower bounds), turns out to be convenient also in the constructive approximation of discrete functionals.

In view of the aforementioned properties of discrete functionals, we may split the residual into a discrete and a non-discrete part. Splitting the standard residual estimator in the alternative local indicators (4.22) is in a similar spirit. To see this, we introduce  $\Pi_{\mathcal{T}} \ell \in H^{-1}(\Omega)$  given by

$$\langle \Pi_{\mathcal{T}} \ell, w \rangle := \sum_{F \in \mathcal{F}} \int_F (\Pi_F^{m_1} g) w + \sum_{T \in \mathcal{T}} \int_T (\Pi_T^{m_2} f) w, \quad w \in H_0^1(\Omega), \quad (4.26)$$

for all  $\ell \in H^{-1}(\Omega)$  admitting the representation

$$\langle \ell, w \rangle = \sum_{F \in \mathcal{F}} \int_F g w + \sum_{T \in \mathcal{T}} \int_T f w, \quad w \in H_0^1(\Omega)$$

with suitable density functions  $g$  and  $f$ . Then the splitting of the alternative indicators (4.22) corresponds to writing

$$R_{\mathcal{T}} = \Pi_{\mathcal{T}} R_{\mathcal{T}} + (I - \Pi_{\mathcal{T}}) R_{\mathcal{T}}. \quad (4.27)$$

Moreover, Remark 4.14 (oscillation degrees - data oscillation reduction) discusses in particular conditions for the identity

$$(I - \Pi_{\mathcal{T}}) R_{\mathcal{T}} = f - \Pi_{\mathcal{T}} f, \quad (4.28)$$

which follows from the property that  $\Pi_{\mathcal{T}}$  reproduces the functionals in Remark 4.18 (differential operator and discrete functionals); compare with (4.20), which in terms of  $\Pi_{\mathcal{T}}$  reads  $\Pi_{\mathcal{T}}(\Delta u_{\mathcal{T}}) = \Delta u_{\mathcal{T}}$ , where  $\Delta$  is now the distributional Laplacian.

The fact that the definition of  $\Pi_{\mathcal{T}}$  requires the extra regularity  $f \in L^2(\Omega)$

and  $\mathbf{A} \in W_{\infty}^1(\Omega; \mathbb{R}^{d \times d})$  not only excludes applications but, in light of Remark 4.11, (nonasymptotic overestimation) entails overestimation. To circumvent this flaw, we therefore aim at constructing a new approximation operator  $P_{\mathcal{T}}$  that is defined for all functionals  $\ell \in H^{-1}(\Omega)$ . Furthermore, we want this operator to be a projection onto  $\mathbb{F}_{\mathcal{T}}$  so that the counterpart

$$(I - P_{\mathcal{T}})R_{\mathcal{T}} = f - P_{\mathcal{T}}f$$

of (4.28) holds under the same conditions.

To summarize, our plan is to develop a quasi-optimal a posteriori error analysis by constructing a locally computable, linear projection

$$P_{\mathcal{T}} : H^{-1}(\Omega) \rightarrow \mathbb{F}_{\mathcal{T}} \subset H^{-1}(\Omega)$$

onto the discrete functionals that induces a splitting

$$R_{\mathcal{T}} = P_{\mathcal{T}}R_{\mathcal{T}} + (I - P_{\mathcal{T}})R_{\mathcal{T}} \quad (4.29)$$

of the residual into a *discretized residual*  $P_{\mathcal{T}}R_{\mathcal{T}}$ , which can be easily quantified, as well as an *oscillatory residual*  $(I - P_{\mathcal{T}})R_{\mathcal{T}}$ , which under the conditions of Remark 4.18 (differential operator and discrete functionals) reduces to an oscillation of the right-hand side  $f$ .

The proof of an upper bound of the error will then involve a triangle inequality applied to the right-hand side of (4.29). The following remark provides criteria to prevent overestimation in such a context, and is followed by a comparison of the two approaches represented by (4.27) and (4.29).

**Remark 4.19 (avoiding overestimation)** Overestimation can be often avoided by ensuring two relatively simple conditions. In order to discuss them informally, consider the model inequality

$$|\cdot| \leq |\cdot|_1 + |\cdot|_2, \quad (4.30)$$

where  $|\cdot|$ ,  $|\cdot|_i$ ,  $i = 1, 2$ , are seminorms and denote the domain and kernel of  $|\cdot|$ , respectively, by  $\text{dom } |\cdot|$  and  $\text{ker } |\cdot|$  etc.

The first condition, the *kernel condition*, is that zero is not overestimated:

$$\text{ker } |\cdot| \subset \text{ker } |\cdot|_1 \cap \text{ker } |\cdot|_2. \quad (4.31a)$$

The second condition, the *domain condition*, is that a finite value is never bounded by  $\infty$ , or in other, still informal, words: if the evaluation of the left-hand side is (or can be uniquely defined to be) a finite value, the same holds for the right-hand side:

$$\text{dom } |\cdot| \subset \text{dom } |\cdot|_1 \cap \text{dom } |\cdot|_2. \quad (4.31b)$$

[Kreuzer *et al.* n.d.] provide a precise version of the domain condition (4.31b), show that, given inequality (4.30), the two conditions (4.31) are also sufficient for equivalence, and discuss further applications of this viewpoint.

In order to illustrate the application of Remark 4.19, let us consider only the special case of the Poisson equation, i.e.  $\mathbf{A} = \mathbf{I}$ ,  $c = 0$ , and linear elements, i.e.  $n = 1$ . We start with the upper bound in terms of the standard residual estimator in Theorem 4.9 and view it as a function of the right-hand side  $f$ . Then the domain condition is violated as the left-hand side is defined for any  $f \in H^{-1}(\Omega)$ , while the right-hand side is defined only for  $f \in L^2(\Omega)$ . Also the kernel condition is not verified: the left-hand side vanishes whenever  $f = -\Delta v$  for some  $v \in \mathbb{V}_{\mathcal{T}}$ , while the right-hand side vanishes only for  $f = 0$ . The splitting in the alternative local indicators (4.22) does not worsen this situation, i.e. it does not add further instances in which kernel and domain condition are missed. Note however that the oscillation indicators alone are in conflict with the domain condition and, therefore, another PDE indicator cannot cure the overestimation. Finally, for the outlined approach, the splitting (4.29) and the required properties for the operator  $P_{\mathcal{T}}$  ensure both kernel and domain condition.

#### 4.4 Testing discrete functionals

The  $H^{-1}$ -projection  $P_{\mathcal{T}}$  onto the discrete functionals  $\mathbb{F}_{\mathcal{T}}$  will be defined by means of a Petrov-Galerkin-type approach. This section prepares its definition by individuating a suitable *test space*  $\mathbb{V}_{\mathcal{T}}^+$ . The key property of  $\mathbb{V}_{\mathcal{T}}^+$  is that the dual pairing  $\langle \cdot, \cdot \rangle$  in  $H^{-1}(\Omega)$  is nondegenerate on the product  $\mathbb{F}_{\mathcal{T}} \times \mathbb{V}_{\mathcal{T}}^+$ . Doing so, the degrees  $(m_1, m_2)$  of the discrete functionals will be parameters that are omitted in the notation. The construction of the test space  $\mathbb{V}_{\mathcal{T}}^+$  proceeds in two steps. First, we locally associate to the degrees of freedom in  $\mathbb{F}_{\mathcal{T}}$  certain functions on  $\Omega$ . For the degrees of freedom on the skeleton, this will involve a suitable *extension operator*. Second, we turn the ensuing functions into admissible test functions with the help of a *cut-off*.

The degrees of freedom in  $\mathbb{F}_{\mathcal{T}}$  are given by density polynomials over element and faces. For an element  $T \in \mathcal{T}$ , if we extend such a density polynomial  $q_T$  by 0 off  $T$ , it is already a function on  $\Omega$ . For a polynomial  $q_F$  associated with a face  $F \in \mathcal{F}$ , we employ the following extension operator  $E_F$  mapping a function  $v$  on  $F$  to a function on  $\omega_F$ , the union of all elements  $T$  containing  $F$ .

Given such an element  $T \subset \omega_F$ , write  $z_0, \dots, z_d$  for its vertices,  $z_d$  being the one opposite to  $F$ , denote by  $b_F := \frac{1}{d} \sum_{i=0}^{d-1} z_i$  the barycenter of  $F$ , and set

$$(E_F v)(x) := v \left( \phi_d(x) b_F + \sum_{i=0}^{d-1} \phi_{z_i}(x) z_i \right) \quad x \in T,$$

and extend by 0 off  $\omega_F$ . Note that the definition of  $E_F$  is affine invariant and does not depend on the enumeration of the vertices of  $F$ . The next lemma collects two useful properties of this extension operator.

**Lemma 4.20 (extending from faces)** *Let  $F \in \mathcal{F}$  be a face. For any function*



$v \in L^2(F)$ , we have

$$\|E_F v\|_{L^2(\omega_F)} \lesssim h_F^{\frac{1}{2}} \|v\|_{L^2(F)},$$

where  $h_F$  stands for the diameter of  $F$  and the hidden constant depends only on  $d$  and the shape regularity coefficient  $\sigma$ . Furthermore, if  $v$  is a polynomial, then  $E_F v$  is a continuous piecewise polynomial of the same degree.

*Proof.* [1] In view of  $(E_F v)^2 = E_F(v^2)$ , we may show the inequality by verifying

$$\int_{\omega_F} E_F w \lesssim h_F \int_F w \quad (4.32)$$

for any positive function  $w : F \rightarrow \mathbb{R}$ , which amounts to  $L^1$ -stability. To this end, we shall use a standard argument involving the following reference situation, which slightly differs from the common one with  $T_d = \{x = (x_1, \dots, x_d) \in \mathbb{R}^d \mid 0 \leq x_i \leq 1, \sum_{i=1}^d x_i \leq 1\}$  and  $b_d := (d+1)^{-1}(1, \dots, 1) \in \mathbb{R}^d$ . Let the reference face  $\widehat{F} := T_{d-1} - b_{d-1} \subset \mathbb{R}^{d-1}$  be a translation of  $T_{d-1}$  and let the reference simplex  $\widehat{T} \subset \mathbb{R}^d$  be the convex hull of  $\widehat{F} \times \{0\}$  and the canonical basis vector  $e_d = (0, \dots, 0, 1) \in \mathbb{R}^d$ . The barycenter of  $\widehat{F} \times \{0\}$  is then the origin in  $\mathbb{R}^d$  and the barycentric coordinate of the vertex  $e_d$  of  $\widehat{T}$  is  $x_d$ . Fixing an element  $T$  with  $T \subset \omega_F$ , denote by  $G_T : \widehat{T} \rightarrow T$  a bi-affine map sending vertices of  $\widehat{F} \times \{0\}$  into vertices of  $F$  and  $e_d$  into the vertex of  $T$  opposite to  $F$ , and write  $G_F : \widehat{F} \times \{0\} \rightarrow F$  for the restriction  $G_T|_{\widehat{F} \times \{0\}}$ . The pullbacks of  $E_F w$  and  $w$  satisfy

$$G_T^*(E_F w)(x', x_d) = G_F^* w(x', 0)$$

for all  $x = (x', x_d) \in \widehat{T} = \{y = (y', y_d) \in \widehat{F} \times \mathbb{R} \mid 0 \leq y_d \leq 1 - |y' + b_{d-1}|_1\}$ , where  $|z'|_1 = \sum_{i=1}^{d-1} |z'_i|$  stands for the  $\ell_1$ -norm in  $\mathbb{R}^{d-1}$ . Consequently, the transformation rule, the fact that the Jacobians of  $G_T$  and  $G_F$  are constant, the Fubini theorem,  $w \geq 0$ , and  $|\widehat{F}|/|\widehat{T}| = |T_{d-1}|/|T_d| = d$  yield

$$\begin{aligned} \int_T E_F w &= \frac{|T|}{|\widehat{T}|} \int_{\widehat{T}} G_T^*(E_F w) = \frac{|T|}{|\widehat{T}|} \int_{\widehat{F}} \int_0^{1-|x'+b_{d-1}|_1} G_F^* w(x', 0) dx_d dx' \\ &\leq \frac{|T|}{|\widehat{T}|} \int_{\widehat{F}} G_F^* w(\cdot, 0) = \frac{|T||\widehat{F}|}{|\widehat{T}||F|} \int_F w = d \frac{|T|}{|F|} \int_F w. \end{aligned}$$

Since the hidden constant in  $|T|/|F| \lesssim h_F$  depends only on the shape regularity coefficient  $\sigma$ , this implies the  $L^1$ -stability bound (4.32) and so also the claimed  $L_2$ -stability is proven.

[2] The second statement for polynomial arguments of  $E_F$  is a direct consequence of its definition.  $\square$

In view of the above different treatment of elements and faces, we need two types of cut-off functions: one for elements denoted by  $\phi_T$  and another one for faces denoted by  $\phi_F$ . Possible choices are the element and face bubbles from (4.14)

and (4.17). Since other choices will be useful in Sect. 4.8 below, we shall rely henceforth only on the following properties.

**Assumption 4.21 (abstract cut-off)** The *cut-off functions*  $\phi_T$ ,  $T \in \mathcal{T}$ , and  $\phi_F$ ,  $F \in \mathcal{F}$ , satisfy

$$\text{supp } \phi_T = T, \quad 0 \leq \phi_T \leq 1, \quad \text{supp } \phi_F = \omega_F, \quad 0 \leq \phi_F \leq 1,$$

and act in an affine-equivalent manner on the element level: there exists a finite-dimensional linear space  $\mathbb{S}^+ \subset L^\infty(T_d)$  of functions defined on the reference element  $T_d$  such that  $G_T^* \phi_T$  does not depend on  $T$ ,  $G_F^* \phi_F$  does not depend on  $F$ , and

$$\begin{aligned} \forall T \in \mathcal{T}, \forall q \in \mathbb{P}_{m_2}(T) \quad G_T^*(q\phi_T) &\in \mathbb{S}^+, \\ \forall F \in \mathcal{F}, \forall q \in \mathbb{P}_{m_1}(F), \forall T \in \mathcal{T} \quad G_T^*\left(\left((E_F q)\phi_F\right)|_T\right) &\in \mathbb{S}^+, \end{aligned}$$

where  $G_T$  is a bi-affine map from the reference element  $T_d$  to the generic element  $T$  and  $G_T^*(v) = v \circ G_T$  denotes the pullback of a function  $v : T \rightarrow \mathbb{R}$  via  $G_T$ .

In the case of the bubble functions (4.14) and (4.17), Assumption 4.21 holds with  $\mathbb{S}^+ = \mathbb{P}_{\max\{m_1+d-1, m_2+d\}}(T_d)$  as the extension operators  $E_F$ ,  $F \in \mathcal{F}$ , preserve the polynomial degree.

**Lemma 4.22 (properties of cut-off)** *If the cut-off functions  $\phi_T$ ,  $T \in \mathcal{T}$ , and  $\phi_F$ ,  $F \in \mathcal{F}$ , satisfy Assumption 4.21, then we have*

$$\|q\|_{L^2(T)} \lesssim \|q\phi_T^{1/2}\|_{L^2(T)} \quad \text{and} \quad \|\nabla(q\phi_T)\|_{L^2(T)} \lesssim h_T^{-1}\|q\phi_T\|_{L^2(T)}$$

for all  $q \in \mathbb{P}_{m_2}(T)$  as well as

$$\|q\|_{L^2(F)} \lesssim \|q\phi_F^{1/2}\|_{L^2(F)} \quad \text{and} \quad \|\nabla((E_F q)\phi_F)\|_{L^2(\omega_F)} \lesssim h_F^{-1}\|(E_F q)\phi_F\|_{L^2(\omega_F)}$$

for all  $q \in \mathbb{P}_{m_1}(F)$ . The hidden constants depend only on  $d$ , the shape regularity coefficient  $\sigma$ , the degrees  $(m_1, m_2)$  of the discrete functionals, and the space  $\mathbb{S}^+$ .

*Proof.* [1] To verify the first claimed inequality, we start by noting that, thanks to  $\text{supp } \phi_T = T$ , we have  $\phi_{T_d} := G_T^*(\phi_T) > 0$  in the interior of  $T_d$ . Hence,  $\|\cdot\|_{L^2(T_d)}$  and  $\|\cdot\|_{\phi_{T_d}^{1/2}} = \|\cdot\phi_{T_d}^{1/2}\|_{L^2(T_d)}$  are norms on  $\mathbb{P}_{m_2}(T_d)$  and, thanks to  $\dim \mathbb{P}_{m_2}(T_d) < \infty$ , are equivalent. A standard round trip to the reference element and  $G_T^* q G_T^* \phi_T^{1/2} = G_T^*(q\phi_T^{1/2})$  thus yield

$$\begin{aligned} \|q\|_{L^2(T)} &\lesssim h_T^{d/2} \|G_T^* q\|_{L^2(T_d)} \lesssim h_T^{d/2} \|G_T^* q G_T^* \phi_T^{1/2}\|_{L^2(T_d)} \\ &\lesssim \|q\phi_T^{1/2}\|_{L^2(T)}, \end{aligned}$$

and the first claimed inequality is established. The third one is proved along the same lines, but with a round trip to the reference face.

[2] For the other claimed inequalities note that  $\|\nabla \cdot\|_{L^2(T_d)}$  and  $\inf_{c \in \mathbb{R}} \|\cdot - c\|_{L^2(T_d)}$  are equivalent norms on the finite-dimensional quotient space  $\mathbb{S}^+/\mathbb{R}$ . Consequently,

further round trips to the reference element give

$$\begin{aligned} \|\nabla(q\phi_T)\|_{L^2(T)} &\lesssim h_T^{-1+d/2} \|\nabla G_T^*(q\phi_T)\|_{L^2(T_d)} \lesssim h_T^{-1+d/2} \inf_{c \in \mathbb{R}} \|G_T^*(q\phi_T) - c\|_{L^2(T_d)} \\ &\lesssim h_T^{-1+d/2} \|G_T^*(q\phi_T)\|_{L^2(T_d)} \lesssim h_T^{-1} \|q\phi_T\|_{L^2(T)} \end{aligned}$$

and

$$\begin{aligned} \|\nabla(E_F(q)\phi_T)\|_{L^2(\omega_F)}^2 &= \sum_{T \subset \omega_F} \|\nabla(E_F(q)\phi_T)\|_{L^2(T)}^2 \\ &\lesssim h_F^{-1} \sum_{T \subset \omega_F} \|E_F(q)\phi_T\|_{L^2(T)}^2 = h_F^{-1} \|E_F(q)\phi_T\|_{L^2(\omega_F)}^2 \end{aligned}$$

and the proof is completed.  $\square$

These preparations lead to the following test space for discrete functionals.

**Definition 4.23 (test space for discrete functionals)** Using the cut-off functions from Assumption 4.21, we associate to the space  $\mathbb{F}_{\mathcal{T}}$  of discrete functionals the following *test space*:

$$\begin{aligned} \mathbb{V}_{\mathcal{T}}^+ := \mathbb{V}^+(\mathcal{T}) &:= \text{span} \left( \{q_T \phi_T \mid q_T \in \mathbb{P}_{m_2}(T), T \in \mathcal{T}\} \right. \\ &\quad \left. \bigcup \{E_F(q_F)\phi_F \mid q_F \in \mathbb{P}_{m_1}(F), F \in \mathcal{F}\} \right). \end{aligned}$$

If  $\omega$  is a subdomain of  $\Omega$  meshed by  $\mathcal{T}_{\omega}$ , then  $\mathbb{V}^+(\mathcal{T}_{\omega})$  is the test space for  $\mathbb{F}(\mathcal{T}_{\omega})$ .

Similarly as for the space  $\mathbb{F}_{\mathcal{T}}$  of discrete functionals, the test space over a subdomain  $\omega$  meshed by  $\mathcal{T}_{\omega}$  can be obtained from the global test space, namely

$$\mathbb{V}^+(\mathcal{T}_{\omega}) = \mathbb{V}^+(\mathcal{T}) \cap H_0^1(\omega). \quad (4.33)$$

#### 4.5 A projection onto discrete functionals

Having the test space  $\mathbb{V}_{\mathcal{T}}^+$  from Definition 4.23 at our disposal, we are now ready to construct a  $H^{-1}$ -projection  $P_{\mathcal{T}}$  as suggested in Sect. 4.3. Like in the previous section, the degrees  $(m_1, m_2)$  of the discrete functionals in  $\mathbb{F}_{\mathcal{T}}$  are hidden parameters.

**Definition 4.24 (projection onto discrete functionals)** Given the discrete functionals  $\mathbb{F}_{\mathcal{T}}$  and the test space  $\mathbb{V}_{\mathcal{T}}^+$ , we define a *projection*  $P_{\mathcal{T}} : H^{-1}(\Omega) \rightarrow \mathbb{F}_{\mathcal{T}}$  by

$$\langle P_{\mathcal{T}}\ell, w \rangle = \langle \ell, w \rangle \quad \forall w \in \mathbb{V}_{\mathcal{T}}^+. \quad (4.34)$$

The well-posedness of this definition and algebraic properties of  $P_{\mathcal{T}}$  are verified in the following Lemma 4.25. Moreover, a representation of  $P_{\mathcal{T}}$  in form of a quasi-interpolation operator is given in Corollary 4.61 below.

The *polynomial densities* of  $P_{\mathcal{T}}\ell$  are denoted by  $P_T\ell := P_{\mathcal{T},T}\ell$ ,  $T \in \mathcal{T}$ , and  $P_F\ell := P_{\mathcal{T},F}\ell$ ,  $F \in \mathcal{F}$ , so that

$$\langle P_{\mathcal{T}}\ell, w \rangle = \sum_{T \in \mathcal{T}} \int_T P_T\ell w + \sum_{F \in \mathcal{F}} \int_F P_F\ell w. \quad (4.35)$$

In the next lemma we show in particular that  $P_{\mathcal{T}}$  is a local operator. In order to formulate this, we shall use  $\mathcal{T}$ -meshed local subdomains, i.e.  $\mathcal{T}$ -meshed subdomains  $\omega$  for which there exists a mesh element  $T \in \mathcal{T}$  with  $\omega \subset \omega_T$ . For the next lemma addressing algebraic properties of the operator  $P_{\mathcal{T}}$ , recall the notation  $\omega_F = \cup_{T \in \mathcal{T}_F} T$  from (4.7) for an interior face  $F \in \mathcal{F}$ .

**Lemma 4.25 (algebraic properties)** *The operator  $P_{\mathcal{T}}$  is a local linear projection onto the subspace  $\mathbb{F}_{\mathcal{T}}$  of discrete functionals. More precisely, for any local subdomain  $\omega$  meshed by  $\mathcal{T}$ , there is a linear projection  $P_{\omega} : H^{-1}(\omega) \rightarrow \mathbb{F}(\mathcal{T}_{\omega})$  such that*

$$P_{\mathcal{T}}\ell|_{H^{-1}(\omega)} = P_{\omega} \left( \ell|_{H_0^1(\omega)} \right) \in \mathbb{F}(\mathcal{T}_{\omega})$$

for all  $\ell \in H^{-1}(\Omega)$ .

*Proof.* [1] We first show that the degrees of freedom in Definition 4.23 of  $\mathbb{V}_{\mathcal{T}}^+$  are linearly independent. To this end, we fix an element  $T \in \mathcal{T}$ , denote by  $F_1, \dots, F_l$ ,  $l \leq d+1$  its faces that are in  $\mathbb{F}_{\mathcal{T}}$  and write  $\phi_0 := \phi_T$ ,  $\phi_i := \phi_{F_i}$  and  $E_i := E_{F_i}$  for  $i = 1, \dots, l$ . We then claim that, for all  $q_0 \in \mathbb{P}_{m_2}(T) \setminus \{0\}$  and all  $q_i \in \mathbb{P}_{m_1}(F_i) \setminus \{0\}$ ,  $i = 1, \dots, l$ , we have

$$\alpha_0 q_0 \phi_0 + \sum_{i=1}^l \alpha_i E_i(q_i) \phi_i = 0 \text{ in } T \implies \alpha_0 = \dots = \alpha_l = 0. \quad (4.36)$$

In light of  $\text{supp } \phi_T = T$  and  $\text{supp } \phi_F = \omega_F$ , we observe  $\phi_0|_{\partial T} = 0$  and  $\phi_i|_{\partial T \setminus F_i} = 0$  for  $i = 1, \dots, l$ . We thus evaluate the hypothesis of (4.36) first on the faces  $F_1, \dots, F_l$  and then in the element  $T$ . This gives  $\alpha_i = 0$  for  $i = 0, \dots, l$  and (4.36) is verified.

[2] Next, we discuss the well-posedness of (4.34). The linear independence (4.36) and (4.25) lead to

$$\dim \mathbb{V}_{\mathcal{T}}^+ = \#\mathcal{T} \dim \mathbb{P}_{m_2} + \#\mathcal{F} \dim \mathbb{P}_{m_1} = \dim \mathbb{F}_{\mathcal{T}}. \quad (4.37)$$

Thus, it suffices to show the implication

$$\ell \in \mathbb{F}_{\mathcal{T}} : \langle \ell, w \rangle = 0 \ \forall w \in \mathbb{V}_{\mathcal{T}}^+ \implies \ell = 0. \quad (4.38)$$

For that purpose, let  $\ell \in \mathbb{F}_{\mathcal{T}}$  and let  $q_T$ ,  $T \in \mathcal{T}$ , and  $q_F$ ,  $F \in \mathcal{F}$  be the determining polynomials. For any element  $T \in \mathcal{T}$ , the choice  $w = q_T \phi_T \in H_0^1(T)$  implies

$$0 = \langle \ell, w \rangle = \int_T (q_T)^2 \phi_T, \quad \text{i.e.} \quad q_T = 0.$$

Thus,  $\ell$  does not have contributions from elements. Regarding faces, any choice  $w = (E_F q_F) \phi_F \in H_0^1(\omega_F)$ ,  $F \in \mathcal{F}$ , therefore gives

$$0 = \langle \ell, w \rangle = \int_F (q_F)^2 \phi_F, \quad \text{i.e.} \quad q_F = 0,$$

and implication (4.38) is established. Combining (4.37) and (4.38), we have that

the dual pairing in  $H^{-1}(\Omega)$  is nondegenerate on  $\mathbb{F}_{\mathcal{T}} \times \mathbb{V}_{\mathcal{T}}^+$ , which in turn ensures that  $P_{\mathcal{T}}$  is well-defined. The Petrov-Galerkin character of the definition (4.34) then ensures that  $P_{\mathcal{T}}$  is a linear projection onto  $\mathbb{F}_{\mathcal{T}}$ .

[3] It remains to show that  $P_{\mathcal{T}}$  is a local operator. Given any local subdomain  $\omega$  meshed by  $\mathcal{T}_{\omega}$ , we can apply the preceding proof to  $\mathcal{T}_{\omega}$  instead of  $\mathcal{T}$ . This shows that the dual pairing in  $H^{-1}(\omega)$  is nondegenerate on  $\mathbb{F}(\mathcal{T}_{\omega}) \times \mathbb{V}^+(\mathcal{T}_{\omega})$  and ensures a local projection operator  $P_{\omega} : H^{-1}(\omega) \rightarrow \mathbb{F}(\mathcal{T}_{\omega})$ . Taking into account (4.23) and (4.33), we note

$$P_{\mathcal{T}}|_{H^{-1}(\omega)} = P_{\omega},$$

which completes the proof.  $\square$

The verification of (4.38), which amounts to a proof of uniqueness, suggests the following approach to compute  $P_{\mathcal{T}}\ell$ ,  $\ell \in H^{-1}(\Omega)$ .

**Remark 4.26 (local computation)** Let  $\ell \in H^{-1}(\Omega)$ . Recalling (4.35), the polynomials  $P_T\ell$ ,  $T \in \mathcal{T}$ , and  $P_F\ell$ ,  $F \in \mathcal{F}$  can be computed by solving first

$$\int_T P_T\ell q\phi_T = \langle \ell, q\phi_T \rangle \quad \forall T \in \mathcal{T}, q \in \mathbb{P}_{m_2}(T), \quad (4.39)$$

and then

$$\int_F P_F\ell q\phi_F = \langle \ell, q\phi_F \rangle - \sum_{T \subset \omega_F} \int_T P_T\ell q\phi_F \quad \forall F \in \mathcal{F}, q \in \mathbb{P}_{m_1}(F). \quad (4.40)$$

This amounts to two block-diagonal linear systems with, respectively,  $\#\mathcal{T}$  blocks of size  $\dim \mathbb{P}_{m_2}$  and  $\#\mathcal{F}$  blocks of size  $\dim \mathbb{P}_{m_1}$ . Each block and each corresponding right-hand side arises from local computations.

**Remark 4.27 (star localization vs locality of  $P_{\mathcal{T}}$ )** Stars  $\omega_z$ ,  $z \in \mathcal{V}$ , are meshed local subdomains. Lemma 4.25 thus shows that, for any vertex  $z \in \mathcal{V}$ , there is a linear projection  $P_z : H^{-1}(\omega_z) \rightarrow \mathbb{F}(\mathcal{T}_z)$  such that

$$P_{\mathcal{T}}\ell|_{H^{-1}(\omega_z)} = P_z \left( \ell|_{H_0^1(\omega_z)} \right) \in \mathbb{F}(\mathcal{T}_z)$$

for all  $\ell \in H^{-1}(\Omega)$ . The stars also appear in the localizing upper bound of the global residual norm in Corollary 4.6. As they are minimal subdomains therein, cf. Remark 4.7, it may appear that a finer localization with smaller domains cannot be exploited in a posteriori analysis. Although this is true in the context of upper bounds, the increased locality of  $P_{\mathcal{T}}$  is useful in the context of lower bounds; see the reduced lower bound (5.17), which follows from the interior vertex property introduced in Definition 4.50, and is crucial to derive the contraction result (5.26).

We have already mentioned that we shall use  $P_{\mathcal{T}}$  to split the residual. In light of the bounds for the residual norm in Corollary 4.6 (star localization of residual norms), this should be done in a locally stable manner. In order to formulate and

employ the local stability properties of  $P_{\mathcal{T}}$ , the following notation is useful. Given a local subdomain  $\omega$  meshed by  $\mathcal{T}_{\omega}$ , we define the  $\mathbb{V}^+(\mathcal{T}_{\omega})$ -discrete dual norm by

$$\|\ell\|_{\mathbb{V}^+(\mathcal{T}_{\omega})}^* := \sup_{w \in \mathbb{V}^+(\mathcal{T}_{\omega}), \|\nabla w\|_{L^2(\omega)}=1} \langle \ell, w \rangle \quad \ell \in H^{-1}(\omega). \quad (4.41)$$

In view of  $\mathbb{V}^+(\mathcal{T}_{\omega}) \subset H_0^1(\omega)$ , we have  $\|\ell\|_{\mathbb{V}^+(\mathcal{T}_{\omega})}^* \leq \|\ell\|_{H^{-1}(\omega)}$  for all  $\ell \in H^{-1}(\omega)$ .

**Lemma 4.28 (local  $H^{-1}$ -stability)** *The projection  $P_{\mathcal{T}}$  is locally  $H^{-1}$ -stable: for any local subdomain  $\omega$  meshed by  $\mathcal{T}_{\omega}$ , we have*

$$\|P_{\mathcal{T}}\|_{\mathcal{L}(H^{-1}(\omega))} = \sup_{\ell \in \mathbb{F}(\mathcal{T}_{\omega})} \frac{\|\ell\|_{H^{-1}(\omega)}}{\|\ell\|_{\mathbb{V}^+(\mathcal{T}_{\omega})}^*} \leq C_{\text{IStb}},$$

where  $C_{\text{IStb}} = C_{\text{IStb}}(d, \sigma, m_1, m_2)$  depends only on  $d$ , the shape regularity coefficient  $\sigma$  from (3.9), the degrees  $(m_1, m_2)$  of the discrete functionals and the space  $\mathbb{S}^+$ .

*Proof.* [1] We start by verifying the ' $\leq$ '-part of the claimed identity for the operator norm. The definition of the operator norm leads to

$$\|P_{\omega}\|_{\mathcal{L}(H^{-1}(\omega))} = \sup_{\ell \in H^{-1}(\omega)} \frac{\|P_{\omega}\ell\|_{H^{-1}(\omega)}}{\|\ell\|_{H^{-1}(\omega)}} \leq \sup_{\ell \in H^{-1}(\omega)} \frac{\|P_{\omega}\ell\|_{H^{-1}(\omega)}}{\|\ell\|_{\mathbb{V}^+(\mathcal{T}_{\omega})}^*}.$$

We now notice that  $\|\ell\|_{\mathbb{V}^+(\mathcal{T}_{\omega})}^* = \|P_{\omega}\ell\|_{\mathbb{V}^+(\mathcal{T}_{\omega})}^*$  in view of (4.41) and (4.34). Hence,

$$\|P_{\omega}\|_{\mathcal{L}(H^{-1}(\omega))} \leq \sup_{\ell \in H^{-1}(\omega)} \frac{\|P_{\omega}\ell\|_{H^{-1}(\omega)}}{\|P_{\omega}\ell\|_{\mathbb{V}^+(\mathcal{T}_{\omega})}^*} = \sup_{\ell \in \mathbb{F}(\mathcal{T}_{\omega})} \frac{\|\ell\|_{H^{-1}(\omega)}}{\|\ell\|_{\mathbb{V}^+(\mathcal{T}_{\omega})}^*}$$

because the projection  $P_{\omega}$  is onto  $\mathbb{F}(\mathcal{T}_{\omega})$ .

[2] Next, we show that  $\|P_{\mathcal{T}}\|_{\mathcal{L}(H^{-1}(\omega))}$  is uniformly bounded. Let  $\ell \in \mathbb{F}(\mathcal{T}_{\omega})$  be a discrete functional, namely

$$\langle \ell, w \rangle = \sum_{T \in \mathcal{T}_{\omega}} \int_T q_T w + \sum_{F \in \mathcal{F}_{\omega}} \int_F q_F w \quad \forall w \in H_0^1(\omega).$$

We proceed in two steps that are quite similar to the classical standard residual estimates. Arguing as in (4.11), and using Lemma 2.2 (first Poincaré inequality) in the domain  $\omega$  of diameter about  $h_T$  for  $w \in H_0^1(\omega)$ , we obtain

$$|\langle \ell, w \rangle| \lesssim \left( \sum_{T \in \mathcal{T}_{\omega}} h_T^2 \|q_T\|_{L^2(T)}^2 + \sum_{F \in \mathcal{F}_{\omega}} h_F \|q_F\|_{L^2(F)}^2 \right)^{1/2} \|\nabla w\|_{L^2(\omega)}$$

whence

$$\|\ell\|_{H^{-1}(\omega)}^2 \lesssim \sum_{T \in \mathcal{T}_{\omega}} h_T^2 \|q_T\|_{L^2(T)}^2 + \sum_{F \in \mathcal{F}_{\omega}} h_F \|q_F\|_{L^2(F)}^2. \quad (4.42)$$

Here we do not exploit that  $\ell$  is discrete in  $\omega$ ; this will be crucial in the second step, when we bound each term on the right-hand side, in a manner reminding

the derivation of classical lower bounds. For any  $T \in \mathcal{T}_\omega$ , we write  $\mathbb{V}^+(T)^*$  as shorthand for  $\mathbb{V}^+(\{T\})^*$  and exploit Lemma 4.22 (properties of cut-off) to deduce

$$\begin{aligned} \|q_T\|_{L^2(T)}^2 &\lesssim \int_T (q_T)^2 \phi_T = \langle \ell, q_T \phi_T \rangle \leq \|\ell\|_{\mathbb{V}^+(T)^*} \|\nabla(q_T \phi_T)\|_{L^2(T)} \\ &\lesssim \|\ell\|_{\mathbb{V}^+(T)^*} h_T^{-1} \|q_T \phi_T\|_{L^2(T)} \leq \|\ell\|_{\mathbb{V}^+(T)^*} h_T^{-1} \|q_T\|_{L^2(T)}, \end{aligned}$$

whence

$$h_T \|q_T\|_{L^2(T)} \lesssim \|\ell\|_{\mathbb{V}^+(T)^*}. \quad (4.43)$$

For an interior face  $F \in \mathcal{F}_\omega$ , we proceed similarly, taking into account also Lemma 4.20 and that  $\mathbb{V}^+(T) \subset \mathbb{V}^+(\mathcal{T}_F)$  entails  $\|\ell\|_{\mathbb{V}^+(T)^*} \leq \|\ell\|_{\mathbb{V}^+(\mathcal{T}_F)^*}$  for  $T \in \mathcal{T}_F$ . We thus obtain

$$\begin{aligned} \|q_F\|_{L^2(F)}^2 &\lesssim \int_F (q_F)^2 \phi_F = \langle \ell, (E_F q_F) \phi_F \rangle - \sum_{T \subset \omega_F} \int_T q_T (E_F q_F) \phi_F \\ &\lesssim \|\ell\|_{\mathbb{V}^+(\mathcal{T}_F)^*} \|\nabla((E_F q_F) \phi_T)\|_{L^2(\omega_F)} + \sum_{T \subset \omega_F} \|q_T\|_{L^2(T)} \|(E_F q_F) \phi_T\|_{L^2(T)} \\ &\lesssim \|\ell\|_{\mathbb{V}^+(\mathcal{T}_F)^*} h_T^{-1} \|(E_F q_F) \phi_T\|_{L^2(\omega_F)} \lesssim \|\ell\|_{\mathbb{V}^+(\mathcal{T}_F)^*} h_T^{-\frac{1}{2}} \|q_F\|_{L^2(F)}, \end{aligned}$$

i.e.

$$h_F^{\frac{1}{2}} \|q_F\|_{L^2(F)} \lesssim \|\ell\|_{\mathbb{V}^+(\mathcal{T}_F)^*}. \quad (4.44)$$

The number of elements and interior faces in the local subdomain  $\omega$  is uniformly bounded by  $d$  and the shape regularity coefficient  $\sigma$ . Hence, inequalities (4.43) and (4.44) together with the inclusions  $\mathbb{V}^+(T) \subset \mathbb{V}^+(\mathcal{T}_\omega)$  for  $T \in \mathcal{T}_\omega$  and  $\mathbb{V}^+(\mathcal{T}_F) \subset \mathbb{V}^+(\mathcal{T}_\omega)$  for  $F \in \mathcal{F}_\omega$  imply

$$\sum_{T \in \mathcal{T}_\omega} h_T^2 \|q_T\|_{L^2(T)}^2 + \sum_{F \in \mathcal{F}_\omega} h_F \|q_F\|_{L^2(F)}^2 \lesssim \|\ell\|_{\mathbb{V}^+(\mathcal{T}_\omega)^*}^2. \quad (4.45)$$

Combing (4.42) and (4.45) shows that the ratio  $\|\ell\|_{H^{-1}(\omega)} / \|\ell\|_{\mathbb{V}^+(\mathcal{T}_\omega)^*}$  for  $\ell \in \mathbb{F}(\mathcal{T}_\omega)$  is bounded by a universal constant depending on  $d, \sigma, m_1, m_2$  and  $\mathbb{S}^+$ .

[3] It remains to complete the proof of the claimed identity for the operator norm. To this end, we first introduce an operator  $Q_\omega : H_0^1(\omega) \rightarrow \mathbb{V}^+(\mathcal{T}_\omega)$  by

$$\langle \ell, Q_\omega w \rangle = \langle \ell, w \rangle \quad \forall \ell \in \mathbb{F}(\mathcal{T}_\omega).$$

As the one for  $P_\omega$ , this definition is well-posed because the pair  $(\mathbb{F}(\mathcal{T}_\omega), \mathbb{V}^+(\mathcal{T}_\omega))$  is nondegenerate for the dual pairing of  $H^{-1}(\omega)$ ; cf. (4.37) and (4.38) of the proof of Lemma 4.25 (algebraic properties). By the Petrov-Galerkin character of the definition,  $Q_\omega$  is a linear projection onto  $\mathbb{V}^+(\mathcal{T}_\omega)$ . Given arbitrary  $\ell \in H^{-1}(\omega)$  and  $w \in H_0^1(\omega)$ , the definitions of  $Q_\omega$  and  $P_\omega$  imply

$$\langle P_\omega \ell, w \rangle = \langle P_\omega \ell, Q_\omega w \rangle = \langle \ell, Q_\omega w \rangle,$$

that is  $Q_\omega = P_\omega^*$  is the (Hilbert) adjoint to  $P_\omega$ . In other words: the adjoint  $P_\omega^*$

is a projection onto  $\mathbb{V}^+(\mathcal{T}_\omega)$ . With this, we can prove the missing inequality. Let  $\ell \in \mathbb{F}(\mathcal{T}_\omega)$  be discrete. In fact,

$$\langle \ell, w \rangle = \langle P_\omega \ell, w \rangle = \langle \ell, P_\omega^* w \rangle \quad \Rightarrow \quad \|\ell\|_{H^{-1}(\omega)} = \sup_{w \in H_0^1(\omega)} \frac{\langle \ell, P_\omega^* w \rangle}{\|\nabla w\|_{L^2(\omega)}}$$

leads to

$$\|\ell\|_{H^{-1}(\omega)} \leq \|\ell\|_{\mathbb{V}^+(\mathcal{T}_\omega)^*} \|P_\omega^*\|_{\mathcal{L}(H_0^1(\omega))} = \|\ell\|_{\mathbb{V}^+(\mathcal{T}_\omega)^*} \|P_\omega\|_{\mathcal{L}(H^{-1}(\omega))}.$$

This concludes the proof.  $\square$

**Remark 4.29 (failing global  $H^{-1}$ -stability)** For Lebesgue norms, local stability of linear operators in terms of shape regularity entails that their respective global stability is uniform under mesh refinement. The fact that the first part of Lemma 4.5 (localization of  $H^{-1}$ -norm) needs a condition to be true, may lead one to suspect that this implication might not be true in general for the  $H^{-1}$ -norm. This suspicion is confirmed by Example 4.63 below, where we show that  $\|P_{\mathcal{T}}\|_{\mathcal{L}(H^{-1}(\Omega))}$  can tend to  $\infty$  under mesh refinement.

The proof of Lemma 4.28 provides all nontrivial ingredients to allow the approximate computation of  $\|\ell\|_{H^{-1}(\omega)}$  whenever  $\ell \in H^{-1}(\Omega)$  is discrete in  $\omega$ .

**Corollary 4.30 (quantifying  $H^{-1}$ -norms of discrete functionals)** *Let  $\omega \subset \Omega$  be a local subdomain meshed by  $\mathcal{T}_\omega$  and  $\ell \in H^{-1}(\Omega)$  be discrete in  $\omega$ , given by the polynomials  $q_T$  for  $T \in \mathcal{T}_\omega$ , and  $q_F$  for  $F \in \mathcal{F}_\omega$ , where  $F \in \mathcal{F}_\omega$  are the interior faces in  $\omega$ . We then have*

$$\|\ell\|_{H^{-1}(\omega)}^2 \approx \sum_{T \in \mathcal{T}_\omega} h_T^2 \|q_T\|_{L^2(T)}^2 + \sum_{F \in \mathcal{F}_\omega} h_F \|q_F\|_{L^2(F)}^2,$$

where the hidden constants depend on  $d$ , the shape regularity coefficient  $\sigma$ , the degrees  $(m_1, m_2)$  of the discrete functionals, and the space  $\mathbb{S}^+$  appearing in Assumption 4.21.

*Proof.* This is a consequence of (4.42) and (4.45), the latter requiring  $\ell$  to be discrete, along with the fact that  $\|\ell\|_{\mathbb{V}^+(\mathcal{T}_\omega)^*} \leq \|\ell\|_{H^{-1}(\omega)^*}$  for any  $\ell \in H^{-1}(\omega)$ .  $\square$

**Corollary 4.31 (local near-best approximation)** *The projection  $P_{\mathcal{T}}$  yields local near-best approximations: for any functional  $\ell \in H^{-1}(\Omega)$  and any local subdomain  $\omega$  meshed by  $\mathcal{T}_\omega$ , we have*

$$\|\ell - P_{\mathcal{T}}\ell\|_{H^{-1}(\omega)} \leq C_{\text{lstb}} \inf_{\chi \in \mathbb{F}(\mathcal{T}_\omega)} \|\ell - \chi\|_{H^{-1}(\omega)},$$

where  $C_{\text{lstb}}$  is the constant of Lemma 4.28 (local  $H^{-1}$  stability).

*Proof.* Fix a local subdomain  $\omega$  meshed by  $\mathcal{T}_\omega$  and let  $\chi \in \mathbb{F}(\mathcal{T}_\omega)$  be arbitrary. Thanks to Lemma 4.25 (algebraic properties), we have  $P_\omega \chi = \chi$  and

$$(\ell - P_{\mathcal{T}}\ell)|_{H_0^1(\omega)} = (I - P_\omega)\ell|_{H_0^1(\omega)} = (I - P_\omega)(\ell - \chi)|_{H_0^1(\omega)}.$$



As  $P_\omega$  is a nontrivial projection on the Hilbert space  $H^{-1}(\omega)$ , [Szyld 2006] ensures

$$\|I - P_\omega\|_{\mathcal{L}(H^{-1}(\omega))} = \|P_\omega\|_{\mathcal{L}(H^{-1}(\omega))} \leq C_{\text{Istb}}. \quad (4.46)$$

Hence

$$\|\ell - P_{\mathcal{T}}\ell\|_{H^{-1}(\omega)} \leq C_{\text{Istb}}\|\ell - \chi\|_{H^{-1}(\omega)}$$

concludes the proof because  $\chi \in \mathbb{F}(\mathcal{T}_\omega)$  is arbitrary.  $\square$

We illustrate the approximation of possible parts of the residual with the projection  $P_{\mathcal{T}}$  in a series of three remarks. For that purpose, the approximation quality is to be measured with a local  $H^{-1}(\omega)$ -norm, and it is instructive to compare with the operator  $\Pi_{\mathcal{T}}$  from (4.26). Recall that the operator  $\Pi_{\mathcal{T}}$  is used implicitly in the standard approach (see Sect. 4.2) to approximate the discrete functionals  $\mathbb{F}_{\mathcal{T}}$ .

**Remark 4.32 (approximating functions)** For functions, the local error with  $P_{\mathcal{T}}$  is uniformly dominated by the one with  $\Pi_{\mathcal{T}}$ . More precisely, if  $m_2 \geq 0$  and  $\ell \in H^{-1}(\Omega)$  satisfies  $\langle \ell, w \rangle = \int_\Omega f w$  where  $f \in L^p(\Omega)$  with  $p > 2d/(2+d)$ , then Corollary 4.31 (local near-best approximations) and  $\Pi_{\mathcal{T}}f = \sum_{T \in \mathcal{T}} (\Pi_T^{m_2} f) \chi_T \in \mathbb{F}(\mathcal{T}_\omega)$  imply, for any local meshed subdomain  $\omega$ ,

$$\|\ell - P_{\mathcal{T}}\ell\|_{H^{-1}(\omega)} \lesssim \|f - \Pi_{\mathcal{T}}f\|_{H^{-1}(\omega)}.$$

Observe that, although  $\ell$  is a function,  $P_{\mathcal{T}}\ell$  is typically not a function. This property might look undesirable but it is crucial for an advantage of  $P_{\mathcal{T}}$  over  $\Pi_{\mathcal{T}}$  and closely related to the fact that the opposite inequality does not hold; cf. Remark 4.34 about stability.

Furthermore, supposing  $f \in L^2(\Omega)$  and combining the preceding inequality with Lemma 2.2 (first Poincaré inequality) gives

$$\|\ell - P_{\mathcal{T}}\ell\|_{H^{-1}(\omega_T)}^2 \lesssim \sum_{T' \subset \omega_T} h_{T'}^2 \|f - \Pi_{T'}^{m_2} f\|_{L^2(\omega_{T'})}^2, \quad (4.47)$$

which establishes that the local  $P_{\mathcal{T}}$ -oscillation of functions is uniformly dominated by its classical  $\Pi_{\mathcal{T}}$ -counterpart but not vice versa.

In Sect. 7.3.1 this case is considered in the context of adaptive approximation.

**Remark 4.33 (approximating admissible functionals)** For functionals allowing for the application of  $\Pi_{\mathcal{T}}$ , the local error with  $P_{\mathcal{T}}$  is again uniformly dominated by the one with  $\Pi_{\mathcal{T}}$ . In view of the previous remark, let us consider only  $\ell \in H^{-1}(\Omega)$  such that  $\langle \ell, w \rangle = \int_\Sigma g w$  where  $g \in L^p(\Sigma)$  with  $\Sigma := \cup_{F \in \mathcal{F}} F$  and  $p > 2(d-1)/d$ . Note that we have again  $\Pi_{\mathcal{T}}\ell \in \mathbb{F}(\mathcal{T}_\omega)$  as  $\langle \Pi_{\mathcal{T}}\ell, w \rangle = \sum_{F \in \mathcal{F}} \int_F (\Pi_F^{m_1} g) w$  for all  $w \in H_0^1(\Omega)$ . Corollary 4.31 (local near-best approximations) thus ensures, for any local meshed subdomain  $\omega$ ,

$$\|\ell - P_{\mathcal{T}}\ell\|_{H^{-1}(\omega)} \lesssim \|\ell - \Pi_{\mathcal{T}}\ell\|_{H^{-1}(\omega)}.$$

Moreover, supposing  $g \in L^2(\Sigma)$  and combining the scaled trace theorem (4.12)

with Lemma 2.2 (first Poincaré inequality) yields

$$\|\ell - P_{\mathcal{T}}\ell\|_{H^{-1}(\omega_T)}^2 \lesssim \sum_{\substack{F \subseteq \bar{\omega}_T \\ F \not\subseteq \partial\omega_T}} h_F \|g - \Pi_F^{m_1} g\|_{L^2(F)}^2.$$

Also this case will be revisited in the context of adaptive approximation, namely in Sect. 7.3.3.

**Remark 4.34 (stability of approximation)** The error with  $P_{\mathcal{T}}$  is stable, while the one with  $\Pi_{\mathcal{T}}$  is not. To see this by example, we restrict to  $(m_1, m_2) = (0, 0)$ , fix some interior face  $F \in \mathcal{F}$  and, for  $\varepsilon > 0$  sufficiently small, consider

$$\langle \ell_\varepsilon, w \rangle := \int_{\Omega} f_\varepsilon w = \frac{1}{2\varepsilon} \int_{-\varepsilon}^{\varepsilon} \int_F w(y + s\mathbf{n}_F) dy ds, \quad w \in H_0^1(\Omega),$$

where  $f_\varepsilon = (2\varepsilon)^{-1} \chi_{F_\varepsilon}$  is a multiple of the characteristic function of  $F_\varepsilon := \{x + s\mathbf{n}_F \mid x \in F, -\varepsilon < s < \varepsilon\}$ . As

$$\begin{aligned} \langle \ell_\varepsilon - \delta_F, w \rangle &= \frac{1}{2\varepsilon} \int_{-\varepsilon}^{\varepsilon} \int_F \int_0^s \partial_{\mathbf{n}_F} w(y + t\mathbf{n}_F) dt dy ds \\ &\leq \frac{1}{2\varepsilon} \int_{-\varepsilon}^{\varepsilon} \int_{F_\varepsilon} |\nabla w(x)| dx ds \leq |F_\varepsilon|^{1/2} \|\nabla w\|_{L^2(\Omega)}, \end{aligned}$$

the functions  $(\ell_\varepsilon)_{\varepsilon>0}$  tend to the proper functional  $\delta_F$ :

$$\ell_\varepsilon \rightarrow \delta_F \text{ in } H^{-1}(\Omega).$$

Combining the convergence with the local stability of  $P_{\mathcal{T}}$ , see (4.46), yields

$$\begin{aligned} \left| \|\ell_\varepsilon - P_{\mathcal{T}}\ell_\varepsilon\|_{H^{-1}(\omega_F)} - \|\delta_F - P_{\mathcal{T}}\delta_F\|_{H^{-1}(\omega_F)} \right| &\leq \|(I - P_{\mathcal{T}})(\ell_\varepsilon - \delta_F)\|_{H^{-1}(\omega_F)} \\ &\lesssim \|\ell_\varepsilon - \delta_F\|_{H^{-1}(\omega_F)} \rightarrow 0, \end{aligned}$$

the stability of the error with  $P_{\mathcal{T}}$ . Furthermore, since  $P_{\mathcal{T}}\delta_F = \delta_F$ , the stability entails here  $\|\ell_\varepsilon - P_{\mathcal{T}}\ell_\varepsilon\|_{H^{-1}(\omega_F)} \rightarrow 0$ . For  $\Pi_{\mathcal{T}}$  however, the approximation on the skeleton and in the volume are independent of each other. Hence, combining  $\Pi_{\mathcal{T}}\delta_F = \delta_F$ , which follows from (4.26), with  $\lim_{\varepsilon \rightarrow 0} \Pi_T^0 f_\varepsilon = |F|/(2|T|)$  for the two elements  $T \in \mathcal{T}$  containing  $F$ , leads to

$$\|\delta_F - \Pi_{\mathcal{T}}\delta_F\|_{H^{-1}(\omega_F)} = 0 < \lim_{\varepsilon \rightarrow 0} \|\ell_\varepsilon - \Pi_{\mathcal{T}}\ell_\varepsilon\|_{H^{-1}(\omega_F)}.$$

Measuring the error in weighted  $L^2$ -norms instead of the  $H^{-1}$ -norm results in a more dramatic instability. Indeed, denoting by  $1_K$  and  $0_K$  the constant functions on a simplex  $K$  equal to 1 or 0, the two sides translate in

$$\begin{aligned} h_F^{\frac{1}{2}} \|1_F - \Pi_F^0 1_F\|_{L^2(F)} + \sum_{T \in \mathcal{T}_F} h_T \|0_T - \Pi_T^0 0_T\|_{L^2(T)} &= 0, \\ \lim_{\varepsilon \rightarrow 0} \left( h_F^{\frac{1}{2}} \|0_F - \Pi_F^0 0_F\|_{L^2(F)} + \sum_{T \in \mathcal{T}_F} h_T \|f_\varepsilon - \Pi_T^0 f_\varepsilon\|_{L^2(T)} \right) &= \infty. \end{aligned}$$

Note that such a transformation of volume contributions into contributions on the skeleton may occur by perturbation in the right-hand side or, in the opposite direction, by an improvement of the Galerkin approximation thanks to refinement.

In view of this instability of  $\Pi_{\mathcal{T}}$ , the inequalities in the preceding Remarks 4.32 and 4.33 cannot be reversed - a fact that can be inferred also from Remark 4.19.

The above perturbations of  $\delta_F$  are in the domain of  $\Pi_{\mathcal{T}}$ . For the functionals

$$\langle \widehat{\ell}_{\varepsilon}, w \rangle := \int_F w(y + \varepsilon \mathbf{n}_F) dy, \quad w \in H_0^1(\Omega)$$

however, it is not clear how to directly apply  $\Pi_{\mathcal{T}}$  for  $\varepsilon \neq 0$ . To the contrary, the approximations  $P_{\mathcal{T}} \widehat{\ell}_{\varepsilon}$  are defined and stable around 0. Noteworthy,  $P_{\mathcal{T}} \widehat{\ell}_{\varepsilon}$  uses volume contributions to compensate for the displacement in the representation of the singular contribution.

#### 4.6 Discretized and oscillatory residual

We now turn to the proper a posteriori analysis, that is we shall derive *upper* and *lower bounds of the error*, implementing the following plan, which is motivated in Sect. 4.3. We use the projection  $P_{\mathcal{T}}$  onto discrete functionals  $\mathbb{F}_{\mathcal{T}}$  to split the residual into discretized and oscillatory parts. Then the quantification of the *oscillatory residual* is reduced to *data* oscillation through suitable choices of the degrees  $(m_1, m_2)$  of the discrete functionals. The *discretized residual* can be quantified in various manners; see Sections 4.2 and 4.9 below.

We start by introducing indicators reflecting the announced splitting of the residual into discretized and oscillatory parts. They are vertex-indexed and, given  $z \in \mathcal{V}$ , defined by

$$\begin{aligned} \mathcal{E}_{\mathcal{T}}^{\text{abs}}(z)^2 &:= \eta_{\mathcal{T}}^{\text{abs}}(z)^2 + \text{osc}_{\mathcal{T}}(R_{\mathcal{T}}, z)^2 \quad \text{with} \\ \eta_{\mathcal{T}}^{\text{abs}}(z) &:= \|P_{\mathcal{T}} R_{\mathcal{T}}\|_{H^{-1}(\omega_z)} \quad \text{and} \quad \text{osc}_{\mathcal{T}}(R_{\mathcal{T}}, z) := \|(I - P_{\mathcal{T}})R_{\mathcal{T}}\|_{H^{-1}(\omega_z)}. \end{aligned} \quad (4.48)$$

Note that these quantities are not proper indicators: they still need to be quantified in a computable manner. By using ‘abs’ (shorthand for ‘abstract’), we hint at the fact that  $\eta_{\mathcal{T}}^{\text{abs}}$  can be quantified by various approaches.

**Lemma 4.35 (splitting of local residual norm)** *For any vertex  $z \in \mathcal{V}$ , the local residual norm is equivalent to the abstract indicator from (4.48):*

$$\frac{1}{\sqrt{2}C_{\text{lstb}}} \mathcal{E}_{\mathcal{T}}^{\text{abs}}(z) \leq \|R_{\mathcal{T}}\|_{H^{-1}(\omega_z)} \leq \sqrt{2} \mathcal{E}_{\mathcal{T}}^{\text{abs}}(z),$$

where  $C_{\text{lstb}}$  is the constant of Lemma 4.28 (local  $H^{-1}$  stability).

*Proof.* As announced, we use the linear projection  $P_{\mathcal{T}}$  in order to split the residual into a discretized and an oscillatory part:

$$R_{\mathcal{T}} = P_{\mathcal{T}} R_{\mathcal{T}} + (I - P_{\mathcal{T}}) R_{\mathcal{T}}. \quad (4.49)$$

The upper bound of the local residual norm then readily follows from the triangle inequality:

$$\|R_{\mathcal{T}}\|_{H^{-1}(\omega_z)} \leq \eta_{\mathcal{T}}^{\text{abs}}(z) + \text{osc}_{\mathcal{T}}(R_{\mathcal{T}}, z) \leq \sqrt{2}\mathcal{E}_{\mathcal{T}}^{\text{abs}}(z).$$

To show the lower bound, we exploit the local stability of  $P_{\mathcal{T}}$ , cf. (4.46), to obtain

$$\eta_{\mathcal{T}}^{\text{abs}}(z) = \|P_{\mathcal{T}}R_{\mathcal{T}}\|_{H^{-1}(\omega_z)} \leq C_{\text{lstb}}\|R_{\mathcal{T}}\|_{H^{-1}(\omega_z)}$$

and

$$\text{osc}_{\mathcal{T}}(R_{\mathcal{T}}, z) = \|(I - P_{\mathcal{T}})R_{\mathcal{T}}\|_{H^{-1}(\omega_z)} \leq C_{\text{lstb}}\|R_{\mathcal{T}}\|_{H^{-1}(\omega_z)}.$$

Squaring both inequalities, summing them, and then taking the square root finishes the proof.  $\square$

In many a posteriori analyses, this lemma is replaced by steps breaking a possible true equivalence between error and estimator. Therefore, the following remark points out the key ingredients.

**Remark 4.36 (ensuring proper equivalence)** The fact that the projection  $P_{\mathcal{T}}$  and so also  $I - P_{\mathcal{T}}$  are linear and locally bounded operators precludes overestimation; see also Remark 4.19. Comparing with Sect. 4.2 and  $\Pi_{\mathcal{T}}$  in (4.27), we see that the local stability in  $H^{-1}$  is crucial to that end and, in view of Remark 4.34, requires discrete functionals with contributions on the skeleton.

Next, we want to simplify the residual oscillations  $\text{osc}_{\mathcal{T}}(R_{\mathcal{T}}, z)$ ,  $z \in \mathcal{V}$ , in the spirit of Remark 4.14. This will be dependent on the coefficients  $\mathbf{A}$  and  $c$  of the differential operator and involve the following ‘polynomial degrees’:

$$n_{\mathbf{A}} := \min \left\{ k \in \mathbb{N}_0 \mid \mathbf{A} \in (\mathbb{S}_{\mathcal{T}}^{k,-1})^{d \times d} \right\}, \quad (4.50a)$$

$$n_c := \min \left\{ k \in \mathbb{N}_0 \cup \{-1\} \mid c \in \mathbb{S}_{\mathcal{T}}^{k,-1} \right\}, \quad (4.50b)$$

where we use the convention  $\min \emptyset = \infty$ . We shall say that the differential operator  $-\text{div}(\mathbf{A}\nabla \cdot) + c(\cdot)$  in (2.5) has *discrete coefficients* whenever  $\max\{n_{\mathbf{A}}, n_c\} < \infty$ , otherwise it has *nondiscrete coefficients*.

**Lemma 4.37 (data oscillation reduction for discrete coefficients)** *If the coefficients  $\mathbf{A}$  and  $c$  are discrete, the choices*

$$m_1 = n_{\mathbf{A}} + n - 1,$$

$$m_2 = \max\{n - 2 + n_{\mathbf{A}}, \widetilde{m}_c\} \quad \text{with} \quad \widetilde{m}_c = \begin{cases} n + n_c, & \text{if } c \neq 0, \\ 0, & \text{otherwise} \end{cases}$$

*ensure that the oscillatory residual reduces to data oscillation of the right-hand side:*

$$(I - P_{\mathcal{T}})R_{\mathcal{T}} = f - P_{\mathcal{T}}f.$$

*Proof.* The choices for  $m_1$  and  $m_2$  yield, for any face  $F \in \mathcal{F}$  and any element  $T \in \mathcal{T}$ ,

$$[[A \nabla u_{\mathcal{T}}]] \cdot \mathbf{n}_F \in \mathbb{P}_{m_1}(F) \quad \text{and} \quad \operatorname{div}(A \nabla u_{\mathcal{T}})|_T \in \mathbb{P}_{m_2}(T).$$

Furthermore, if  $c \neq 0$ , we also have  $cu_{\mathcal{T}}|_T \in \mathbb{P}_{m_2}(T)$  and the claimed identity follows from  $-\operatorname{div}(A \nabla u_{\mathcal{T}}) + cu_{\mathcal{T}} \in \mathbb{F}_{\mathcal{T}}$ .  $\square$

**Remark 4.38 (Poisson equation with linear elements)** In the case of the Poisson equation with linear elements, the choices in Lemma 4.37 lead to  $m_1 = 0$  and  $m_2 = 0$ . Alternatively, one may use  $m_1 = 0$  and  $m_2 = -1$  (recall we have set  $\mathbb{P}_{-1}(T) = \{0\}$ ), cf. [Diening, Kreuzer and Stevenson 2016] or [Siebert and Veeres 2007]. The choice here leads to an oscillation for which the standard oscillation indicators  $h_T \|f - \Pi_T f\|_T$ ,  $T \in \mathcal{T}$ , can be used as a surrogate; see also Remark 4.43 about surrogates.

If one of the coefficients,  $A$  or  $c$ , is nondiscrete, the range of the finite element space  $\mathbb{V}_{\mathcal{T}}$  under the differential operator  $-\operatorname{div}(A \nabla \cdot) + c(\cdot)$  consists of functionals whose densities are not piecewise polynomial. Consequently, the oscillatory residual cannot be reduced to the oscillation  $f - P_{\mathcal{T}}f$ , or to any other oscillation of  $f$  involving discrete functionals with piecewise polynomial densities. The next result illustrates the idea of a non-perfect remedy, namely bounding the residual oscillation defined in (4.48) in terms of data oscillation and discrete stability. For its formulation, we define global  $H^{-1}$ -oscillations by

$$\operatorname{osc}_{\mathcal{T}}(\ell)^2 := \sum_{z \in \mathcal{V}} \operatorname{osc}_{\mathcal{T}}(\ell, z)^2, \quad \ell \in H^{-1}(\Omega), \quad (4.51)$$

which, in contrast to  $\|(I - P_{\mathcal{T}})\ell\|_{H^{-1}(\Omega)}$ , is bounded in terms of  $\ell$ ; cf. Remark 4.29 (failing global  $H^{-1}$ -stability).

**Lemma 4.39 (surrogate data oscillation reduction)** Let  $m_A := \min\{n_A, n - 1\}$  and  $m_c := \min\{n_c, n - 1\}$  and define  $m_1$  and  $m_2$  as in Lemma 4.37, but replacing  $n_A$  and  $n_c$ , respectively, with  $m_A$  and  $m_c$ . Given any approximations

$$\widehat{A} \in (\mathbb{S}_{\mathcal{T}}^{m_A, -1})^{d \times d} \quad \text{and} \quad \widehat{c} \in \mathbb{S}_{\mathcal{T}}^{m_c, -1},$$

we then have, for all vertices  $z \in \mathcal{V}$ ,

$$\begin{aligned} \operatorname{osc}_{\mathcal{T}}(R_{\mathcal{T}}, z) &\leq \operatorname{osc}_{\mathcal{T}}(f, z) + C_{\text{lstb}} C(d, \sigma) \|A - \widehat{A}\|_{L^\infty(\omega_z)} \|\nabla u_{\mathcal{T}}\|_{L^2(\omega_z)} \\ &\quad + C_{\text{lstb}} C(d, \sigma) \|h(c - \widehat{c})\|_{L^\infty(\omega_z)} \|u_{\mathcal{T}}\|_{L^2(\omega_z)} \end{aligned}$$

and thus

$$\begin{aligned} \operatorname{osc}_{\mathcal{T}}(R_{\mathcal{T}})^2 &\leq 3 \operatorname{osc}_{\mathcal{T}}(f)^2 \\ &\quad + \frac{3(d+1)}{\alpha^2} \|f\|_{H^{-1}(\Omega)}^2 \left( \|A - \widehat{A}\|_{L^\infty(\Omega)}^2 + C_P^2 C(d, \sigma) \|h(c - \widehat{c})\|_{L^\infty(\Omega)}^2 \right) \end{aligned}$$

where  $C_{\text{lstb}}$  is the constant from Lemma 4.28 (local  $H^{-1}$ -stability),  $h$  the meshsize

function defined by  $h|_T = h_T$  for all  $T \in \mathcal{T}$ ,  $\alpha$  is the coercivity constant from (2.29), and  $C_P$  is the constant in Lemma 2.2 (first Poincaré inequality).

The bounds of Lemma 4.39 are obviously not convenient if  $A$  or  $c$  are not continuous. We therefore implement the underlying idea in Sect. 5.4 differently.

*Proof.* [1] To verify the local bound, let  $z \in \mathcal{V}$  be any vertex. By linearity of  $P_{\mathcal{T}}$ , we obtain

$$\begin{aligned} \text{osc}_{\mathcal{T}}(R_{\mathcal{T}}, z) &\leq \text{osc}_{\mathcal{T}}(f, z) \\ &\quad + \|(I - P_{\mathcal{T}})(-\text{div}(A \nabla u_{\mathcal{T}}))\|_{H^{-1}(\omega_z)} + \|(I - P_{\mathcal{T}})(cu_{\mathcal{T}})\|_{H^{-1}(\omega_z)} \end{aligned}$$

and it remains to bound appropriately the two terms involving the coefficients  $A$  and  $c$ . As the definitions of  $m_1$  and  $m_2$  ensure  $-\text{div}(\widehat{A} \nabla u_{\mathcal{T}}) \in \mathbb{F}_{\mathcal{T}}$ , Corollary 4.31 (local near-best approximation), the scaled trace theorem (4.12) and Lemma 2.2 (first Poincaré inequality) give

$$\begin{aligned} \|(I - P_{\mathcal{T}})(-\text{div}(A \nabla u_{\mathcal{T}}))\|_{H^{-1}(\omega_z)} &\leq C_{\text{IStb}} \|-\text{div}((A - \widehat{A}) \nabla u_{\mathcal{T}})\|_{H^{-1}(\omega_z)} \\ &\lesssim C_{\text{IStb}} \|A - \widehat{A}\|_{L^\infty(\omega_z)} \|\nabla u_{\mathcal{T}}\|_{L^2(\omega_z)}. \end{aligned}$$

As  $\widehat{c}u_{\mathcal{T}} \in \mathbb{F}_{\mathcal{T}}$  thanks to the definition of  $m_2$ , a similar argument using again Lemma 2.2 and  $\text{diam } \omega_z \leq Ch_z$  on  $\omega_z$  provides

$$\begin{aligned} \|(I - P_{\mathcal{T}})(cu_{\mathcal{T}})\|_{H^{-1}(\omega_z)} &\leq C_{\text{IStb}} \|(c - \widehat{c})u_{\mathcal{T}}\|_{H^{-1}(\omega_z)} \\ &\leq CC_{\text{IStb}} \|h(c - \widehat{c})\|_{L^\infty(\omega_z)} \|u_{\mathcal{T}}\|_{L^2(\omega_z)}, \end{aligned}$$

and the local bound is verified.

[2] To show the global bound, we square the local bound and sum it over all vertices  $z \in \mathcal{V}$  to obtain

$$\begin{aligned} \text{osc}_{\mathcal{T}}(R_{\mathcal{T}})^2 &\leq 3 \sum_{z \in \mathcal{V}} \text{osc}_{\mathcal{T}}(f, z)^2 + 3(d+1)C \|A - \widehat{A}\|_{L^\infty(\Omega)}^2 \|\nabla u_{\mathcal{T}}\|_{L^2(\Omega)}^2 \\ &\quad + 3(d+1)C \|h(c - \widehat{c})\|_{L^\infty(\Omega)}^2 \|u_{\mathcal{T}}\|_{L^2(\Omega)}^2. \end{aligned}$$

Hence, Lemma 2.2 (first Poincaré inequality) on  $\Omega$  and discrete stability,

$$\|u_{\mathcal{T}}\|_{L^2(\Omega)} \leq C_P \|\nabla u_{\mathcal{T}}\|_{L^2(\Omega)} \leq \frac{C_P}{\alpha} \|f\|_{H^{-1}(\Omega)},$$

finish the proof.  $\square$

The following remarks set Lemma 4.35 (splitting of local residual norm) and the accompanying results Lemma 4.37 and Lemma 4.39 on the reduction to data oscillation in the context of adaptive algorithms.

**Remark 4.40 (structure of splitting)** Combining Lemma 4.35 (splitting of local residual norms) with Lemma 4.37 or Lemma 4.39 about reduction to data oscillation

thus provides an abstract estimator with the following two global parts:

$$\eta_{\mathcal{T}}^{\text{abs}}(u_{\mathcal{T}})^2 := \sum_{z \in \mathcal{V}} \eta_{\mathcal{T}}^{\text{abs}}(u_{\mathcal{T}}, z)^2$$

and, writing  $\mathcal{D} = (A, c, f)$  for the data of the partial differential equation,

$$\text{osc}_{\mathcal{T}}^{\text{abs}}(\mathcal{D})^2 := \text{osc}_{\mathcal{T}}(f)^2,$$

$$\text{osc}_{\mathcal{T}}^{\text{abs}}(\mathcal{D})^2 := \text{osc}_{\mathcal{T}}(f)^2 + C_1 \max_{z \in \mathcal{V}} \|A - \widehat{A}\|_{L^\infty(\omega_z)}^2 + C_2 \max_{z \in \mathcal{V}} \|h(c - \widehat{c})\|_{L^\infty(\omega_z)}^2,$$

the latter provided  $(A, c)$  are not discrete. It is important to note the different nature of these two parts. The first part  $\eta_{\mathcal{T}}^{\text{abs}}(u_{\mathcal{T}})$ , the *abstract PDE indicator*,

- is strictly related to the structure of the underlying PDE,
- involves only discrete functionals from  $\mathbb{F}_z$ , and
- the evaluation of its local indicators  $\eta_{\mathcal{T}}^{\text{abs}}(u_{\mathcal{T}}, z)$  requires the global computation of the discrete solution  $u_{\mathcal{T}}$ .

In contrast, the second part  $\text{osc}_{\mathcal{T}}^{\text{abs}}(\mathcal{D})$ , the *oscillation (indicator)*,

- depends only on the data  $\mathcal{D}$  of the differential operator,
- involves non-discrete functionals, and
- the evaluation of its local indicators  $\text{osc}_{\mathcal{T}}(f, z)$ ,  $\|A - \widehat{A}\|_{L^\infty(\omega_z)}$ ,  $\|h(c - \widehat{c})\|_{L^\infty(\omega_z)}$ ,  $z \in \mathcal{V}$ , is completely local.

The respective properties ‘discrete nature’ and ‘local dependence’ of the two parts are the key advantage over the whole local residual indicators  $\|R_{\mathcal{T}}\|_{H^{-1}(\omega_z)}$ ,  $z \in \mathcal{V}$ , and will be instrumental in the algorithmic design in Sections 5 and 6 below.

**Remark 4.41 (minimal regularity and regularizing  $P_{\mathcal{T}}$ )** It is worth noting that the results in this section do not involve any regularity beyond (4.2) and that the projection  $P_{\mathcal{T}}$  has a regularizing effect. In particular, we have

$$\text{Im } P_{\mathcal{T}} = \mathbb{F}_{\mathcal{T}} \subset H^{-\frac{1}{2}-\varepsilon}(\Omega) \quad \text{for any small } \varepsilon > 0$$

thanks to the trace theorem in fractional Sobolev spaces. As a consequence, most techniques for a posteriori error estimation can be directly applied to the discretized residual  $P_{\mathcal{T}}R_{\mathcal{T}}$ , without any special twisting and under natural regularity assumptions.

**Remark 4.42 (reduction vs surrogate reduction)** The kernel condition of Remark 4.19 (avoiding overestimation) is not verified for the bounds in Lemma 4.39 (surrogate data oscillation reduction). These bounds may thus exhibit overestimation and cannot be reversed. If we use the right-hand side of an overestimating bound as a part of an estimator, we shall call that part a *surrogate*. This label marks a crucial difference between the cases represented by Lemma 4.39 (surrogate data oscillation reduction) and Lemma 4.37 (data oscillation reduction), which is free of any overestimation.

**Remark 4.43 (surrogate data oscillation)** Surrogates for data oscillation indicators can be useful in order to provide more direct access for computation. For example, if  $f \in L^2(\Omega)$ , the bound (4.47) by the classical  $\Pi_{\mathcal{T}}$ -oscillation can be approximated by numerical integration. In such a context, it is useful to take the following points into account:

- Computable surrogates, i.e. computable upper bounds, for data oscillation indicators are in general impossible. In fact, generic data from an infinite-dimensional space will not be completely seen by the finite information available at any stage of a computation; cf. also [Kreuzer and Veiser 2021, Lemma 2 and Corollary 5] illustrating this fact for  $\text{osc}_{\mathcal{T}}(f)$  with the help of orthogonality. Hence, computable surrogates will hinge on additional a priori information on the given data. We postpone a discussion of examples to Section 7.3.
- As a general rule, surrogates should be applied last. This avoids that other parts of the estimators are also affected by overestimation; see Remark 4.46 (standard vs modified residual estimator) below.

#### 4.7 Modified residual estimation

In view of the splitting into PDE and oscillation indicators and the discussion on the computability of the latter, it remains to quantify the abstract PDE indicators  $\eta_{\mathcal{T}}^{\text{abs}}(u_{\mathcal{T}}, z)$ ,  $z \in \mathcal{V}$ . To this end, we can employ Corollary 4.30 (quantifying  $H^{-1}$ -norms of discrete functionals), resulting in a modification of the standard residual estimator  $\mathcal{E}_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, \mathcal{D})$  from Section 4.2. Alternative quantifications by other techniques of a posteriori error estimation are discussed in Section 4.9 below. Doing so, for simplicity, we consider only the case given by the following assumption.

**Assumption 4.44 (discrete coefficients and discrete functionals)** Suppose that the coefficients  $A$  and  $c$  in (2.5) are discrete and choose the degrees  $(m_1, m_2)$  of the discrete functionals in  $\mathbb{F}_{\mathcal{T}}$  according to Lemma 4.37 (data oscillation reduction for discrete coefficients).

For nondiscrete coefficients, one essentially has to invoke Lemma 4.39 (surrogate data oscillation reduction) instead of Lemma 4.37 in order to reduce to data oscillation.

We shall employ the bisection method in order to refine the mesh. Since this method is based upon the subdivision of elements, it is convenient to split the estimator into contributions associated with elements and not with vertices as in Sect. 4.6.

To define the modified residual estimator, we recall the representation (4.35) of the  $H^{-1}$ -projection  $P_{\mathcal{T}}$ , and we use Assumption 4.44 (discrete coefficients and



discrete functionals) to set

$$\begin{aligned}
\mathcal{E}_{\mathcal{T}}^2 &:= \sum_{T \in \mathcal{T}} \mathcal{E}_{\mathcal{T}}(T)^2 \quad \text{with} \\
\mathcal{E}_{\mathcal{T}}(T)^2 &:= \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f, T)^2 := \eta_{\mathcal{T}}(u_{\mathcal{T}}, T)^2 + \text{osc}_{\mathcal{T}}(f, T)^2, \\
\eta_{\mathcal{T}}(u_{\mathcal{T}}, T)^2 &:= h_T \sum_{F \subset \partial T \setminus \partial \Omega} \| [\mathbf{A} \nabla u_{\mathcal{T}}] \cdot \mathbf{n}_F - P_F f \|_{L^2(F)}^2 \\
&\quad + h_T^2 \| P_T f - cu_{\mathcal{T}} + \text{div}(\mathbf{A} \nabla u_{\mathcal{T}}) \|_{L^2(T)}^2, \\
\text{osc}_{\mathcal{T}}(f, T)^2 &:= \| f - P_{\mathcal{T}} f \|_{H^{-1}(\omega_T)}^2.
\end{aligned} \tag{4.52}$$

Clearly, this is a variant of the standard residual estimator in (4.10), where the main differences are given by the corrections  $P_F f$ ,  $F \in \mathcal{F}$ , of the jump residual and the replacement of  $f|_T$  by  $P_T f$ ,  $T \in \mathcal{T}$ , in the PDE indicator. As shown by the following theorem and remarks, the modification leads to more accurate a posteriori bounds.

**Theorem 4.45 (modified residual estimator)** *Under Assumption 4.44, the modified residual estimator (4.52) is equivalent to the error: more precisely, we have*

$$C_L \mathcal{E}_{\mathcal{T}} \leq \| \nabla(u - u_{\mathcal{T}}) \|_{L^2(\Omega)} \leq C_U \mathcal{E}_{\mathcal{T}},$$

where the constants  $C_U \geq C_L > 0$  depend only on the coefficients  $(\mathbf{A}, c)$ , the shape regularity coefficient  $\sigma$  from (3.9), the polynomial degree  $n$ , and  $d$ .

*Proof.* To derive the upper bound, we use the ones of Lemma 4.1 (error and residual), Corollary 4.6 (star localization of residual norm), Lemma 4.35 (splitting of local residual norms), Corollary 4.30 (quantifying  $H^{-1}$ -norms of discrete functionals) with stars and obtain

$$\begin{aligned}
\| \nabla(u - u_{\mathcal{T}}) \|_{L^2(\Omega)}^2 &\lesssim \| R_{\mathcal{T}} \|_{H^{-1}(\Omega)}^2 \lesssim \sum_{z \in \mathcal{V}} \| R_{\mathcal{T}} \|_{H^{-1}(\omega_z)}^2 \\
&\lesssim \sum_{z \in \mathcal{V}} \mathcal{E}_{\mathcal{T}}^{\text{abs}}(z)^2 = \sum_{z \in \mathcal{V}} \eta_{\mathcal{T}}^{\text{abs}}(u_{\mathcal{T}}, z)^2 + \sum_{z \in \mathcal{V}} \text{osc}_{\mathcal{T}}(f, z)^2 \\
&\lesssim \sum_{z \in \mathcal{V}} \sum_{T \in \mathcal{T}_z} \eta_{\mathcal{T}}(u_{\mathcal{T}}, T)^2 + \sum_{z \in \mathcal{V}} \text{osc}_{\mathcal{T}}(f, z)^2
\end{aligned}$$

with  $\mathcal{T}_z = \{T \in \mathcal{T} \mid T \ni z\}$ . As a given mesh element appears in the star meshes  $\mathcal{T}_z$  for at most  $d + 1$  vertices, we have

$$\sum_{z \in \mathcal{V}} \sum_{T \in \mathcal{T}_z} \eta_{\mathcal{T}}(u_{\mathcal{T}}, T)^2 \leq (d + 1) \sum_{T \in \mathcal{T}} \eta_{\mathcal{T}}(u_{\mathcal{T}}, T)^2$$

for the first sum and Lemma 4.8 (localization re-indexing) yields

$$\sum_{z \in \mathcal{V}} \text{osc}_{\mathcal{T}}(f, z)^2 \lesssim \sum_{T \in \mathcal{T}} \text{osc}_{\mathcal{T}}(f, T)^2$$

for the second sum. Inserting the last two inequalities in the previous one, we conclude the upper bound:

$$\|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)}^2 \lesssim \sum_{T \in \mathcal{T}} \eta_{\mathcal{T}}(u_{\mathcal{T}}, T)^2 + \sum_{T \in \mathcal{T}} \text{osc}_{\mathcal{T}}(f, T)^2 = \sum_{T \in \mathcal{T}} \mathcal{E}_{\mathcal{T}}(T)^2.$$

To show the lower bound, fix a mesh element  $T \in \mathcal{T}$ . Applying the local lower bounds in Corollary 4.30, Lemma 4.28 (local  $H^{-1}$  stability) and Lemma 4.1 on the local meshed subdomain  $\tilde{\omega}_T$  defined in (3.12) yields for the PDE indicator

$$\eta_{\mathcal{T}}(u_{\mathcal{T}}, T) \lesssim \|P_{\mathcal{T}} R_{\mathcal{T}}\|_{H^{-1}(\tilde{\omega}_T)} \lesssim \|R_{\mathcal{T}}\|_{H^{-1}(\tilde{\omega}_T)}. \quad (4.53)$$

In the case of the oscillation indicator, we exploit  $-cu_{\mathcal{T}} + \text{div}(A \nabla u_{\mathcal{T}}) \in \mathbb{F}_{\mathcal{T}}$  with the help of Lemma 4.25 (algebraic properties) and apply Lemma 4.28 on the local meshed subdomain  $\omega_T$ :

$$\text{osc}_{\mathcal{T}}(f, T) = \|f - P_{\mathcal{T}} f\|_{H^{-1}(\omega_T)} = \|(I - P_{\mathcal{T}}) R_{\mathcal{T}}\|_{H^{-1}(\omega_T)} \lesssim \|R_{\mathcal{T}}\|_{H^{-1}(\omega_T)}.$$

Thanks to  $\tilde{\omega}_T \subset \omega_T$ , combining the last two inequalities gives the desired local lower bound

$$\begin{aligned} \mathcal{E}_{\mathcal{T}}(T)^2 &= \eta_{\mathcal{T}}(u_{\mathcal{T}}, T)^2 + \text{osc}_{\mathcal{T}}(f, T)^2 \\ &\lesssim \|R_{\mathcal{T}}\|_{H^{-1}(\tilde{\omega}_T)}^2 + \|R_{\mathcal{T}}\|_{H^{-1}(\omega_T)}^2 \lesssim \|R_{\mathcal{T}}\|_{H^{-1}(\omega_T)}^2. \end{aligned} \quad (4.54)$$

As the number of patches  $\omega_T$ ,  $T \in \mathcal{T}$ , containing a given mesh element is uniformly bounded by  $d$  and the shape regularity coefficient  $\sigma$ , summing this bound over all mesh elements yields the global lower bound

$$\sum_{T \in \mathcal{T}} \mathcal{E}_{\mathcal{T}}(T)^2 \lesssim \sum_{T \in \mathcal{T}} \|R_{\mathcal{T}}\|_{H^{-1}(\omega_T)}^2 \lesssim \|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)}^2$$

with the help of Lemma 4.5 (localization of  $H^{-1}$  norm) and Lemma 4.1. Thus the equivalence of error and estimator is established.  $\square$

A detailed comparison of the modified residual estimator with the standard one is in order.

**Remark 4.46 (modified vs standard residual estimator)** We compare the modified residual estimator (4.52) with the standard one given by (4.10a) and the local split indicators (4.22). As a common characterizing feature, both residual estimators use properly scaled  $L^2$ -norms of jump and element residual, ready for numerical integration. However, we observe the following differences:

- while the modified estimator  $\mathcal{E}_{\mathcal{T}}$  is defined under the natural regularity assumptions (4.2), the standard estimator  $\mathcal{E}_{\mathcal{T}}^{\text{std}}$  requires  $A \in W_{\infty}^1(\Omega; \mathbb{R}^{d \times d})$  and  $f \in L^2(\Omega)$  in addition;
- while the modified estimator  $\mathcal{E}_{\mathcal{T}}$  is truly equivalent to the error, the standard estimator may overestimate it, limited however by Proposition 4.12 (partial lower bound).

By the domain test in Remark 4.19 (avoiding overestimation), we know that these two points are interrelated. However, also the kernel test is at play in the overestimation. Indeed, revisiting the proof of Theorem 4.9 (upper bound with standard residual estimator), we can replace the scaled  $L^2$ -norms of the element residuals on a star  $\omega_z$  by  $\|r\|_{H^{-1}(\omega_z)}$  and the resulting vertex-oriented variant of the residual estimator with unsplit local indicators is defined for all  $f \in H^{-1}(\Omega)$ . Overestimation can, however, still occur non-asymptotically as well as asymptotically; cf. [Cohen *et al.* 2012]. Indeed, in the case of the Poisson equation and linear finite elements, the kernel test is obviously not satisfied. This shows that the splitting in jump and element residual is quite delicate and highlights the crucial role of the modifications of the standard residual estimator: not only do they allow for stability in line with Remark 4.34 (stability of approximation) but also imply the kernel test.

To conclude this comparison, let us illustrate the second point of Remark 4.43 (surrogate data oscillation), namely that surrogates should be applied last. Using (4.47) in Remark 4.32 (approximating functions), we may replace in the modified residual estimator the  $H^{-1}$ -oscillation  $\text{osc}_{\mathcal{T}}(f)$  by the standard oscillation  $\text{osc}_{\mathcal{T}}^{\text{std}}(f)$ , which can be readily approximated with numerical integration. Doing so, we first split the residual with  $P_{\mathcal{T}}$  and then apply  $\Pi_{\mathcal{T}}$  to obtain the surrogate. Note however that, if we apply  $\Pi_{\mathcal{T}}$  earlier to split the residual, the crucial modifications will not appear and, therefore, also the PDE indicator of the standard residual estimator exhibits overestimation.

#### 4.8 Bounds for corrections and reduction of PDE estimator

In the following sections we shall use the modified residual estimator  $\mathcal{E}_{\mathcal{T}}$  from (4.52) in adaptive algorithms. In their convergence analyses, not only its relationship with the error is important, but also its relationship with the norm  $\|\nabla(u_{\mathcal{T}_*} - u_{\mathcal{T}})\|_{L^2(\Omega)}$  of (possible) *corrections*, where  $u_{\mathcal{T}_*}$  is the Galerkin approximation to  $u$  over some refinement  $\mathcal{T}_*$  of  $\mathcal{T}$ . This section establishes corresponding upper and lower bounds, as well as related results about the *global PDE indicator*

$$\eta_{\mathcal{T}}(u_{\mathcal{T}}, f)^2 := \sum_{T \in \mathcal{T}} \eta_{\mathcal{T}}(u_{\mathcal{T}}, T)^2 \quad (4.55)$$

and the *global oscillation*

$$\text{osc}_{\mathcal{T}}(f)^2 := \sum_{T \in \mathcal{T}} \text{osc}_{\mathcal{T}}(f, T)^2. \quad (4.56)$$

When it is important to indicate that the oscillations are measured in  $H^{-1}$ , we use the notation

$$\text{osc}_{\mathcal{T}}(f)_{-1} \quad \text{and} \quad \text{osc}_{\mathcal{T}}(f, T)_{-1}.$$

Let  $\mathcal{T}_*$  be a conforming mesh that is a *refinement* of  $\mathcal{T}$ , i.e. for any element  $T \in \mathcal{T}$ , there exists a submesh  $\mathcal{T}_{*,T}$  of  $\mathcal{T}_*$  such that  $T = \cup\{T_* : T_* \in \mathcal{T}_{*,T}\}$ . The

Galerkin approximation in  $\mathbb{V}_{\mathcal{T}_*}$  is characterized by

$$u_{\mathcal{T}_*} \in \mathbb{V}_{\mathcal{T}_*} : \quad \mathcal{B}[u_{\mathcal{T}_*}, w] = \langle f, w \rangle \quad \forall w \in \mathbb{V}_{\mathcal{T}_*}.$$

Hence, the discrete solution  $u_{\mathcal{T}}$  on the original mesh  $\mathcal{T}$  is not only a Galerkin approximation to the exact solution  $u$  satisfying (2.7) but also to  $u_{\mathcal{T}_*}$ . The norm  $\|\nabla(u_{\mathcal{T}_*} - u_{\mathcal{T}})\|_{L^2(\Omega)}$  of the correction therefore can be viewed as the error in approximating  $u_{\mathcal{T}_*}$  on the mesh  $\mathcal{T}$ . This viewpoint suggests considering the variant

$$\langle R_{\mathcal{T}}, w \rangle = \mathcal{B}[u_{\mathcal{T}_*} - u_{\mathcal{T}}, w] \quad \forall w \in \mathbb{V}_{\mathcal{T}_*} \quad (4.57)$$

of the error-residual identity (4.3) and introducing the discrete dual norm

$$\|R_{\mathcal{T}}\|_{(\mathbb{V}_{\mathcal{T}_*})^*} := \sup_{w \in \mathbb{V}_{\mathcal{T}_*}} \frac{\langle R_{\mathcal{T}}, w \rangle}{\|\nabla w\|_{L^2(\Omega)}} \quad (4.58)$$

of the residual as a counterpart of  $\|R_{\mathcal{T}}\|_{H^{-1}(\Omega)}$ . Arguing as in the proof of Lemma 4.1 (error and residual), we thus readily obtain the following quantitative relationship between the correction and the residual.

**Lemma 4.47 (correction and residual)** *If  $\mathcal{T}_*$  is a refinement of the mesh  $\mathcal{T}$ , the norm of the correction  $u_{\mathcal{T}_*} - u_{\mathcal{T}}$  is equivalent to the discrete residual norm. More precisely,*

$$\frac{1}{\|\mathcal{B}\|} \|R_{\mathcal{T}}\|_{(\mathbb{V}_{\mathcal{T}_*})^*} \leq \|\nabla(u_{\mathcal{T}_*} - u_{\mathcal{T}})\|_{L^2(\Omega)} \leq \frac{1}{\alpha} \|R_{\mathcal{T}}\|_{(\mathbb{V}_{\mathcal{T}_*})^*}$$

where  $\|\mathcal{B}\| \geq \alpha > 0$ , are, respectively, the continuity and coercivity constant of the bilinear form  $\mathcal{B}$ .

We first exploit the upper bound in Lemma 4.47. As the inclusion  $\mathbb{V}_{\mathcal{T}_*} \subset H_0^1(\Omega)$  implies

$$\|R_{\mathcal{T}}\|_{(\mathbb{V}_{\mathcal{T}_*})^*} \leq \|R_{\mathcal{T}}\|_{H^{-1}(\Omega)}, \quad (4.59)$$

Theorem 4.45 (modified residual estimator) immediately yields the upper bound

$$\|\nabla(u_{\mathcal{T}_*} - u_{\mathcal{T}})\|_{L^2(\Omega)} \leq C_U \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f). \quad (4.60)$$

This bound however appears to be not accurate in view of the use of (4.59). We shall sharpen it by following the line of its proof but exploiting the *full* orthogonality

$$\langle R_{\mathcal{T}}, w \rangle = 0 \quad \forall w \in \mathbb{V}_{\mathcal{T}} \quad (4.61)$$

with suitably tuned Scott-Zhang interpolation [Scott and Zhang 1990].

In order to prepare the use of this interpolation, denote by  $\mathcal{N}$  the Lagrange nodes of order  $n$  of the mesh  $\mathcal{T}$  and by  $\mathcal{F}$  and  $\mathcal{F}_*$ , respectively, the  $(d-1)$ -dimensional faces of  $\mathcal{T}$  and  $\mathcal{T}_*$ , including boundary ones. Given a node  $z \in \mathcal{N}$ , fix a face  $F_z \in \mathcal{F}$  such that  $F_z$  contains  $z$  and the following conditions are met:

$$z \in \partial\Omega \implies F_z \subset \partial\Omega,$$

$$\{F \in \mathcal{F} \cap \mathcal{F}_* \mid F \ni z\} \neq \emptyset \implies F_z \in \mathcal{F}_*.$$

Furthermore, denote by  $\psi_z^*$  the polynomial in  $\mathbb{P}_n(F_z)$  satisfying

$$\forall y \in \mathcal{N} \quad \int_{F_z} \psi_z^* \psi_y = \delta_{yz},$$

where  $\{\psi_y\}_{y \in \mathcal{N}}$  is the Lagrange basis of  $\mathbb{S}_{\mathcal{T}}^{n,0}$ , and define

$$I_{\mathcal{T}} w = \sum_{z \in \mathcal{N}} \left( \int_{F_z} \psi_z^* w \right) \psi_z. \quad (4.62)$$

The two conditions on the fixed face  $F_z$  then ensure, respectively,

$$w \in H_0^1(\Omega) \implies I_{\mathcal{T}} w \in \mathbb{V}_{\mathcal{T}}, \quad (4.63a)$$

$$w \in \mathbb{V}_{\mathcal{T}_*} \text{ and } T \in \mathcal{T} \cap \mathcal{T}_* \implies I_{\mathcal{T}} w = w \text{ on } T. \quad (4.63b)$$

In particular, if  $w \in \mathbb{V}_{\mathcal{T}_*}$ , its approximation  $I_{\mathcal{T}} w \in \mathbb{V}_{\mathcal{T}}$  is an admissible test function and coincides with  $w$  whenever possible. Finally,  $I_{\mathcal{T}}$  has the following stability and approximation properties, where the hidden constants depend only on  $d$ ,  $n$ , and the shape regularity coefficient  $\sigma$ : for any element  $T \in \mathcal{T}$  and any face  $F \in \mathcal{F}$ ,

$$\|\nabla I_{\mathcal{T}} w\|_{L^2(T)} \lesssim \|\nabla w\|_{L^2(\omega_T)}, \quad (4.64a)$$

$$\|w - I_{\mathcal{T}} w\|_{L^2(T)} \lesssim h_T \|\nabla w\|_{L^2(\omega_T)}, \quad (4.64b)$$

$$\|w - I_{\mathcal{T}} w\|_{L^2(F)} \lesssim h_F^{\frac{1}{2}} \|\nabla w\|_{L^2(\omega_F)}. \quad (4.64c)$$

The sharpening of the simple upper bound (4.60) lies in the fact that only a part of the estimator in (4.52) will be invoked. To formulate this, we define

$$\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f, \tilde{\mathcal{T}}) := \left( \sum_{T \in \tilde{\mathcal{T}}} \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f, T)^2 \right)^{1/2} \quad (4.65)$$

where  $\tilde{\mathcal{T}} \subset \mathcal{T}$  is a subset of elements in  $\mathcal{T}$ . In the same vein, we shall denote  $\eta_{\mathcal{T}}(u_{\mathcal{T}}, f, \tilde{\mathcal{T}})$  and  $\text{osc}_{\mathcal{T}}(f, \tilde{\mathcal{T}})$ .

**Theorem 4.48 (upper bound for corrections)** *Let Assumption 4.44 hold and let  $\mathcal{T}_*$  be a refinement of the mesh  $\mathcal{T}$ . The correction  $u_{\mathcal{T}_*} - u_{\mathcal{T}}$  is bounded in terms of the indicators of the refined elements  $\mathcal{T} \setminus \mathcal{T}_*$ :*

$$\|\nabla(u_{\mathcal{T}} - u_{\mathcal{T}_*})\|_{L^2(\Omega)} \leq \tilde{C}_U \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f, \mathcal{T} \setminus \mathcal{T}_*),$$

where the constant  $\tilde{C}_U > 0$  depends only on the dimension  $d$ , the coefficients  $\mathbf{A}$  and  $c$ , the polynomial degree  $n$ , and the shape regularity coefficient  $\sigma$  from (3.9).

*Proof.* 1 *Localization and splitting of the residual norm.* In light of Lemma 4.47, it suffices to bound the discrete residual norm  $\|R_{\mathcal{T}}\|_{(\mathbb{V}_{\mathcal{T}_*})^*}$ . Given  $w \in \mathbb{V}_{\mathcal{T}_*}$ , we prepare the localization of the residual by full orthogonality (4.61) and split it with

help of the projection  $P_{\mathcal{T}}$  on discrete functionals:

$$\begin{aligned} |\langle R_{\mathcal{T}}, w \rangle| &= |\langle R_{\mathcal{T}}, w - I_{\mathcal{T}}w \rangle| \\ &\leq |\langle P_{\mathcal{T}}R_{\mathcal{T}}, w - I_{\mathcal{T}}w \rangle| + |\langle f - P_{\mathcal{T}}f, w - I_{\mathcal{T}}w \rangle|, \end{aligned}$$

where we used the identity  $R - P_{\mathcal{T}}R = f - P_{\mathcal{T}}f$  in the last step. In light of

$$\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f, \mathcal{T} \setminus \mathcal{T}_*)^2 = \eta_{\mathcal{T}}(u_{\mathcal{T}}, f, \mathcal{T} \setminus \mathcal{T}_*)^2 + \text{osc}_{\mathcal{T}}(f, \mathcal{T} \setminus \mathcal{T}_*)^2,$$

it remains to bound the two terms with discretized residual  $P_{\mathcal{T}}R_{\mathcal{T}}$  and the oscillation of  $f$  appropriately.

**[2] Bounding the discretized residual.** We adopt the notation (4.35) for the densities of  $P_{\mathcal{T}}$ , and exploit the piecewise nature of the discretized residual and the local invariance (4.63b) of  $I_{\mathcal{T}}$  to deduce

$$\begin{aligned} &\langle P_{\mathcal{T}}R_{\mathcal{T}}, w - I_{\mathcal{T}}w \rangle \\ &= \sum_{T \in \mathcal{T} \setminus \mathcal{T}_*} \left( \int_T (P_T R_{\mathcal{T}})(w - I_{\mathcal{T}}w) + \frac{1}{2} \sum_{F \subset \partial T \setminus \partial \Omega} \int_F (P_F R_{\mathcal{T}})(w - I_{\mathcal{T}}w) \right). \end{aligned}$$

Invoking the local approximation properties (4.64b) and (4.64c) of  $I_{\mathcal{T}}$  leads to the desired bound for the discretized residual:

$$\begin{aligned} |\langle P_{\mathcal{T}}R_{\mathcal{T}}, w - I_{\mathcal{T}}w \rangle| &\lesssim \sum_{T \in \mathcal{T} \setminus \mathcal{T}_*} \eta_{\mathcal{T}}(u_{\mathcal{T}}, f, T) \|\nabla w\|_{L^2(\omega_T)} \\ &\lesssim \eta_{\mathcal{T}}(u_{\mathcal{T}}, f, \mathcal{T} \setminus \mathcal{T}_*) \|\nabla w\|_{L^2(\Omega)}. \end{aligned}$$

**[3] Bounding the oscillation.** We need to split the oscillation into suitable local contributions and first proceed similarly to the proof of Lemma 4.5 (localization  $H^{-1}$  norm) (i). Writing

$$w - I_{\mathcal{T}}w = \sum_{z \in \mathcal{V}} (w - I_{\mathcal{T}}w)\phi_z \quad \text{and} \quad \Omega_0 := \bigcup_{T \in \mathcal{T} \setminus \mathcal{T}_*} T,$$

we have  $(w - I_{\mathcal{T}}w)\phi_z \in H_0^1(\omega_z \cap \Omega_0)$  thanks to (4.63b) and, for any  $T \subset \omega_z \cap \Omega_0$

$$\begin{aligned} \|\nabla((w - I_{\mathcal{T}}w)\phi_z)\|_{L^2(T)} &\leq \|\phi_z \nabla(w - I_{\mathcal{T}}w)\|_{L^2(T)} + \|(w - I_{\mathcal{T}}w)\nabla\phi_z\|_{L^2(T)} \\ &\leq \|\nabla(w - I_{\mathcal{T}}w)\|_{L^2(T)} + C(d)\sigma \|\nabla w\|_{L^2(\omega_T)} \\ &\lesssim \|\nabla w\|_{L^2(\omega_T)} \end{aligned}$$

by means of  $0 \leq \phi_z \leq 1$ ,  $|\nabla\phi_z| \leq C(d)\sigma h_T^{-1}$ , (4.64a), and (4.64b). Hence, we get

$$\begin{aligned} |\langle f - P_{\mathcal{T}}f, w - I_{\mathcal{T}}w \rangle| &\leq \sum_{z \in \mathcal{V}} |\langle f - P_{\mathcal{T}}f, (w - I_{\mathcal{T}}w)\phi_z \rangle| \\ &\leq \sum_{z \in \mathcal{V}} \|f - P_{\mathcal{T}}f\|_{H^{-1}(\omega_z \cap \Omega_0)} \|\nabla((w - I_{\mathcal{T}}w)\phi_z)\|_{L^2(\omega_z \cap \Omega_0)} \\ &\lesssim \sum_{z \in \mathcal{V}} \|f - P_{\mathcal{T}}f\|_{H^{-1}(\omega_z \cap \Omega_0)} \|\nabla w\|_{L^2(\cup_{T \subseteq \omega_z \cap \Omega_0} \omega_T)} \end{aligned}$$

$$\lesssim \left( \sum_{z \in \mathcal{V}} \|f - P_{\mathcal{T}} f\|_{H^{-1}(\omega_z \cap \Omega_0)}^2 \right)^{1/2} \|\nabla w\|_{L^2(\Omega)}.$$

Since

$$\sum_{z \in \mathcal{V}} \|f - P_{\mathcal{T}} f\|_{H^{-1}(\omega_z \cap \Omega_0)}^2 \leq \sum_{T \in \mathcal{T} \setminus \mathcal{T}_*} \|f - P_{\mathcal{T}} f\|_{H^{-1}(\omega_T)}^2,$$

the oscillation of  $f$  is therefore bounded by

$$|\langle f - P_{\mathcal{T}} f, w - I_{\mathcal{T}} w \rangle| \leq \text{osc}_{\mathcal{T}}(f, \mathcal{T} \setminus \mathcal{T}_*) \|\nabla w\|_{L^2(\Omega)}$$

and the proof is complete.  $\square$

Proposition 4.12 (partial lower bound) as well as Lemma 4.35 (splitting of local residual norm) illustrate that the test space  $\mathbb{V}_{\mathcal{T}}^+$  is closely related to lower bounds for the error. This observation suggests establishing lower bounds for the correction  $\|\nabla(u_{\mathcal{T}} - u_{\mathcal{T}_*})\|_{L^2(\Omega)}$  by ensuring conditions like

$$\mathbb{V}^+(\mathcal{T}_{\omega}) \subset \mathbb{V}(\mathcal{T}_*),$$

where  $\omega$  is a  $\mathcal{T}$ -mesh subdomain. Inspecting the construction of  $\mathbb{V}_{\mathcal{T}}^+$ , we realize that such conditions can be achieved if  $\max\{m_1, m_2\} \leq n - 1$  and the cut-off is implemented with hat functions of a virtual refinement of  $\mathcal{T}$ .

**Lemma 4.49 (cut-off by refined hat functions)** *Let  $\mathcal{T}_+$  be the minimal bisection refinement of  $\mathcal{T}$  such that the relative interior of each element  $T \in \mathcal{T}$  and each face  $F \in \mathcal{F}$  of the original mesh  $\mathcal{T}$  contains at least one vertex from  $\mathcal{T}_+$ . Then there exist hat functions  $\phi_T$ ,  $T \in \mathcal{T}$ , and  $\phi_F$ ,  $F \in \mathcal{F}$ , in  $\mathbb{S}^{1,0}(\mathcal{T}_+)$  satisfying Assumption 4.21 if  $\max\{m_1, m_2\} \leq n - 1$ .*

*Proof.* The details of the proof depend on bisection and we therefore restrict to the case  $d = 2$ ; for  $d > 2$ , the following reference situation used to define the hat functions is replaced by several ones with “tagged” reference simplices. Let  $\widehat{T} = T_2$  be the reference element in  $\mathbb{R}^2$  with the standard enumeration of its vertices  $\widehat{z}_0 = 0$ ,  $\widehat{z}_1 = e_1$ , and  $\widehat{z}_2 = e_2$ . Furthermore, let  $\widehat{\mathcal{T}}_+$  be the mesh obtained by applying 5 bisections so that vertices in the interiors of  $\widehat{T}$  and of its faces are generated. Denote by  $\widehat{\phi}_{\widehat{T}}$ ,  $\widehat{\phi}_{F'}$ ,  $F' \subset \widehat{T}$ , the four hat functions in  $\mathbb{S}^{1,0}(\widehat{\mathcal{T}}_+)$  associated with these generated vertices. Given an arbitrary element  $T \in \mathcal{T}$ , let  $H_T$  denote the bi-affine map  $T \rightarrow \widehat{T}$  preserving the numbering of the vertices for bisection and define the pull-backs

$$\phi_T := H_T^*(\widehat{\phi}_{\widehat{T}}), \quad \phi_F|_T := H_T^*(\widehat{\phi}_{H_T(F)}), \quad F \subset T,$$

and extend by 0 off  $T$  or  $\omega_F$ . As the extension operators  $E_F$  preserve the polynomial degree, see Lemma 4.20 (extending from faces),  $\max\{m_1, m_2\} \leq n - 1$ , and

$$\{H_T^{-1}(\widehat{T}_+) \mid \widehat{T}_+ \in \widehat{\mathcal{T}}_+\} = \{T_+ \in \mathcal{T}_+ \mid T_+ \subset T\},$$

the hat functions  $\phi_T$ ,  $T \in \mathcal{T}$ , and  $\phi_F$ ,  $F \in \mathcal{F}$ , then satisfy Assumption 4.21 with  $G_T = H_T^{-1}$  and  $\mathbb{S}^+ = \mathbb{S}^{n,0}(\widehat{\mathcal{T}}_+)$ .  $\square$

**Definition 4.50 (interior vertex property)** A mesh element  $T \in \mathcal{T}$  satisfies the *interior vertex property* with respect to  $\mathcal{T}_* \geq \mathcal{T}$  whenever each interior face  $F \subset \partial T \setminus \partial\Omega$  of  $T$  and each element in  $\widetilde{\omega}_T$  (defined in (3.12)) have in their relative interiors at least one vertex from  $\mathcal{T}_*$ .

A set  $\mathcal{M} \subset \mathcal{T}$  satisfies the interior vertex property with respect to a refinement  $\mathcal{T}_* \geq \mathcal{T}$  if each element  $T \in \mathcal{M}$  satisfies the interior vertex property.

The interior vertex property is valid upon enforcing a fixed number  $b$  of bisections ( $b = 3, 6$  for  $d = 2, 3$ ). An immediate consequence is the following lower bound for corrections.

**Theorem 4.51 (lower bound for corrections)** Suppose  $\mathbf{A}$  is piecewise constant over  $\mathcal{T}$  and  $c = 0$ , define  $P_{\mathcal{T}}$  with the help of the cut-off functions in Lemma 4.49, and denote by  $\mathcal{M}$  the subset of elements in  $\mathcal{T}$  satisfying the interior vertex property with respect to  $\mathcal{T}_*$ . Then

$$\sum_{T \in \mathcal{M}} \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f, T)^2 \leq C \|\nabla(u_{\mathcal{T}} - u_{\mathcal{T}_*})\|_{L^2(\Omega)}^2 + \sum_{T \in \mathcal{M}} \|f - P_{\mathcal{T}}f\|_{H^{-1}(\omega_T)}^2,$$

where  $C$  depends on  $d$ , the shape regularity coefficient  $\sigma$ , the coefficients  $(\mathbf{A}, c)$ , and on the polynomial degree  $n$ .

*Proof.* [1] We first show a local bound with the PDE indicator  $\eta_{\mathcal{T}}(u_{\mathcal{T}}, T)$ . In view of  $\mathbf{A} \in \mathbb{S}_{\mathcal{T}}^{0,-1}$  and  $c = 0$ , we choose  $m_1 = n - 1$  and  $m_2 = n - 2$  as the degrees for the discrete functionals. We can thus apply Lemma 4.49 and construct  $P_{\mathcal{T}}$  with the refined hat functions. Let  $T \in \mathcal{M}$  and so, using the interior vertex property and the notation associated with  $\widetilde{\omega}_T$  in (3.12), we deduce  $\mathbb{V}^+(\widetilde{\mathcal{T}}_T) \subset \mathbb{V}(\mathcal{T}_*, T) := \mathbb{V}(\mathcal{T}_*) \cap H_0^1(\widetilde{\omega}_T)$ , where  $\widetilde{\mathcal{T}}_T := \{T \in \mathcal{T} \mid T \subset \widetilde{\omega}_T\}$ . Combining this with inequality (4.45) and Definition 4.24 (projection onto discrete functionals), we conclude

$$\eta_{\mathcal{T}}(u_{\mathcal{T}}, T) \lesssim \|P_{\mathcal{T}}R_{\mathcal{T}}\|_{\mathbb{V}^+(\widetilde{\mathcal{T}}_T)^*} = \|R_{\mathcal{T}}\|_{\mathbb{V}^+(\widetilde{\mathcal{T}}_T)^*} \leq \|R_{\mathcal{T}}\|_{\mathbb{V}(\mathcal{T}_*, T)^*}.$$

[2] To collect the local bounds of the first step, we first show that, for any  $\ell \in \mathbb{V}(\mathcal{T}_*)^*$ ,

$$\sum_{T \in \mathcal{T}} \|\ell\|_{\mathbb{V}(\mathcal{T}_*, T)^*}^2 \leq (d+2) \|\ell\|_{\mathbb{V}(\mathcal{T}_*)^*}^2.$$

To this end, we just repeat the proof of Lemma 4.5 (localization of  $H^{-1}$ -norm), replacing the spaces  $H_0^1(\omega_i)$  and  $H_0^1(\Omega)$ , respectively, with  $\mathbb{V}(\mathcal{T}_*, T)$  and  $\mathbb{V}(\mathcal{T}_*)$ . Hence, squaring and summing the bound of the first step as well as using Lemma 4.1 (error and residual) yield

$$\sum_{T \in \mathcal{M}} \eta_{\mathcal{T}}(u_{\mathcal{T}}, T)^2 \lesssim \sum_{T \in \mathcal{M}} \|R_{\mathcal{T}}\|_{\mathbb{V}(\mathcal{T}_*, T)^*}^2 \lesssim \|\nabla(u_{\mathcal{T}_*} - u_{\mathcal{T}})\|_{L^2(\Omega)}^2. \quad (4.66)$$



[3] We finally prove the claimed bound by simply inserting (4.66):

$$\begin{aligned} \sum_{T \in \mathcal{M}} \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f, T)^2 &= \sum_{T \in \mathcal{M}} (\eta_{\mathcal{T}}(u_{\mathcal{T}}, T)^2 + \text{osc}_{\mathcal{T}}(f, T)^2) \\ &\leq C \|\nabla(u_{\mathcal{T}_*} - u_{\mathcal{T}})\|_{L^2(\Omega)}^2 + \sum_{T \in \mathcal{M}} \text{osc}_{\mathcal{T}}(f, T)^2 \end{aligned}$$

and the proof is finished.  $\square$

**Remark 4.52 (oscillation and correction)** In general, fixing first the finer mesh  $\mathcal{T}_*$ , it is impossible to bound oscillation indicators  $\text{osc}_{\mathcal{T}}(f, T)$  by some suitable correction. Indeed, these indicators can contain contributions to  $f$  and so to  $R_{\mathcal{T}}$  of “arbitrarily high frequency”, while the correction can control only contributions of the residual  $R_{\mathcal{T}}$  with frequencies representable over  $\mathcal{T}_*$ ; cf. (4.58).

Monotonicity properties of the error  $|u - u_{\mathcal{T}}|_{H_0^1(\Omega)} = \|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)}$  and the PDE error estimator  $\eta_{\mathcal{T}}(u_{\mathcal{T}}) = \eta_{\mathcal{T}}(u_{\mathcal{T}}, f)$  with respect to  $\mathcal{T}$  would be useful but fail to hold. To investigate this issue, we consider two admissible meshes  $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$ , the latter being a refinement of the former  $\mathcal{T}_* \geq \mathcal{T}$ , and a third admissible mesh  $\widehat{\mathcal{T}} \leq \mathcal{T}$ . We further assume that data  $\mathcal{D} = (A, c, f)$  is discrete over  $\widehat{\mathcal{T}}$  in the sense that  $\mathcal{D} \in \mathbb{D}_{\widehat{\mathcal{T}}}$  where

$$\mathbb{D}_{\widehat{\mathcal{T}}} := [\mathbb{S}_{\widehat{\mathcal{T}}}^{n-1, -1}]^{d \times d} \times \mathbb{S}_{\widehat{\mathcal{T}}}^{n-1, -1} \times \mathbb{F}_{\widehat{\mathcal{T}}}$$

and  $\mathcal{D}$  does not change in the transition from  $\mathcal{T}$  to  $\mathcal{T}_*$  irrespective of the degree of local refinement; in particular  $f = P_{\widehat{\mathcal{T}}} f \in \mathbb{F}_{\widehat{\mathcal{T}}}$ . We will later denote discrete data as  $\widehat{\mathcal{D}} = (\widehat{A}, \widehat{c}, \widehat{f})$  to distinguish it from exact data  $\mathcal{D}$ , and to study their discrepancy, but we prefer to keep the simple notation  $\mathcal{D} = \widehat{\mathcal{D}}$  now because there is no reason for confusion. In particular, this implies that the bilinear form in (2.8) and forcing function are the same for both Galerkin solutions  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}$  and  $u_{\mathcal{T}_*} \in \mathbb{V}_{\mathcal{T}_*}$ , whence the energy errors are monotone according to (3.8)

$$\|u - u_{\mathcal{T}_*}\|_{\Omega} \leq \|u - u_{\mathcal{T}}\|_{\Omega},$$

but not  $|u - u_{\mathcal{T}}|_{H_0^1(\Omega)}$ . Moreover,  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$  is not monotone because the discrete solution  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}$  changes with the mesh. It is thus useful to quantify the behavior of  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$  in terms of  $\mathcal{T}$  and  $u_{\mathcal{T}}$  following [Cascón *et al.* 2008]; see also [Morin, Siebert and Veeser 2008]. We do this next.

The first lemma exploits the structure of the PDE residual estimator, namely the presence of a positive power of the local meshsize, and expresses the *reduction* of  $\eta_{\mathcal{T}_*}(v, f)$  relative to  $\eta_{\mathcal{T}}(v, f)$  for fixed functions  $v \in \mathbb{V}_{\mathcal{T}}$  and  $f \in \mathbb{F}_{\mathcal{T}}$ . This quantitative property is instrumental in studying convergence of AFEMs for coercive problems in Section 6 as well as discontinuous Galerkin methods in Section 9 and inf-sup stable problems in Section 10.

**Lemma 4.53 (reduction property of the estimator)** *If the elements of  $\mathcal{M} \subset \mathcal{T}$*

are bisected at least  $b \geq 1$  times to refine  $\mathcal{T}$  into  $\mathcal{T}_*$ , and  $\lambda = 1 - 2^{-b/d}$ , then

$$\eta_{\mathcal{T}_*}(v, f, \mathcal{T}_*)^2 \leq \eta_{\mathcal{T}}(v, f, \mathcal{T})^2 - \lambda \eta_{\mathcal{T}}(v, f, \mathcal{M})^2 \quad \forall v \in \mathbb{V}_{\mathcal{T}}, f \in \mathbb{F}_{\mathcal{T}}. \quad (4.67)$$

*Proof.* Given  $T \in \mathcal{T}$ , we rewrite (4.52) as follows

$$\eta_{\mathcal{T}}(v, T)^2 = h_T j_{\mathcal{T}}(v, T)^2 + h_T^2 r_{\mathcal{T}}(v, T)^2$$

with  $\eta_{\mathcal{T}}(v, T) = \eta_{\mathcal{T}}(v, f, T)$  and

$$j_{\mathcal{T}}(v, T)^2 = j_{\mathcal{T}}(v, f, T)^2 = \sum_{\substack{F \in \mathcal{F} \\ F \subset \partial T}} \| [A \nabla v] \cdot \mathbf{n}_F - P_{\mathcal{T}} f \|_{L^2(F)}^2,$$

$$r_{\mathcal{T}}(v, T) = r_{\mathcal{T}}(v, f, T) = \| P_{\mathcal{T}} f + \operatorname{div}(A \nabla v) - cv \|_{L^2(T)},$$

where  $f = P_{\mathcal{T}} f = P_{\mathcal{T}_*} f \in \mathbb{F}_{\mathcal{T}}$  does not change from  $\mathcal{T}$  to  $\mathcal{T}_*$ . We readily have

$$\sum_{T_* \in \mathcal{T}_*, T_* \subseteq T} \eta_{\mathcal{T}_*}(v, T_*)^2 \leq \eta_{\mathcal{T}}(v, T)^2.$$

because  $h_{T_*} \leq h_T$  for all  $T_* \subset T$  and  $T_* \in \mathcal{T}_*$ . If, in addition,  $T$  is bisected at least  $b$  times, then any such  $T_*$  satisfies  $h_{T_*} \leq 2^{-\frac{b}{d}} h_T$ , whence

$$\sum_{T_* \in \mathcal{T}_*, T_* \subseteq T} \eta_{\mathcal{T}_*}(v, T_*)^2 \leq 2^{-\frac{b}{d}} \eta_{\mathcal{T}}(v, T)^2.$$

Therefore, adding over  $T \in \mathcal{T}$  we obtain

$$\eta_{\mathcal{T}_*}(v)^2 = \sum_{T \in \mathcal{T}} \sum_{T_* \in \mathcal{T}_*, T_* \subseteq T} \eta_{\mathcal{T}_*}(v, T_*)^2 \leq 2^{-\frac{b}{d}} \sum_{T \in \mathcal{M}} \eta_{\mathcal{T}}(v, T)^2 + \sum_{T \in \mathcal{T} \setminus \mathcal{M}} \eta_{\mathcal{T}}(v, T)^2$$

which implies the assertion (4.67).  $\square$

The next result complements Lemma 4.53 in that it expresses the Lipschitz continuity of  $\eta_{\mathcal{T}}(v, f)$  with respect to the argument  $v \in \mathbb{V}_{\mathcal{T}}$  for fixed  $\mathcal{T}$  and  $f \in \mathbb{F}_{\mathcal{T}}$ .

**Lemma 4.54 (Lipschitz property of the estimator)** *Let  $\mathcal{T}$  and  $f \in \mathbb{F}_{\mathcal{T}}$  be fixed. There exists a constant  $C_{\text{Lip}}$  proportional to  $\|A\|_{L^\infty(\Omega)} + \|c\|_{L^\infty(\Omega)}$  such that*

$$|\eta_{\mathcal{T}}(v, f) - \eta_{\mathcal{T}}(w, f)| \leq C_{\text{Lip}} |v - w|_{H_0^1(\Omega)} \quad \forall v, w \in \mathbb{V}_{\mathcal{T}}. \quad (4.68)$$

*Proof.* Since  $\eta_{\mathcal{T}}(v) = \eta_{\mathcal{T}}(v, f)$  is the  $\ell^2$ -norm of the vector  $(\eta_{\mathcal{T}}(v, T))_{T \in \mathcal{T}} \in \mathbb{R}^{\#\mathcal{T}}$ , applying the triangle inequality gives

$$\begin{aligned} |\eta_{\mathcal{T}}(v) - \eta_{\mathcal{T}}(w)|^2 &\leq \sum_{T \in \mathcal{T}} |\eta_{\mathcal{T}}(v, T) - \eta_{\mathcal{T}}(w, T)|^2 \\ &\leq \sum_{T \in \mathcal{T}} h_T |j_{\mathcal{T}}(v, T) - j_{\mathcal{T}}(w, T)|^2 + h_T^2 |r_{\mathcal{T}}(v, T) - r_{\mathcal{T}}(w, T)|^2. \end{aligned}$$

We first consider the jump terms and apply again the triangle inequality followed

by an inverse estimate to find that

$$\begin{aligned} |j_{\mathcal{T}}(v, T) - j_{\mathcal{T}}(w, T)|^2 &\leq \sum_{\substack{F \in \mathcal{F} \\ F \subset \partial T}} \| [\mathbf{A} \nabla(v - w)] \cdot \mathbf{n}_F \|_{L^2(F)}^2 \\ &\lesssim h_T^{-1} \|\mathbf{A}\|_{L^\infty(\Omega)}^2 \|\nabla(v - w)\|_{L^2(\omega_T)}^2, \end{aligned}$$

where  $\omega_T$  is the patch of  $T$ . A similar reasoning for the element residuals yields

$$\begin{aligned} |r_{\mathcal{T}}(v, T) - r_{\mathcal{T}}(w, T)|^2 &\lesssim \|c(v - w)\|_{L^2(T)}^2 + \|\operatorname{div}(\mathbf{A} \nabla(v - w))\|_{L^2(T)}^2 \\ &\lesssim \|c\|_{L^\infty(\Omega)}^2 \|(v - w)\|_{L^2(T)}^2 + h_T^{-2} \|\mathbf{A}\|_{L^\infty(\Omega)}^2 \|\nabla(v - w)\|_{L^2(T)}^2, \end{aligned}$$

because  $A$  is piecewise polynomial. Finally, adding over  $T \in \mathcal{T}$  and applying Lemma 2.2 (first Poincaré inequality) concludes the proof.  $\square$

Since the estimator  $\eta_{\mathcal{T}}(v, f)$  depends explicitly on  $P_{\mathcal{T}}f$ , and  $P_{\mathcal{T}}f$  may change with  $\mathcal{T}$ , it is crucial to account for the variations of  $\eta_{\mathcal{T}}(v, f)$  while keeping  $\mathcal{T}$  and  $v \in \mathbb{V}_{\mathcal{T}}$  fixed. This is the purpose of our next result.

**Lemma 4.55 (estimator dependence on discrete forcing)** *Let  $\mathcal{T}$  and  $v \in \mathbb{V}_{\mathcal{T}}$  be fixed. Then there exists a constant  $C_{\text{Lip}}$  such that*

$$|\eta_{\mathcal{T}}(v, f) - \eta_{\mathcal{T}}(v, g)| \leq C_{\text{Lip}} \left( \sum_{T \in \mathcal{T}} \|f - g\|_{H^{-1}(\omega_T)}^2 \right)^{1/2} \quad \forall f, g \in \mathbb{F}_{\mathcal{T}}. \quad (4.69)$$

*Proof.* We proceed elementwise, as in Lemma 4.54, except that after applying the triangle inequality we end up with the weighted  $L^2$ -norms

$$h_T^2 \|f - g\|_{L^2(T)}^2 + h_T \|f - g\|_{L^2(\partial T)}^2, \quad \forall T \in \mathcal{T}.$$

Extending these norms to patches  $\omega_T$  and its interior faces  $\sigma_T$ , and appealing to Corollary 4.30 (quantifying  $H^{-1}$ -norms of discrete functionals), we deduce

$$h_T^2 \|f - g\|_{L^2(\omega_T)}^2 + h_T \|f - g\|_{L^2(\sigma_T)}^2 \approx \|f - g\|_{H^{-1}(\omega_T)}^2,$$

Adding over  $T \in \mathcal{T}$  finishes the proof.  $\square$

In the subsequent applications of Lemma 4.54 the discrete coefficients  $(\mathbf{A}, c)$  may change with the change of the supporting mesh  $\widehat{\mathcal{T}}$  but they will always be uniformly bounded in  $L^\infty(\Omega)$ ; hence the constant  $C_{\text{Lip}}$  is uniformly bounded as well. Upon combining Lemmas 4.53, 4.54 and 4.55 we obtain the following crucial property.

**Proposition 4.56 (estimator reduction)** *Given  $\mathcal{T} \in \mathbb{T}$  and a subset  $\mathcal{M} \subset \mathcal{T}$  of elements marked for refinement, let  $\text{REFINE}$  be the procedure discussed in Section 3.5 that bisects the elements of  $\mathcal{M}$  at least  $b$  times and  $\mathcal{T}_* = \text{REFINE}(\mathcal{T}, \mathcal{M})$  be the resulting conforming mesh. Let the coefficients  $(\mathbf{A}, c)$  be discrete and fixed. Then for  $\lambda = 1 - 2^{-b/d}$ , for all  $v \in \mathbb{V}_{\mathcal{T}}$ ,  $v_* \in \mathbb{V}_{\mathcal{T}_*}$ ,  $f \in \mathbb{F}_{\mathcal{T}}$ ,  $f_* \in \mathbb{F}_{\mathcal{T}_*}$ , and any  $\delta > 0$*

$$\eta_{\mathcal{T}_*}(v_*, f_*)^2 \leq (1 + \delta) (\eta_{\mathcal{T}}(v, f, \mathcal{T})^2 - \lambda \eta_{\mathcal{T}}(v, f, \mathcal{M})^2)$$

$$+ 2(1 + \delta^{-1}) C_{Lip}^2 \left( |v_* - v|_{H_0^1(\Omega)}^2 + \sum_{T_* \in \mathcal{T}_*} \|f_* - f\|_{H^{-1}(\omega_{T_*})}^2 \right),$$

where  $C_{Lip}$  is the constant in Lemmas 4.54 and 4.55.

*Proof.* For any  $\delta > 0$ , write

$$\eta_{\mathcal{T}_*}(v_*, f_*, \mathcal{T}_*)^2 \leq (1 + \delta) \eta_{\mathcal{T}_*}(v, f, \mathcal{T}_*)^2 + (1 + \delta^{-1}) (\eta_{\mathcal{T}_*}(v_*, f_*, \mathcal{T}_*) - \eta_{\mathcal{T}_*}(v, f, \mathcal{T}_*))^2$$

and apply Lemma 4.53 to the first term and Lemmas 4.54 and 4.55 to the second one combined with a triangle inequality.  $\square$

We finish this section by investigating the behavior of the global oscillation under refinement.

**Lemma 4.57 (quasi-monotonicity of oscillation)** *If  $f \in H^{-1}(\Omega)$  and  $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$  with  $\mathcal{T}_* \geq \mathcal{T}$ , then*

$$\text{osc}_{\mathcal{T}_*}(f) \leq C_{\text{osc}} \text{osc}_{\mathcal{T}}(f),$$

where  $C_{\text{osc}}$  depends only on the shape regularity coefficient  $\sigma$  and  $d$ .

*Proof.* Given  $T \in \mathcal{T}$ , let  $T_* \in \mathcal{T}_*$  such that  $T_* \subset T$ . Since  $\mathcal{T}_*$  is a refinement of  $\mathcal{T}$ , this implies that the patch  $\omega_{\mathcal{T}_*}(T_*)$  in  $\mathcal{T}_*$  around  $T_*$  is contained in the patch  $\omega_{\mathcal{T}}(T)$  in  $\mathcal{T}$  around  $T$ . Thanks to Lemma 4.31 (local near-best approximation), we derive

$$\text{osc}_{\mathcal{T}_*}(f, T_*)^2 = \|(I - P_{\mathcal{T}_*})f\|_{H^{-1}(\omega_{\mathcal{T}_*}(T_*))}^2 \leq C_{\text{Istb}}^2 \|(I - P_{\mathcal{T}})f\|_{H^{-1}(\omega_{\mathcal{T}}(T))}^2$$

and therefore, with the help of (ii) of Lemma 4.5 (localization of  $H^{-1}$ -norm),

$$\sum_{T_* \subset T} \text{osc}_{\mathcal{T}_*}(f, T_*)^2 \leq C_{\text{Istb}}^2 C_{\text{ovrl}} \|(I - P_{\mathcal{T}})f\|_{H^{-1}(\omega_{\mathcal{T}}(T))}^2 = C_{\text{Istb}}^2 C_{\text{ovrl}} \text{osc}_{\mathcal{T}}(f, T)^2,$$

where  $C_{\text{ovrl}}$  is bounded in terms of the shape regularity coefficient  $\sigma$  and  $d$ . Hence, summing over  $T \in \mathcal{T}$  yields

$$\text{osc}_{\mathcal{T}_*}(f)^2 = \sum_{T \in \mathcal{T}} \sum_{T_* \subset T} \text{osc}_{\mathcal{T}_*}(f, T_*)^2 \leq C_{\text{Istb}}^2 C_{\text{ovrl}} \text{osc}_{\mathcal{T}}(f)^2$$

and the proof is finished.  $\square$

#### 4.9 Alternative estimators

In Sect. 4.7 we used the  $H^{-1}$ -projection  $P_{\mathcal{T}}$  to derive a posteriori bounds for the error in the spirit of the standard residual estimator. The goal of this section is to illustrate that the approach with  $P_{\mathcal{T}}$  can be combined also with other techniques of a posteriori error estimation, generalizing and expanding the discussion in [Kreuzer and Veese 2021, Sect. 4] with the  $H^{-1}$ -projection  $P_{\mathcal{T}}$ .

Alternative techniques have been developed with the desire to reduce or even circumvent that constants spoil the relationship between error and estimator. In the

framework of the aforementioned approach, we shall see that the various techniques based upon

- *local (discrete) problems*,
- *hierarchy*,
- *flux equilibration*

amount to different ways of quantifying a local norm of the discretized residual  $P_{\mathcal{T}}R_{\mathcal{T}}$ . This observation is useful for comparing the techniques and for a common treatment in the following sections about adaptive algorithms.

As in Sect. 4.7 on modified residual estimation, we shall consider only the case given by Assumption 4.44 (discrete coefficients and discrete functionals). For the hidden constants in the results of this section, it is useful to keep in mind Remark 4.4 (constants in error-residual relationship).

Theorem 4.45 (modified residual estimator) analyzed an element-indexed version of the residual estimator. For the sake of simplicity, we shall refrain here from such an element-indexed setting and remain in the vertex-indexed setting of the abstract analysis of Sect. 4.6. In order to facilitate the comparison with the other estimators below, we offer the following vertex-indexed variant of Theorem 4.45. For  $z \in \mathcal{V}$ , we set  $\mathcal{F}_z := \{F \in \mathcal{F} \mid z \in F\}$  and  $\mathcal{T}_z := \{T \in \mathcal{T} \mid z \in T\}$ . Given the Galerkin approximation  $u_{\mathcal{T}}$  from (4.1), define the PDE indicator by

$$\begin{aligned} \eta_{\mathcal{T}}^{\text{res}}(u_{\mathcal{T}}) &:= \sum_{z \in \mathcal{V}} \eta_{\mathcal{T}}^{\text{res}}(u_{\mathcal{T}}, z)^2 \quad \text{with} \\ \eta_{\mathcal{T}}^{\text{res}}(u_{\mathcal{T}}, z)^2 &:= \sum_{F \in \mathcal{F}_z} h_F \|P_F R_{\mathcal{T}}\|_{L^2(F)}^2 + \sum_{T \in \mathcal{T}_z} h_T^2 \|P_T R_{\mathcal{T}}\|_{L^2(T)}^2, \end{aligned} \quad (4.70a)$$

where  $R_{\mathcal{T}} = f + \text{div}(\mathbf{A} \nabla u_{\mathcal{T}}) - cu_{\mathcal{T}} \in H^{-1}(\Omega)$  is the residual and  $P_F$ ,  $F \in \mathcal{F}$  and  $P_T$ ,  $T \in \mathcal{T}$  yield the polynomial densities of  $P_{\mathcal{T}}$ ; see Definition 4.24. The vertex-indexed modified residual estimator is then

$$\mathcal{E}_{\mathcal{T}}^{\text{res}} := \mathcal{E}_{\mathcal{T}}^{\text{res}}(u_{\mathcal{T}}, f)^2 := \eta_{\mathcal{T}}^{\text{res}}(u_{\mathcal{T}})^2 + \text{osc}_{\mathcal{T}}(f)^2. \quad (4.70b)$$

**Theorem 4.58 (vertex-indexed modified residual estimator)** *Suppose Assumption 4.44. The modified residual estimator (4.70b) is equivalent to the error:*

$$\frac{\min\{1, C_{\text{L},\text{res}}\}}{C_{\text{IStb}}^{\star} C_d} \mathcal{E}_{\mathcal{T}}^{\text{res}} \lesssim \|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)} \lesssim \max\{1, C_{\text{U},\text{res}}\} C_d C_{\text{loc}} \mathcal{E}_{\mathcal{T}}^{\text{res}},$$

while its PDE indicator (4.70a) is locally equivalent to the discretized residual: for all vertices  $z \in \mathcal{V}$ ,

$$C_{\text{L},\text{res}} \eta_{\mathcal{T}}^{\text{res}}(u_{\mathcal{T}}, z) \leq \|P_{\mathcal{T}} R_{\mathcal{T}}\|_{H^{-1}(\omega_z)} \leq C_{\text{U},\text{res}} \eta_{\mathcal{T}}^{\text{res}}(u_{\mathcal{T}}, z).$$

Here,  $C_{\text{L},\text{res}}$  and  $C_{\text{U},\text{res}}$  are the hidden constants of Corollary 4.30 on stars,  $C_{\text{IStb}}^{\star}$  is the stability constant of  $P_{\mathcal{T}}$  on stars from Lemma 4.28,  $C_{\text{loc}}$  is the constant

from Corollary 4.6,  $C_d = \sqrt{2(d+1)}$ , and the hidden constants depend only on the error-residual relationship in Lemma 4.1.

*Proof.* Local equivalence is a reformulation of Corollary 4.30 (quantifying  $H^{-1}$ -norms of discrete functionals) on stars. The global bounds follow by combining local equivalence with Lemma 4.1 (error and residual), Corollary 4.6 (star localization of residual norm), and Lemma 4.35 (splitting of local residual norms).  $\square$

### Adjoint projection

The projection  $P_{\mathcal{T}}$  relates residual  $R_{\mathcal{T}}$  and discretized residual  $P_{\mathcal{T}}R_{\mathcal{T}}$ . In order to exploit this relationship on the test space  $H_0^1(\Omega)$ , we shall need the adjoint  $P_{\mathcal{T}}^*$  to the projection  $P_{\mathcal{T}}$ . Curiously, operators employed in this vein appeared first; see, e.g., [Morin, Nochetto and Siebert 2003] and [Veese 2002].

Given  $w \in H_0^1(\Omega)$ , the function  $P_{\mathcal{T}}^*w$  can be directly defined by requiring

$$P_{\mathcal{T}}^*w \in \mathbb{V}_{\mathcal{T}}^+ : \quad \langle \ell, P_{\mathcal{T}}^*w \rangle = \langle \ell, w \rangle \quad \forall \ell \in \mathbb{F}_{\mathcal{T}} \quad (4.71)$$

This definition is well-posed thanks to Lemma 3.1 (discrete inf-sup condition) and Lemma 4.25 (algebraic properties), especially (4.37) and (4.38). Clearly,  $P_{\mathcal{T}}^*$  is a linear projection onto the finite-dimensional subspace  $\mathbb{V}_{\mathcal{T}}^+ \subset H_0^1(\Omega)$ . A representation as interpolation operator will be derived in Corollary 4.61 below. Using both definitions of  $P_{\mathcal{T}}$  and  $P_{\mathcal{T}}^*$ , we see that they are actually adjoint:

$$\langle P_{\mathcal{T}}\ell, w \rangle = \langle P_{\mathcal{T}}\ell, P_{\mathcal{T}}^*w \rangle = \langle \ell, P_{\mathcal{T}}^*w \rangle \quad \forall \ell \in H^{-1}(\Omega), w \in H_0^1(\Omega). \quad (4.72)$$

Consequently, Lemmas 4.25 and 4.28 (local  $H^{-1}$  stability) show that  $P_{\mathcal{T}}^*$  is a local operator with

$$\|P_{\mathcal{T}}^*\|_{\mathcal{L}(H_0^1(\omega))} = \|P_{\mathcal{T}}\|_{\mathcal{L}(H^{-1}(\omega))} \leq C_{\text{lstb}}. \quad (4.73)$$

The choice  $\ell = R_{\mathcal{T}}$  in (4.72) leads to

$$\langle P_{\mathcal{T}}R_{\mathcal{T}}, w \rangle = \langle P_{\mathcal{T}}R_{\mathcal{T}}, P_{\mathcal{T}}^*w \rangle = \langle R_{\mathcal{T}}, P_{\mathcal{T}}^*w \rangle \quad \forall w \in H_0^1(\Omega),$$

where the two identities show that the discretized residual  $P_{\mathcal{T}}R_{\mathcal{T}}$  can be analyzed with discrete test functions in  $\mathbb{V}_{\mathcal{T}}^+$  only, cf. the norm equivalence in Lemma 4.28. Restricting to discrete test functions in  $\mathbb{V}_{\mathcal{T}}^+ = \text{Im } P_{\mathcal{T}}^*$ , we obtain:

$$\langle P_{\mathcal{T}}R_{\mathcal{T}}, w \rangle = \langle R_{\mathcal{T}}, w \rangle \quad \forall w \in \mathbb{V}_{\mathcal{T}}^+. \quad (4.74)$$

### An estimator based upon local problems

Local dual norms can be quantified by solving local problems. Requiring computability of these solutions leads to *finite-dimensional* or *discrete* local problems. In other words, we lift the residual to local and finite-dimensional extensions of the finite element space. Starting with [Babuška and Rheinboldt 1978], this idea was used to soften the impact of constants in the relationship between error and estimator; cf. [Verfürth 2013, Remark 1.22] for more references.

Within the approach of Sections 4.1 and 4.6, we can use local discrete problems

to quantify the local  $H^{-1}$ -norms of the discretized residual  $P_{\mathcal{T}}R_{\mathcal{T}}$ . In this manner, constants arise only due to the localization of the residual norm and to the splitting into discretized and oscillatory residual by the  $H^{-1}$ -projection  $P_{\mathcal{T}}$ .

We start by introducing the vertex-oriented PDE indicator. Given the Galerkin approximation  $u_{\mathcal{T}}$  from (4.1), set

$$\eta_{\mathcal{T}}^{\text{lpb}}(u_{\mathcal{T}}) := \sum_{z \in \mathcal{V}} \eta_{\mathcal{T}}^{\text{lpb}}(u_{\mathcal{T}}, z)^2 \quad \text{with} \quad \eta_{\mathcal{T}}^{\text{lpb}}(u_{\mathcal{T}}, z) := \|\nabla v_z\|_{L^2(\omega_z)}, \quad (4.75a)$$

where  $v_z \in \mathbb{V}^+(\mathcal{T}_z)$  is the solution of the local problem

$$\int_{\omega_z} \nabla v_z \cdot \nabla w = \langle R_{\mathcal{T}}, w \rangle \quad \forall w \in \mathbb{V}^+(\mathcal{T}_z). \quad (4.75b)$$

Note that this problem is discrete for  $\dim \mathbb{V}^+(\mathcal{T}_z) < \infty$  and therefore can be solved up to machine precision. The resulting estimator is then

$$\mathcal{E}_{\mathcal{T}}^{\text{lpb}} := \mathcal{E}_{\mathcal{T}}^{\text{lpb}}(u_{\mathcal{T}}, f)^2 := \eta_{\mathcal{T}}^{\text{lpb}}(u_{\mathcal{T}})^2 + \text{osc}_{\mathcal{T}}(f)^2. \quad (4.75c)$$

**Theorem 4.59 (estimator based on local problems)** *Under Assumption 4.44 the estimator (4.75) based on local problems is equivalent to the error, while its PDE indicator is locally equivalent to the discretized residual with constant 1 in the lower bound so that*

$$\frac{1}{C_{\text{lstb}}^{\star} C_d} \mathcal{E}_{\mathcal{T}}^{\text{lpb}} \lesssim \|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)} \lesssim C_{\text{lstb}}^{\star} C_d C_{\text{loc}} \mathcal{E}_{\mathcal{T}}^{\text{lpb}}$$

and, for all vertices  $z \in \mathcal{V}$ ,

$$\eta_{\mathcal{T}}^{\text{lpb}}(u_{\mathcal{T}}, z) \leq \|P_{\mathcal{T}}R_{\mathcal{T}}\|_{H^{-1}(\omega_z)} \leq C_{\text{lstb}}^{\star} \eta_{\mathcal{T}}^{\text{lpb}}(u_{\mathcal{T}}, z)$$

where  $C_{\text{lstb}}^{\star}$  is the stability constant of  $P_{\mathcal{T}}$  on stars from Lemma 4.28,  $C_{\text{loc}}$  from Corollary 4.6,  $C_d = \sqrt{2(d+1)}$  and the hidden constants depend only on the error-residual relationship in Lemma 4.1.

*Proof.* It suffices to show the local equivalence for the PDE indicator; cf. Theorem 4.58 (vertex-indexed modified residual estimator) and note  $C_{\text{lstb}}^{\star} \geq 1$ . Let  $z \in \mathcal{V}$  be any vertex. In view of (4.74), the definition of  $v_z \in \mathbb{V}^+(\mathcal{T})$  readily implies

$$\eta_{\mathcal{T}}^{\text{lpb}}(u_{\mathcal{T}}, z) = \|\nabla v_z\|_{L^2(\omega_z)} = \|P_{\mathcal{T}}R_{\mathcal{T}}\|_{\mathbb{V}^+(\mathcal{T})^*}.$$

Hence Lemma 4.28 on the local  $H^{-1}$ -stability of  $P_{\mathcal{T}}$  yields the asserted local equivalence of PDE indicator and discretized residual  $P_{\mathcal{T}}R_{\mathcal{T}}$ .  $\square$

*A stable biorthogonal system for  $\mathbb{F}_{\mathcal{T}} \times \mathbb{V}_{\mathcal{T}}^+$*

Stable biorthogonal systems induce linear bounded projections, which enjoy near-best approximation thanks to the Lebesgue lemma. Supposing Assumption 4.21 (abstract cut-off), we now outline the construction of such a system for the finite-dimensional product  $\mathbb{F}_{\mathcal{T}} \times \mathbb{V}_{\mathcal{T}}^+$ . The constructed system will induce both projections

$P_{\mathcal{T}}$  and its adjoint  $P_{\mathcal{T}}^*$ . This generalizes the bi-orthogonal system in [Kreuzer and Veiser 2021, Sect. 3.4] to arbitrary degrees of the discrete functionals and provides an alternative approach to  $P_{\mathcal{T}}$ , its local stability as well as its computation. Furthermore, we use it for devising an hierarchical estimator.

The construction is implemented in an affine equivalent manner and our first step consists in setting-up a suitable *reference biorthogonal system*. Let  $\widehat{T} := T_d$  be the reference element,  $\widehat{F} := T_{d-1} \times \{0\} \subset \widehat{T}$  be the reference face, and denote the polynomial degrees in  $\mathbb{F}_{\mathcal{T}}$  by  $m_1 \in \mathbb{N}_0$  and  $m_2 \in \mathbb{N}_0$ . Writing

$$K_1 := \dim \mathbb{P}_{m_1}(\widehat{F}), \quad K_2 := \dim \mathbb{P}_{m_2}(\widehat{T}),$$

assume that we are given orthonormal bases  $q_{(\widehat{F},1)}, \dots, q_{(\widehat{F},K_1)} \in \mathbb{P}_{m_1}(\widehat{F})$  and  $q_{(\widehat{T},1)}, \dots, q_{(\widehat{T},K_2)} \in \mathbb{P}_{m_2}(\widehat{T})$  in the sense that

$$\int_{\widehat{F}} q_{(\widehat{F},k)} q_{(\widehat{F},l)} \phi_{\widehat{F}} = \delta_{kl}, \quad \int_{\widehat{T}} q_{(\widehat{T},k)} q_{(\widehat{T},l)} \phi_{\widehat{T}} = \delta_{kl}. \quad (4.76)$$

for all admissible  $k, l$ , i.e.  $k, l \in \{1, \dots, K_1\}$  or  $\{1, \dots, K_2\}$  depending on the underlying domain. These bases induce the *reference functionals*

$$\widehat{\ell}_{(\widehat{F},k)}(w) := \int_{\widehat{F}} q_{(\widehat{F},k)} w, \quad \widehat{\ell}_{(\widehat{T},k)}(w) := \int_{\widehat{T}} q_{(\widehat{T},k)} w,$$

on  $H^1(\widehat{T})$ , which in turn span the reference space  $\widehat{\mathbb{F}}$ . In order to define complementing test functions, let  $\widehat{E}$  be the extension operator (4.4) associated with the reference face  $\widehat{F}$  adapted to the current situation with the only element  $\widehat{T}$ , and, given some  $v \in L^2(\widehat{T})$ , define  $\widehat{Q}v \in \mathbb{P}_{m_2}$  by

$$\int_{\widehat{T}} q(\widehat{Q}v) \phi_{\widehat{T}} = \int_{\widehat{T}} qv \quad \forall q \in \mathbb{P}_{m_2}. \quad (4.77)$$

We thus define the *reference test functions*

$$\begin{aligned} \widehat{w}_{(\widehat{T},k)} &:= q_{(\widehat{T},k)} \phi_{\widehat{T}}, \quad k = 1, \dots, K_2, \\ \widehat{w}_{(\widehat{F},k)} &:= \widehat{v}_{(\widehat{F},k)} - (\widehat{Q}\widehat{v}_{(\widehat{F},k)}) \phi_{\widehat{T}}, \quad k = 1, \dots, K_1, \end{aligned} \quad (4.78)$$

with  $\widehat{v}_{(\widehat{F},k)} := (\widehat{E}q_{(\widehat{F},k)}) \phi_{\widehat{F}}$ . Note  $\widehat{w}_{(\widehat{F},k)} \neq 0$ . Writing  $\widehat{I} := \{(\widehat{F},k) \mid k = 1, \dots, K_1\} \cup \{(\widehat{T},k) \mid k = 1, \dots, K_2\}$ , we then have

$$\widehat{w}_i \in \widehat{\mathbb{V}}^+ := \left\{ (\widehat{E}q_1) \phi_{\widehat{F}} + q_2 \phi_{\widehat{T}} \mid q_1 \in \mathbb{P}_{m_1}, q_2 \in \mathbb{P}_{m_2} \right\}$$

for all  $i \in \widehat{I}$  and the biorthogonality

$$\forall i, j \in \widehat{I} \quad \langle \widehat{\ell}_i, \widehat{w}_j \rangle = \delta_{ij} \quad (4.79)$$

thanks to (4.76) and (4.77). We thus dispose of a biorthogonal system in the reference product  $\widehat{\mathbb{F}} \times \widehat{\mathbb{V}}^+$ .

Using pull-backs with some minor tweaks, this reference biorthogonal system



induces a global biorthogonal one. To this end, we employ bi-affine maps  $G_F$ ,  $G_T$ , and  $G_{(T,F)}$ . Here, e.g., given a pair  $(T, F) \in \mathcal{T} \times \mathcal{F}$  with  $F \subset T$ , the map  $G_{(T,F)}$  is bi-affine and sends vertices into vertices such that  $G_{(T,F)}(\widehat{T}) = T$  and  $G_{(T,F)}(\widehat{F}) = F$ . The fact that these maps are only unique up to some renumbering of the vertices is irrelevant as all objects in the reference situation on  $(\widehat{T}, \widehat{F})$  are invariant under such renumberings. We denote the respective inverse maps of  $G_F$ ,  $G_T$ , and  $G_{(T,F)}$  by  $H_F$ ,  $H_T$ , and  $H_{(T,F)}$ . Motivated by the transformation rule, we introduce the scaled pull-backs, for  $F \in \mathcal{F}, T \in \mathcal{T}, k$  admissible,

$$q_{(F,k)} := \left( \frac{|\widehat{F}|}{|F|} \right)^{\frac{1}{2}} H_F^* q_{(\widehat{F},k)}, \quad q_{(T,k)} := \left( \frac{|\widehat{T}|}{|T|} \right)^{\frac{1}{2}} H_T^* q_{(\widehat{T},k)}, \quad (4.80)$$

of the reference orthonormal bases in (4.76). These lead to the basis

$$\ell_{(F,k)}(w) := \int_F q_{(F,k)} w, \quad \ell_{(T,k)}(w) := \int_T q_{(T,k)} w, \quad (4.81)$$

of  $\mathbb{F}_{\mathcal{T}}$ , while the associated test functions are given again through pull-backs:

$$w_{(T,k)} := \left( \frac{|\widehat{T}|}{|T|} \right)^{\frac{1}{2}} H_T^* \widehat{w}_{(\widehat{T},k)}, \quad w_{(F,k)|_T} := \left( \frac{|\widehat{F}|}{|F|} \right)^{\frac{1}{2}} H_{(T,F)}^* \widehat{w}_{(\widehat{F},k)} \quad (4.82)$$

for all  $T \in \mathcal{T}, F \in \mathcal{F}$  with  $F \subset T$  and all admissible  $k$ . Note that  $w_{(F,k)} \in H_0^1(\Omega)$ . Finally, we introduce the index set

$$I := (\mathcal{F} \times \{1, \dots, K_1\}) \cup (\mathcal{T} \times \{1, \dots, K_2\}).$$

and observe that  $w_i \in \mathbb{V}_{\mathcal{T}}^+$  for all  $i \in I$ .

**Lemma 4.60 (biorthogonal system)** *The pairs  $(\ell_i, w_i)$ ,  $i \in I$ , provide a stable biorthogonal system of the product  $\mathbb{F}_{\mathcal{T}} \times \mathbb{V}_{\mathcal{T}}^+$ : indeed,  $\langle \ell_i, w_j \rangle = \delta_{ij}$  for all  $i, j \in I$  and, writing  $I_z := \{(S, k) \in I \mid S \ni z\}$  for all  $z \in \mathcal{V}$ ,*

$$\sum_{i \in I_z} \|\ell_i\|_{H^{-1}(\omega_z)} \|\nabla w_i\|_{L^2(\omega_z)} \leq C_{\text{bos}}^*,$$

where the constant  $C_{\text{bos}}^*$  depends only on  $d$ ,  $m_1$ ,  $m_2$  and the shape regularity coefficient  $\sigma$  from (3.9).

*Proof.* [1] We first establish the biorthogonality. Thanks to the transformation rule, the scaled pull-backs (4.80) indeed form local orthonormal bases of  $P_{m_1}(F)$ ,  $F \in \mathcal{F}$ , and  $P_{m_2}(T)$ ,  $T \in \mathcal{T}$ :

$$\int_F q_{(F,k)} q_{(F,l)} \phi_F = \int_{\widehat{F}} q_{(\widehat{F},k)} q_{(\widehat{F},l)} \phi_{\widehat{F}} = \delta_{kl}, \quad \int_T q_{(T,k)} q_{(T,l)} \phi_T = \delta_{kl} \quad (4.83)$$

for all admissible  $k$  and  $l$ . This orthonormality, combined with the local supports of the pairs  $(\ell_i, w_i)$ ,  $i \in I$ , shows the biorthogonality, except for the cases when

$(T, F) \in \mathcal{T} \times \mathcal{F}$  with  $F \subset T$  and  $k, l$  are admissible. Here transformation rule and the definition of  $\widehat{Q}$  imply

$$\begin{aligned} \langle \ell_{(T,k)}, w_{(F,l)} \rangle &= \int_T q_{(T,k)} w_{(F,l)} = \left( \frac{|T| |\widehat{F}|}{|\widehat{T}| |F|} \right)^{\frac{1}{2}} \int_{\widehat{T}} q_{(\widehat{T},k)} \widehat{w}_{(\widehat{F},l)} \\ &= \left( \frac{|T| |\widehat{F}|}{|\widehat{T}| |F|} \right)^{\frac{1}{2}} \int_{\widehat{T}} q_{(\widehat{T},k)} \left( \widehat{v}_{(\widehat{F},l)} - (\widehat{Q} \widehat{v}_{(\widehat{F},l)}) \phi_{\widehat{T}} \right) = 0 \end{aligned}$$

and biorthogonality is verified.

[2] It remains to show the stability bound for any vertex  $z \in \mathcal{V}$ . Given  $i \in I_z$ , we have either  $i = (T, k)$  with  $T \in \mathcal{T}$  or  $i = (F, k)$  with  $F \in \mathcal{F}$ . On the one hand, the functional  $\ell_i$  satisfies

$$\|\ell_{(T,k)}\|_{H^{-1}(\omega_z)} \lesssim h_T \quad \text{or} \quad \|\ell_{(F,k)}\|_{H^{-1}(\omega_z)} \lesssim h_F^{\frac{1}{2}}$$

since, by passing to the reference element and using orthonormality (4.83), (a variant of the) Poincaré inequality (2.2) and the trace inequality on  $\widehat{T}$ , we have

$$\begin{aligned} \langle \ell_{(T,k)}, w \rangle &= \int_T q_{(T,k)} w = \left( \frac{|T|}{|\widehat{T}|} \right)^{\frac{1}{2}} \int_{\widehat{T}} q_{(\widehat{T},k)} G_T^* w \\ &\leq \left( \frac{|T|}{|\widehat{T}|} \right)^{\frac{1}{2}} \|q_{(\widehat{T},k)}\|_{L^2(\widehat{T})} \|G_T^* w\|_{L^2(\widehat{T})} \\ &\lesssim \left( \frac{|T|}{|\widehat{T}|} \right)^{\frac{1}{2}} \left( \int_{\widehat{T}} |q_{(\widehat{T},k)}|^2 \phi_{\widehat{T}} \right) \|\nabla(G_T^* w)\|_{L^2(\widehat{T})} \lesssim h_T \|\nabla w\|_{L^2(T)} \end{aligned}$$

or, with  $T \in \mathcal{T}$  such that  $T \supset F$ ,

$$\begin{aligned} \langle \ell_{(F,k)}, w \rangle &= \int_F q_{(F,k)} w = \left( \frac{|F|}{|\widehat{F}|} \right)^{\frac{1}{2}} \int_{\widehat{F}} q_{(\widehat{F},k)} G_{(T,F)}^* w \\ &\leq \left( \frac{|F|}{|\widehat{F}|} \right)^{\frac{1}{2}} \|q_{(\widehat{F},k)}\|_{L^2(\widehat{F})} \|G_{(T,F)}^* w\|_{L^2(\widehat{F})} \\ &\lesssim \left( \frac{|F|}{|\widehat{F}|} \right)^{\frac{1}{2}} \left( \int_{\widehat{F}} |q_{(\widehat{F},k)}|^2 \phi_{\widehat{F}} \right) \|\nabla(G_{(T,F)}^* w)\|_{L^2(\widehat{T})} \\ &\lesssim \left( \frac{|F| |\widehat{T}|}{|\widehat{F}| |T|} \right)^{\frac{1}{2}} h_T \|\nabla w\|_{L^2(T)} \lesssim h_F^{\frac{1}{2}} \|\nabla w\|_{L^2(T)}. \end{aligned}$$

On the other hand, we obtain that the function  $w_i$  verifies

$$\|\nabla w_{(T,k)}\|_{L^2(T)} = \left( \frac{|\widehat{T}|}{|T|} \right)^{\frac{1}{2}} \|\nabla(H_T^* \widehat{w}_{(\widehat{T},k)})\|_{L^2(T)} \lesssim h_T^{-1} \|\widehat{w}_{(\widehat{T},k)}\|_{L^2(\widehat{T})} \lesssim h_T^{-1}$$

or

$$\begin{aligned} \|\nabla w_{(F,k)}\|_{L^2(T)} &= \left( \frac{|\widehat{F}|}{|F|} \right)^{\frac{1}{2}} \|\nabla(H_{(T,F)}^* \widehat{w}_{(\widehat{T},k)})\|_{L^2(T)} \\ &\lesssim h_T^{-1} \left( \frac{|T|}{|F|} \right)^{\frac{1}{2}} \|\widehat{w}_{(\widehat{T},k)}\|_{L^2(\widehat{T})} \lesssim h_F^{-\frac{1}{2}}. \end{aligned}$$

Using these inequalities to bound  $\|\ell_i\|_{H^{-1}(\omega_z)} \|\nabla w_i\|_{L^2(\omega_z)}$ , summing over all  $i \in I_z$  then establishes the stability bound as the cardinality  $\#I_z$  is uniformly bounded in terms of the shape regularity coefficient  $\sigma$ .  $\square$

**Corollary 4.61 (projections as interpolation operators)** *The biorthogonal system  $(\ell_i, w_i)$ ,  $i \in I$ , induces the  $H^{-1}$ -projection  $P_{\mathcal{T}}$  from Definition 4.24 and its adjoint  $P_{\mathcal{T}}^*$ . Indeed, we have*

$$P_{\mathcal{T}} \ell = \sum_{i \in I} \langle \ell, w_i \rangle \ell_i \quad \text{and} \quad P_{\mathcal{T}}^* w = \sum_{i \in I} \langle \ell_i, w \rangle w_i$$

for all  $\ell \in H^{-1}(\Omega)$  and  $w \in H_0^1(\Omega)$ . The stability of the biorthogonal system then provides an alternative proof of the  $H^{-1}$ -stability on stars of both projection  $P_{\mathcal{T}}$  and  $P_{\mathcal{T}}^*$ , entailing  $C_{\text{ISb}}^* \leq C_{\text{bos}}^*$ .

*Proof.*  $\square$  We only show the identity for  $P_{\mathcal{T}}$ ; the one for  $P_{\mathcal{T}}^*$  can be verified along the same lines. The biorthogonality in Lemma 4.60 readily implies

$$\left\langle \sum_{i \in I} \langle \ell, w_i \rangle \ell_i, w_j \right\rangle = \sum_{i \in I} \langle \ell, w_i \rangle \langle \ell_i, w_j \rangle = \langle \ell, w_j \rangle \quad \forall j \in I.$$

As  $w_j, \mathfrak{x} \in I$  is a basis of  $\mathbb{V}_{\mathcal{T}}^+$ , we conclude the claimed identity for  $P_{\mathcal{T}}$ .

$\square$  To verify the stability statement, we again restrict ourselves to the case of the projection  $P_{\mathcal{T}}$ . Observe first that the proof of the stability of the biorthogonal system does invoke the local stability of  $P_{\mathcal{T}}$ . Thanks to the representation of  $P_{\mathcal{T}}$  and the stability of the biorthogonal system in Lemma 4.60, we have

$$\langle P_{\mathcal{T}} \ell, w \rangle = \sum_{i \in I_z} \langle \ell, w_i \rangle \langle \ell_i, w \rangle \leq C_{\text{bos}}^* \|\ell\|_{H^{-1}(\omega_z)} \|\nabla w\|_{L^2(\omega_z)},$$

where  $I_z := \{(S, k) \in I \mid S \ni z\}$ . The proof is finished.  $\square$

The following two remarks illustrate the practical and theoretical usefulness of the representation formulae.

**Remark 4.62 (alternative computation of  $P_{\mathcal{T}}$ )** A by-product of Corollary 4.61 is a way of computing  $P_{\mathcal{T}}\ell$  for a given functional  $\ell \in H^{-1}(\Omega)$  that “diagonalizes” the approach in Remark 4.26. In fact, given reference orthonormal bases as in (4.76), we can compute the functionals  $\ell_i$ ,  $i \in I$ , and test functions  $w_i$ ,  $i \in I$ , by means of the formulae (4.78), (4.80), (4.81), (4.82) whence, evaluating  $\langle \ell, w_i \rangle$ ,  $i \in I$ , everything in the representation of  $P_{\mathcal{T}}\ell$  in Corollary 4.61 is at our disposal.

**Example 4.63 (global instability of  $P_{\mathcal{T}}$  and  $P_{\mathcal{T}}^*$ )** While the projections  $P_{\mathcal{T}}$  and  $P_{\mathcal{T}}^*$  are locally stable, both may become globally unbounded under mesh refinement. To see this, recall (4.73), note  $\|P_{\mathcal{T}}\|_{\mathcal{L}(H^{-1}(\Omega))} = \|P_{\mathcal{T}}^*\|_{\mathcal{L}(H_0^1(\Omega))}$  and, following the spirit of an example in [Tantardini *et al.* n.d.], consider

$$w := \sum_{z \in \mathcal{V} \cap \Omega} \phi_z \in H_0^1(\Omega).$$

Then, for all *quasi-uniform* meshes  $\mathcal{T}$  with shape regularity coefficient  $\sigma$ , there is a constant  $C$  depending on  $\sigma$  and quasi-uniformity such that

$$\|P_{\mathcal{T}}^*\|_{\mathcal{L}(H_0^1(\Omega))}^2 \geq \frac{\|\nabla(P_{\mathcal{T}}^*w)\|_{L^2(\Omega)}^2}{\|\nabla w\|_{L^2(\Omega)}^2} \geq C \frac{\#\{T \in \mathcal{T} \mid T \cap \partial\Omega = \emptyset\}}{\#\{T \in \mathcal{T} \mid T \cap \partial\Omega \neq \emptyset\}}. \quad (4.84)$$

Obviously, the last term tends to  $\infty$  under uniform refinement.

To prove (4.84), we proceed in several steps, mostly hiding constants depending on quasi-uniformity of  $\mathcal{T}$  and, as usual, the shape regularity coefficient  $\sigma$  and  $d$ .

[1] We first bound  $\|\nabla w\|_{L^2(\Omega)}$  from above. Noting that

$$w = 1 \quad \text{on} \quad \bigcup_{T \cap \partial\Omega = \emptyset} T,$$

the bound  $\|\nabla \phi_z\|_{L^\infty(T)} \lesssim h_T^{-1}$  readily implies

$$\|\nabla w\|_{L^2(\Omega)}^2 = \sum_{T \cap \partial\Omega \neq \emptyset} \|\nabla w\|_{L^2(T)}^2 \lesssim \#\{T \in \mathcal{T} \mid T \cap \partial\Omega \neq \emptyset\} h_{\mathcal{T}}^{d-2}, \quad (4.85)$$

where  $h_{\mathcal{T}}$  stands for the meshsize of  $\mathcal{T}$ .

[2] The lower bound for  $\|\nabla(P_{\mathcal{T}}^*w)\|_{L^2(\Omega)}$  is more involved. We start by showing the following representation for any  $T \in \mathcal{T}$  with  $T \cap \partial\Omega = \emptyset$ :

$$P_{\mathcal{T}}^*w|_T = H_T^* \widehat{v} \quad (4.86)$$

with the fixed function

$$\widehat{v} := \sum_{(\widehat{T}, k) \in \widehat{I}} \left( \int_{\widehat{T}} q_{(\widehat{T}, k)} \right) \widehat{w}_{(\widehat{T}, k)} + \sum_{(F', k)} \left( \frac{|F'|}{|\widehat{F}|} \right)^{\frac{1}{2}} \left( \int_{\widehat{F}} q_{(\widehat{F}, k)} \right) \widehat{w}_{(F', k)} \notin \mathbb{P}_0,$$

where the indices of the second sum vary according to  $F' \subset \widehat{T}$ ,  $k = 1, \dots, K_1$  and

$\widehat{w}_{(F',k)}$  is given by (4.82) with the transformation  $H_{(\widehat{T},F')}$ . Note first that, thanks to  $w = 1$  on  $T$  and (4.80), the coefficients in the expansion of  $P_{\mathcal{T}}^* w|_T$  satisfy

$$\langle \ell_{(T,k)}, w \rangle = \int_T q_{(T,k)} = \left( \frac{|T|}{|\widehat{T}|} \right)^{1/2} \int_{\widehat{T}} q_{(\widehat{T},k)}$$

and, for any  $F \subset T$ ,

$$\langle \ell_{(F,k)}, w \rangle = \int_F q_{(F,k)} = \left( \frac{|F|}{|\widehat{F}|} \right)^{1/2} \int_{\widehat{F}} q_{(\widehat{F},k)}.$$

Combining these identities with (4.82) yields the claimed identity (4.86) and it remains to verify  $\widehat{v} \notin \mathbb{P}_0$ . Suppose  $\widehat{v} = c \in \mathbb{R}$ . As a consequence, for any face  $F' \subset \widehat{T}$  and  $k \in \{1, \dots, K_1\}$ , we have

$$c = \widehat{w}_{(F',k)} = |\widehat{F}|^{1/2} |F'|^{-1/2} \widehat{w}_{(\widehat{F},k)} = |\widehat{F}|^{1/2} |F'|^{-1/2} c.$$

As not all faces of the reference simplex have the same volume, this yields  $c = 0$ . From (4.78) and (4.76), we infer that the coefficients in the definition of  $\widehat{v}$  vanish. In particular,  $\int_{\widehat{T}} q_{(\widehat{T},k)} = 0$  for all  $k = 1, \dots, K_2$  means  $\widehat{Q}1 = 0$ , where  $\widehat{Q}$  is the operator given in (4.77). This however is a contradiction because the restriction of  $\widehat{Q}$  to  $\mathbb{P}_{m_2}$  is injective. Hence,  $\widehat{v} \notin \mathbb{P}_0$  is proven.

[3] We are ready to show the bound for  $\|\nabla(P_{\mathcal{T}}^* w)\|_{L^2(\Omega)}$ . Given any element  $T \in \mathcal{T}$  with  $T \cap \partial\Omega = \emptyset$ , we pass to the reference element to exploit the previous step and obtain

$$\|\nabla(P_{\mathcal{T}}^* w)\|_{L^2(T)} = \|\nabla(H_T^* \widehat{v})\|_{L^2(T)} \gtrsim h_T^{d/2-1} \|\nabla \widehat{v}\|_{L^2(\widehat{T})}$$

with  $\|\nabla \widehat{v}\|_{L^2(\widehat{T})} > 0$  independent of  $T$ . Consequently,

$$\|\nabla(P_{\mathcal{T}}^* w)\|_{L^2(\Omega)}^2 \geq \sum_{T \cap \partial\Omega = \emptyset} \|\nabla(P_{\mathcal{T}}^* w)\|_{L^2(T)}^2 \gtrsim \#\{T \in \mathcal{T} \mid T \cap \partial\Omega = \emptyset\} h_{\mathcal{T}}^{d-2}$$

because  $\mathcal{T}$  is quasi-uniform. Combining this lower bound with the upper bound (4.85) of the first step, we conclude (4.84).

### A hierarchical estimator

Like estimators based upon local problems, hierarchical estimators aim at softening the impact of constants in the lower bound, with the difference that they are explicit. While global higher order extensions were used originally, [Bornemann, Erdmann and Kornhuber 1996] use an extension tailored to the residual structure and derive an upper bound with indicators testing the residual with a basis of the extension. One may expect that such explicit indicators come at the price of increased constants in the upper bound. For the following example, this expectation is confirmed by the inequality  $C_{\text{istb}}^* \leq C_{\text{bos}}^*$ .

Given the Galerkin approximation  $u_{\mathcal{T}}$  from (4.1), the hierarchical PDE indicator is defined by

$$\eta_{\mathcal{T}}^{\text{hier}}(u_{\mathcal{T}}) := \sum_{z \in \mathcal{V}} \eta_{\mathcal{T}}^{\text{hier}}(u_{\mathcal{T}}, z)^2 \quad \text{with} \quad \eta_{\mathcal{T}}^{\text{hier}}(u_{\mathcal{T}}, z) := \max_{i \in I_z} \frac{|\langle R_{\mathcal{T}}, w_i \rangle|}{\|\nabla w_i\|_{L^2(\omega_z)}} \quad (4.87a)$$

with  $I$  and  $I_z$  as in Lemma 4.60 (biorthogonal system). Note that the test functions  $w_i$ ,  $i \in I$ , are available, cf. Remark 4.62, and therefore  $\eta_{\mathcal{T}}^{\text{hier}}(u_{\mathcal{T}})$  is explicit. The resulting estimator is then

$$\mathcal{E}_{\mathcal{T}}^{\text{hier}} := \mathcal{E}_{\mathcal{T}}^{\text{hier}}(u_{\mathcal{T}}, f)^2 := \eta_{\mathcal{T}}^{\text{hier}}(u_{\mathcal{T}})^2 + \text{osc}_{\mathcal{T}}(f)^2. \quad (4.87b)$$

**Theorem 4.64 (hierarchical estimator)** *Suppose the coefficients  $A$  and  $c$  are discrete. The hierarchical estimator (4.87) is equivalent to the error, while its PDE indicator is locally equivalent to the discretized residual with constant 1 in the lower bound:*

$$\frac{1}{C_{\text{lstb}}^{\star} C_d} \mathcal{E}_{\mathcal{T}}^{\text{hier}} \lesssim \|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)} \lesssim C_{\text{bos}}^{\star} C_d C_{\text{loc}} \mathcal{E}_{\mathcal{T}}^{\text{hier}}$$

and, for all vertices  $z \in \mathcal{V}$ ,

$$\eta_{\mathcal{T}}^{\text{hier}}(u_{\mathcal{T}}, z) \leq \|P_{\mathcal{T}} R_{\mathcal{T}}\|_{H^{-1}(\omega_z)} \leq C_{\text{bos}}^{\star} \eta_{\mathcal{T}}^{\text{hier}}(u_{\mathcal{T}}, z),$$

where  $C_{\text{lstb}}^{\star}$  is the stability constant of  $P_{\mathcal{T}}$  on stars from Lemma 4.5,  $C_{\text{loc}}$  from Corollary 4.6,  $C_d = \sqrt{2(d+1)}$ , and the hidden constants depend only on the error-residual relationship in Lemma 4.1.

*Proof.* It suffices to verify the local equivalence for the PDE indicator; cf. Theorem 4.58 (vertex-indexed modified residual estimation). Its lower bound simply follows from (4.74): for all  $i \in I_z$ , we have

$$|\langle R_{\mathcal{T}}, w_i \rangle| = |\langle P_{\mathcal{T}} R_{\mathcal{T}}, w_i \rangle| \leq \|P_{\mathcal{T}} R_{\mathcal{T}}\|_{H^{-1}(\omega_z)} \|\nabla w_i\|_{L^2(\omega_z)}.$$

To show its upper bound, let  $w \in H_0^1(\omega_z)$  and, with the help of Corollary 4.61 (projections as interpolation operators) and Lemma 4.60 (biorthogonal system), we derive

$$\begin{aligned} \langle P_{\mathcal{T}} R_{\mathcal{T}}, w \rangle &= \sum_{i \in I_z} \langle R_{\mathcal{T}}, w_i \rangle \langle \ell_i, w \rangle \\ &\leq \sum_{i \in I_z} \frac{\langle R_{\mathcal{T}}, w_i \rangle}{\|\nabla w_i\|_{L^2(\omega_z)}} \|\nabla w_i\|_{L^2(\omega_z)} \|\ell_i\|_{H^{-1}(\omega_z)} \|\nabla w\|_{L^2(\omega_z)} \\ &\leq C_{\text{bos}}^{\star} \eta_{\mathcal{T}}^{\text{hier}}(u_{\mathcal{T}}, z) \|\nabla w\|_{L^2(\omega_z)} \end{aligned}$$

and the local equivalence is established.  $\square$

**Remark 4.65 (different test functions)** The hierarchical estimator (4.87) does not generalize the one in [Bornemann et al. 1996] as it uses slightly different test functions for edges. The given framework however applies to their variant, too; cf. [Kreuzer and Veese 2021, Sect. 4.1].

*Alternative localization and residual splitting*

Lemma 4.5 (localization of  $H^{-1}$ -norm) is not well suited for reducing or avoiding constants in the upper bounds. The following modification however allows this.

We replace the local spaces  $H_0^1(\omega_z)$ ,  $z \in \mathcal{V}$ , with

$$\mathbb{W}_z := \begin{cases} \{w \in H^1(\omega_z) \mid \int_{\omega_z} w = 0\}, & \text{if } z \in \mathcal{V} \cap \Omega, \\ \{w \in H^1(\omega_z) \mid w = 0 \text{ on } \partial\omega_z \cap \partial\Omega\}, & \text{if } z \in \mathcal{V} \cap \partial\Omega, \end{cases}$$

endow them with the norm  $\|\nabla \cdot\|_{L^2(\omega_z)}$ , and denote by  $\mathbb{W}_z^*$  the respective dual spaces endowed in turn with

$$\|\ell\|_{\mathbb{W}_z^*} := \sup \{ \langle \ell, w \rangle \mid w \in \mathbb{W}_z, \|\nabla w\|_{L^2(\omega_z)} \leq 1 \}. \quad (4.88)$$

**Lemma 4.66 (alternative localization of  $H^{-1}$ -norm)** *Let  $\ell \in H^{-1}(\Omega)$  be any linear functional.*

(i) *If  $\langle \ell, \phi_z \rangle = 0$  for all interior vertices  $z \in \mathcal{V} \cap \Omega$ , then*

$$\|\ell\|_{H^{-1}(\Omega)}^2 \leq (d+1) \sum_{z \in \mathcal{V}} \|\phi_z \ell\|_{\mathbb{W}_z^*}^2.$$

(ii) *We have*

$$\sum_{z \in \mathcal{V}} \|\phi_z \ell\|_{\mathbb{W}_z^*}^2 \leq (d+1) C_{\text{loc}}^2 \|\ell\|_{H^{-1}(\Omega)}^2,$$

where  $C_{\text{loc}}$  is the constant in Lemma 4.5 (i).

*Proof.* The proof is essentially a regrouping of the arguments in Lemma 4.5, where the constant  $C_{\text{loc}}$  in the stability bound (4.6) now arises in the proof of the lower bound from the following argument: we have

$$\|\phi_z \ell\|_{\mathbb{W}_z^*} \leq C_{\text{loc}} \|\ell\|_{H^{-1}(\omega_z)}. \quad (4.89)$$

thanks to

$$\begin{aligned} \langle \phi_z \ell, w \rangle &= \langle \ell, \phi_z w \rangle \leq \|\ell\|_{H^{-1}(\omega_z)} \|\nabla(w \phi_z)\|_{L^2(\omega_z)} \\ &\leq C_{\text{loc}} \|\ell\|_{H^{-1}(\omega_z)} \|\nabla w\|_{L^2(\omega_z)}, \end{aligned}$$

for all  $w \in \mathbb{W}_z$ . □

The question arises whether the inequality (4.89) between the two local dual norms can be reversed. The following lemma reveals that this is only partially possible, covering discrete functionals as arguments.

**Lemma 4.67 (partial equivalence for local dual norms)** *If  $z \in \mathcal{V} \cap \Omega$  is an interior vertex, the functional  $\ell = \phi_z^{-1}$  satisfies*

$$\|\phi_z \ell\|_{\mathbb{W}_z^*} = 0 \quad \text{and} \quad \|\ell\|_{H^{-1}(\omega_z)} > 0.$$

Furthermore, for any vertex  $z \in \mathcal{V}$ ,

$$\|\ell\|_{H^{-1}(\omega_z)} \leq C_{\mathbb{F}} \|\phi_z \ell\|_{\mathbb{W}_z^*} \quad \forall \ell \in \mathbb{F}(\mathcal{T}_z),$$

where the constant  $C_{\mathbb{R}}$  depends only on  $d$ , the shape regularity coefficient  $\sigma$ , and the degrees  $m_1$  and  $m_2$  of the discrete functionals.

*Proof.* [1] We show the claims on the functional  $\ell = \phi_z^{-1}$  for an interior vertex  $z \in \mathcal{V} \cap \Omega$ . By the definition of  $\mathbb{W}_z$ , we have, for all  $w \in \mathbb{W}_z$ ,

$$\langle \phi_z \ell, w \rangle = \int_{\omega_z} w = 0,$$

whence  $\phi_z \ell \in \mathbb{W}_z^*$  with  $\|\phi_z \ell\|_{\mathbb{W}_z^*} = 0$ .

To verify that  $\ell = \phi_z^{-1} \in H^{-1}(\omega_z)$ , we write  $d_z := \text{dist}(\cdot, \partial\omega_z)$  for the distance function of the star boundary and shall use the weighted Poincaré inequality

$$\forall w \in H_0^1(\omega_z) \quad \|wd_z^{-1}\|_{L^2(\omega_z)} \lesssim \|\nabla w\|_{L^2(\omega_z)},$$

which follows from the Hardy inequality; cf. [Sacchi and Veeser 2006, Lemma 3.6]. Consequently, exploiting also  $d_z \leq \phi_z$  on  $\omega_z$ , we obtain, for all  $w \in H_0^1(\Omega)$ ,

$$\langle \ell, w \rangle = \int_{\omega_z} (\phi_z^{-1} d_z)(wd_z^{-1}) \leq |\omega_z|^{1/2} \|wd_z^{-1}\|_{L^2(\omega_z)} \lesssim |\omega_z|^{1/2} \|\nabla w\|_{L^2(\omega_z)}.$$

This and  $\langle \ell, \phi_z \rangle = |\omega_z|$  ensure  $\ell \in H^{-1}(\omega_z)$  with  $\|\ell\|_{H^{-1}(\omega_z)} > 0$ .

[2] We start the proof of the asserted inequality by checking that  $\|\cdot\|_{H^{-1}(\omega_z)}$  is a norm on  $\mathbb{F}(\mathcal{T}_z)$ . To this end, consider  $q_F \in \mathbb{P}_{m_1}(F)$ ,  $F \in \mathcal{F}_z$  and  $q_T \in \mathbb{P}_{m_2}(T)$ ,  $T \in \mathcal{T}_z$  such that, for all  $w \in H_0^1(\omega_z)$ ,

$$0 = \langle \ell, w \rangle := \sum_{F \in \mathcal{F}_z} \int_F q_F w + \sum_{T \in \mathcal{T}_z} \int_T q_T w.$$

We need to show  $\ell = 0$ . Testing with  $w \in H_0^1(T)$ ,  $T \in \mathcal{T}_z$ , the fundamental lemma of the calculus of variations yields  $q_T = 0$  for all  $T \in \mathcal{T}_z$ . Similarly, testing now with  $w \in H_0^1(\omega_F)$ ,  $F \in \mathcal{F}_z$ , gives  $q_F = 0$  for all  $F \in \mathcal{F}_z$ . Thus,  $\ell = 0$  holds.

[3] Next, we check that also  $\|\phi_z \cdot\|_{\mathbb{W}_z^*}$  is a norm on  $\mathbb{F}(\mathcal{T}_z)$ . This time, consider  $q_F \in \mathbb{P}_{m_1}(F)$ ,  $F \in \mathcal{F}_z$  and  $q_T \in \mathbb{P}_{m_2}(T)$ ,  $T \in \mathcal{T}_z$  such that, for all  $w \in \mathbb{W}_z$ ,

$$0 = \langle \phi_z \ell, w \rangle := \sum_{F \in \mathcal{F}_z} \int_F \phi_z q_F w + \sum_{T \in \mathcal{T}_z} \int_T \phi_z q_T w,$$

and again, we need to conclude  $\ell = 0$ . If  $z \in \mathcal{V} \cap \partial\Omega$  is a boundary node, we obtain  $\ell = 0$  by the arguments of the previous step. We are thus left with the case  $z \in \mathcal{V} \cap \Omega$  of interior nodes. Given  $w \in H^1(\omega_z)$ , we set  $c_w := \oint_{\omega_z} w$  and  $c_\ell = |\omega_z|^{-1} \langle \phi_z \ell, 1 \rangle$ , and observe

$$0 = \langle \phi_z \ell, w - c_w \rangle = \langle \phi_z \ell - c_\ell, w - c_w \rangle = \langle \phi_z \ell - c_\ell, w \rangle.$$

Hence, testing with  $w \in H_0^1(T)$ ,  $T \in \mathcal{T}_z$ , we deduce  $\phi_z q_T = c_\ell$  on each  $T \in \mathcal{T}_z$ . This is however only possible if  $c_\ell = 0$  and  $q_T = 0$  for all  $T \in \mathcal{T}_z$ . Therefore, testing



with  $w \in H_0^1(\omega_F)$ ,  $F \in \mathcal{F}_z$ , yields  $q_F = 0$  for all  $F \in \mathcal{F}_z$  and  $\ell = 0$  is established in general.

[4] To conclude the asserted inequality, note that  $\mathbb{F}(\mathcal{T}_z)$  has finite dimension and, for a fixed polynomial degrees  $m_1$  and  $m_2$ , is invariant under continuous piecewise affine transformations. Furthermore, both norms scale in the same manner. We therefore can pass to reference stars and use there the equivalence of norms in finite-dimensional spaces. Transforming the inequality back from the reference star then finishes the proof.  $\square$

The alternative localization entails that we need to adapt Lemma 4.35 (splitting of local residual norms). Relying on the local  $H^{-1}$ -stability of  $P_{\mathcal{T}}$ , Lemma 4.67 (partial equivalence for local dual norms) reveals that the adaptation has to be global.

**Lemma 4.68 (alternative splitting)** *Using the local norms  $\|\cdot\|_{\mathbb{W}_z^*}$ ,  $z \in \mathcal{V}$ , the residual can be split into discretized and oscillatory residual:*

$$\begin{aligned} & \frac{1}{(C_{\text{IStb}}^*)^2 C_d^2 C_{\text{loc}}^2} \sum_{z \in \mathcal{V}} \left( \|\phi_z P_{\mathcal{T}} R_{\mathcal{T}}\|_{\mathbb{W}_z^*}^2 + \|\phi_z (I - P_{\mathcal{T}}) R_{\mathcal{T}}\|_{\mathbb{W}_z^*}^2 \right) \\ & \leq \|R_{\mathcal{T}}\|_{H^{-1}(\Omega)}^2 \leq C_d^2 \sum_{z \in \mathcal{V}} \left( \|\phi_z P_{\mathcal{T}} R_{\mathcal{T}}\|_{\mathbb{W}_z^*}^2 + \|\phi_z (I - P_{\mathcal{T}}) R_{\mathcal{T}}\|_{\mathbb{W}_z^*}^2 \right), \end{aligned}$$

where  $C_{\text{IStb}}^*$  is the stability constant of  $P_{\mathcal{T}}$  on stars from Lemma 4.28, and  $C_d = \sqrt{2(d+1)}$ .

*Proof.* Combine the localization in Lemma 4.66 with the proof of Lemma 4.35, replacing the local norm  $\|\cdot\|_{H^{-1}(\omega_z)}$  in most places but apply (4.89) before using the local  $H^{-1}$ -stability of  $P_{\mathcal{T}}$ .  $\square$

#### An estimator based on flux equilibration

Estimators based on flux equilibration have been designed with the goal to obtain constant 1 in the upper bound. The principal obstruction that computation can access only a finite-dimensional part of infinite-dimensional objects like the residual norm is overcome by means of the *Prager-Synge theorem*. Realizations of this approach can be found, e.g., in [Ainsworth 2010], [Braess, Pillwein and Schöberl 2009], [Ern, Smears and Vohralik 2017], and [Luce and Wohlmuth 2004].

The definition of the PDE indicator needs some preparation. Let  $d \in \{2, 3\}$ , as in the aforementioned works, and let  $z \in \mathcal{V}$  be vertex. Given the operator  $\pi_z : \{\phi_z \ell \mid \ell \in H^{-1}(\Omega)\} \rightarrow \mathbb{W}_z^*$  defined by

$$\pi_z(\phi_z \ell) := \begin{cases} \phi_z \ell - \frac{\langle \ell, \phi_z \rangle}{|\omega_z|}, & \text{if } z \in \mathcal{V} \cap \Omega, \\ \phi_z \ell, & \text{if } z \in \mathcal{V} \cap \partial\Omega. \end{cases}$$

and

$$\gamma_z := \begin{cases} \partial\omega_z, & \text{if } z \in \mathcal{V} \cap \Omega, \\ \partial\omega_z \setminus \partial\Omega & \text{if } z \in \mathcal{V} \cap \partial\Omega, \end{cases}$$

we introduce the local space  $\mathbb{D}_z \neq \emptyset$

$$\mathbb{D}_z := \{\boldsymbol{\xi} \in L^2(\omega_z; \mathbb{R}^d) \mid \operatorname{div} \boldsymbol{\xi} = \pi_z(\phi_z P_{\mathcal{T}} R_{\mathcal{T}}) \text{ and } \boldsymbol{\xi} \cdot \mathbf{n}_F = 0 \text{ on } F, \forall F \subseteq \gamma_z\},$$

and its discretization

$$\mathbb{D}_z(\mathcal{T}) := \{\boldsymbol{\xi} \in \mathbb{D}_z \mid \boldsymbol{\xi} \in \operatorname{RTN}_m(T) \quad \forall T \in \mathcal{T}_z\}$$

with the Raviart-Thomas-Nédélec elements

$$\operatorname{RTN}_m(T) = \{\boldsymbol{\xi} : T \rightarrow \mathbb{R}^d \mid \boldsymbol{\xi}(x) = \mathbf{q}(x) + q(x)x \text{ with } \mathbf{q} \in (\mathbb{P}_m)^d, q \in \mathbb{P}_m\}$$

of order  $m := \max\{m_1, m_2\} + 1$ . Given the Galerkin approximation  $u_{\mathcal{T}}$  from (4.1), the PDE indicator is then given by

$$\eta_{\mathcal{T}}^{\text{feq}}(u_{\mathcal{T}})^2 := \sum_{z \in \mathcal{V}} \eta_{\mathcal{T}}^{\text{feq}}(u_{\mathcal{T}})^2 \quad \text{with} \quad \eta_{\mathcal{T}}^{\text{feq}}(u_{\mathcal{T}}) := \min_{\boldsymbol{\xi} \in \mathbb{D}_z(\mathcal{T})} \|\boldsymbol{\xi}\|_{L^2(\omega_z)} \quad (4.90a)$$

and the total estimator by

$$\mathcal{E}_{\mathcal{T}}^{\text{feq}} := \mathcal{E}_{\mathcal{T}}^{\text{feq}}(u_{\mathcal{T}}, f)^2 := \eta_{\mathcal{T}}^{\text{feq}}(u_{\mathcal{T}})^2 + \|\phi_z(f - P_{\mathcal{T}}f)\|_{\mathbb{W}_z^*}^2. \quad (4.90b)$$

Note that the local PDE indicators  $\eta_{\mathcal{T}}^{\text{feq}}(u_{\mathcal{T}}, z)$  are computable up to machine precision.

**Theorem 4.69 (estimator based on flux equilibration)** *Suppose that the coefficients  $\mathbf{A}$  and  $c$  are discrete and that  $d \in \{2, 3\}$ . The estimator (4.90) based on flux equilibration is equivalent to the error, while its PDE indicator is locally equivalent to the discretized residual with constant 1 in the upper bound for the  $\|\cdot\|_{\mathbb{W}_z^*}$ -norm, so that*

$$\frac{C_{\mathbb{D}}}{C_{\text{ISib}}^* C_d C_{\text{loc}}} \mathcal{E}_{\mathcal{T}}^{\text{feq}} \lesssim \|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)} \lesssim C_d \mathcal{E}_{\mathcal{T}}^{\text{feq}}$$

and, for all vertices  $z \in \mathcal{V}$ ,

$$C_{\mathbb{D}} \eta_{\mathcal{T}}^{\text{feq}}(u_{\mathcal{T}}, z) \leq \|\phi_z P_{\mathcal{T}} R_{\mathcal{T}}\|_{\mathbb{W}_z^*} \leq \eta_{\mathcal{T}}^{\text{feq}}(u_{\mathcal{T}}, z)$$

as well as

$$\frac{C_{\mathbb{D}}}{C_{\text{loc}}} \eta_{\mathcal{T}}^{\text{feq}}(u_{\mathcal{T}}, z) \leq \|P_{\mathcal{T}} R_{\mathcal{T}}\|_{H^{-1}(\omega_z)} \leq C_{\mathbb{F}} \eta_{\mathcal{T}}^{\text{feq}}(u_{\mathcal{T}}, z),$$

where  $C_{\mathbb{D}}$  depends on  $d$  and the shape regularity coefficient  $\sigma$ ,  $C_{\text{ISib}}^*$  is the stability constant of  $P_{\mathcal{T}}$  on stars from Lemma 4.28,  $C_{\text{loc}}$  comes from Lemma 4.5,  $C_d = \sqrt{2(d+1)}$ ,  $C_{\mathbb{F}}$  from Lemma 4.67. and the hidden constants depend only on the error-residual relationship in Lemma 4.1.

*Proof.* [1] We start by verifying the local equivalence for the  $\|\cdot\|_{\mathbb{W}_z^*}$ -norm. Let  $z \in \mathcal{V}$  be any vertex. The Prager-Synge theorem on the star  $\omega_z$  implies

$$\|\phi_z P_{\mathcal{T}} R_{\mathcal{T}}\|_{\mathbb{W}_z^*} = \|\pi_z(\phi_z P_{\mathcal{T}} R_{\mathcal{T}})\|_{\mathbb{W}_z^*} = \min_{\xi \in \mathbb{D}_z} \|\xi\|_{L^2(\omega_z)};$$

cf., e.g., [Verfürth 2013, Proposition 1.40]. Hence the upper bound with constant 1 readily follows the inclusion  $\mathbb{D}_z(\mathcal{T}) \subset \mathbb{D}_z$ , while the lower bound is a consequence of the nontrivial inequality

$$C_{\mathbb{D}} \min_{\xi \in \mathbb{D}_z(\mathcal{T})} \|\xi\|_{L^2(\omega_z)} \leq \min_{\xi \in \mathbb{D}_z} \|\xi\|_{L^2(\omega_z)},$$

where  $C_{\mathbb{D}}$  depends only on  $d$  and the shape regularity coefficient  $\sigma$ ; cf., e.g., [Braess et al. 2009, Theorem 7] and [Ern et al. 2017, Theorem 1.1].

[2] We verify the local equivalence for the  $\|\cdot\|_{H^{-1}(\omega_z)}$ -norm. On the one hand, combining the first equivalence with (4.89), we obtain

$$C_{\mathbb{D}} \eta_{\mathcal{T}}^{\text{feq}}(u_{\mathcal{T}}, z) \leq \|\phi_z P_{\mathcal{T}} R_{\mathcal{T}}\|_{\mathbb{W}_z^*} \leq C_{\text{loc}} \|P_{\mathcal{T}} R_{\mathcal{T}}\|_{H^{-1}(\omega_z)}.$$

On the other hand, using Lemma 4.67 instead of (4.89) yields

$$\|P_{\mathcal{T}} R_{\mathcal{T}}\|_{H^{-1}(\omega_z)} \leq C_{\mathbb{F}} \|\phi_z P_{\mathcal{T}} R_{\mathcal{T}}\|_{\mathbb{W}_z^*} \leq C_{\mathbb{F}} \eta_{\mathcal{T}}^{\text{feq}}(u_{\mathcal{T}}, z)$$

and the equivalence for the  $\|\cdot\|_{H^{-1}(\omega_z)}$  is verified, too.

[3] The global bounds follow by combining Lemmas 4.1 (error and residual), 4.68 (alternative splitting in discretized and oscillatory residual), and 4.37 (data oscillation reduction), as well as the first local equivalence.  $\square$

**Remark 4.70 (improved upper bound)** Applying the Prager-Synge theorem on  $\Omega$ , we can improve the upper bound in Theorem 4.69 to

$$\|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)} \lesssim \|\xi_{\Omega}\|_{L^2(\Omega)} + \sqrt{d+1} \left( \sum_{z \in \mathcal{V}} \|\pi_z(\phi_z(P_{\mathcal{T}} f - f))\|_{\mathbb{W}_z^*}^2 \right)^{1/2} \quad (4.91)$$

with  $\xi_{\Omega} := \sum_{z \in \mathcal{V}} \xi_z$ , where  $\xi_z := \arg\min_{\xi \in \mathbb{D}_z(\mathcal{T})} \|\xi\|_{L^2(\omega_z)}$  are the minimizing vector fields associated with the PDE indicators, extended by 0 off  $\omega_z$ .

To see this, we derive, thanks to the partial orthogonality (4.4) of the residual and Lemma 4.37 (data oscillation reduction),

$$\begin{aligned} \operatorname{div} \xi_{\Omega} &= \sum_{z \in \mathcal{V}} \pi_z(\phi_z P_{\mathcal{T}} R_{\mathcal{T}}) = \sum_{z \in \mathcal{V}} \pi_z(\phi_z R_{\mathcal{T}}) + \sum_{z \in \mathcal{V}} \pi_z(\phi_z(P_{\mathcal{T}} R_{\mathcal{T}} - R_{\mathcal{T}})) \\ &= \sum_{z \in \mathcal{V}} \phi_z R_{\mathcal{T}} + \sum_{z \in \mathcal{V}} \pi_z(\phi_z(P_{\mathcal{T}} f - f)) = R_{\mathcal{T}} + \delta_{\mathcal{T}} \end{aligned}$$

with  $\delta_{\mathcal{T}} := \sum_{z \in \mathcal{V}} \pi_z(\phi_z(P_{\mathcal{T}} f - f))$ . Hence,

$$\|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)} \lesssim \|R_{\mathcal{T}}\|_{H^{-1}(\Omega)} \leq \|R_{\mathcal{T}} + \delta_{\mathcal{T}}\|_{H^{-1}(\Omega)} + \|\delta_{\mathcal{T}}\|_{H^{-1}(\Omega)},$$

inserting  $\xi_{\Omega}$  in the Prager-Synge theorem on  $\Omega$  and Lemma 4.66 (alternative localization of  $H^{-1}$ -norm) establish the claimed bound.

In view of the bound (4.91), the alternative local PDE indicators  $\|\phi_z^{1/2}\xi_\Omega\|_{L^2(\omega_z)}$ ,  $z \in \mathcal{V}$ , may be used in an adaptive context. Note however that this alternative does not necessarily strengthen the link with the local residual as the definition of  $\xi_\Omega$  suggests an increased overlapping in the lower bound.

#### 4.10 Other boundary conditions

This section illustrates that the preceding analysis of homogeneous Dirichlet conditions can be adapted to other boundary conditions. In particular, we discuss

- *Robin and Neumann boundary conditions*, as an example for variationally formulated boundary conditions,
- the *pure Neumann problem*, with its global solvability constraint,
- *non-homogeneous Dirichlet boundary conditions*, formulated in an essential manner.

*Mixed boundary conditions*, suitably discretized, give rise to a posteriori error estimators combining in a straightforward manner the indicators of, for instance, the first and third of the above groups. We therefore omit further details for such a setting.

##### *Robin and Neumann boundary conditions*

The *Robin* bilinear form in (2.13) is coercive and continuous in  $\mathbb{V} := H^1(\Omega)$  provided its coefficient  $p \geq p_0$  on an open subset of  $\partial\Omega$  for some constant  $p_0 > 0$ , according to the norm equivalence (2.31). Consequently, (2.12) admits a unique solution  $u \in \mathbb{V}$ . If  $\mathbb{V}_\mathcal{T} = \mathbb{S}_\mathcal{T}^{n,0}$  is the subspace of  $\mathbb{V}$  of continuous piecewise polynomial functions of degree  $\leq n$ , then the Galerkin counterpart of (4.1) reads

$$u_\mathcal{T} \in \mathbb{V}_\mathcal{T} : \quad \mathcal{B}[u_\mathcal{T}, v] = \ell(v) \quad \forall v \in \mathbb{V}_\mathcal{T},$$

with  $\ell = f + g\delta_{\partial\Omega} \in \mathbb{V}^*$ ; cf. (2.13). Its residual  $R_\mathcal{T} \in \mathbb{V}^*$  is defined as

$$\langle R_\mathcal{T}, w \rangle := \ell(w) - \mathcal{B}[u_\mathcal{T}, w] \quad w \in \mathbb{V},$$

and  $\|R_\mathcal{T}\|_{\mathbb{V}^*}$  is equivalent to the error  $\|u - u_\mathcal{T}\|_{H^1(\Omega)}$  due to Lemma 4.1 (error and residual), whose proof easily extends to  $\mathbb{V}$ .

The global norm  $\|R_\mathcal{T}\|_{\mathbb{V}^*}$  also localizes to all stars  $\omega_z$  because Galerkin orthogonality  $\langle R_\mathcal{T}, \phi_z \rangle = 0$  is now valid also for boundary vertices  $z \in \mathcal{V} \cap \partial\Omega$ . Indeed, the proof of Lemma 4.5 (localization of  $H^{-1}$ -norm) extends with minor modifications, where the local spaces for boundary vertices  $z \in \mathcal{V} \cap \partial\Omega$  are now  $\{v \in H^1(\omega_z) \mid v = 0 \text{ on } \partial\omega_z \setminus \partial\Omega\}$ . Also the proof of Lemma 4.66 (alternative localization of  $H^{-1}$ -norm) is easily modified, using the local space  $\{v \in H^1(\omega_z) \mid \int_{\omega_z} v = 0\}$  at the boundary, too.

The next key step is the construction of a projection  $P_\mathcal{T} : \mathbb{V}^* \rightarrow \mathbb{F}_\mathcal{T}$  that mimics the projection operator  $P_\mathcal{T}$  of Section 4.4. For that purpose, the space of discrete functionals  $\mathbb{F}_\mathcal{T}$  has to include boundary face Dirac masses  $q_F \delta_F$  with densities

$q_F \in \mathbb{P}_{m_1}(F)$  for  $F \subset \partial\Omega$ . Consequently,  $g$  can be approximated on  $\partial\Omega$  similarly to the forcing  $f$  in  $\Omega$ , while the coefficient  $p$  is at play like the coefficient  $c$ . Indeed, considering for simplicity only the case of discrete coefficients  $(A, c, p)$ , the condition  $m_1 \geq n_p + n$  arises in addition to those in Remark 4.14. With these caveats, the tools developed in Sections 4.3, 4.4 and 4.5 give rise to a suitably adapted projection  $P_{\mathcal{T}}$  to split the residual  $R_{\mathcal{T}}$  into a discretized residual  $P_{\mathcal{T}}R_{\mathcal{T}}$  and an oscillatory residual  $(f - P_{\mathcal{T}}f) + (g\delta_{\partial\Omega} - P_{\mathcal{T}}(g\delta_{\partial\Omega}))$ . Here  $g\delta_{\partial\Omega} - P_{\mathcal{T}}(g\delta_{\partial\Omega})$  is supported only on  $\partial\Omega$  and therefore contributes only to the oscillation indicators based upon the aforementioned new local spaces for boundary stars. This modified oscillation  $\text{osc}_{\mathcal{T}}^{\text{Rob}}(\mathcal{D})$  with  $\mathcal{D} = (A, c, p, f, g)$  and be combined with any of the presented PDE indicators, but we focus on residual estimation. In fact, the new discrete residual  $P_{\mathcal{T}}R_{\mathcal{T}}$  leads to a definition of the PDE estimator  $\eta_{\mathcal{T}}^{\text{Rob}}(u_{\mathcal{T}})$  as in (4.52), but with additional contributions related to the boundary faces. Given any boundary face  $F$  of  $\mathcal{T}$ , such a contribution reads

$$h_F \| [A \nabla_{\mathcal{T}} u_{\mathcal{T}}] \cdot \mathbf{n}_F + p u_{\mathcal{T}} - P_F g \|_{L^2(F)}^2$$

and measures the discretized Robin residual. Combining as usual the PDE estimator  $\eta_{\mathcal{T}}^{\text{Rob}}(u_{\mathcal{T}})$  and oscillation  $\text{osc}_{\mathcal{T}}^{\text{Rob}}(\ell)$  yields the total estimator  $\mathcal{E}_{\mathcal{T}}^{\text{Rob}}(u_{\mathcal{T}}, \ell)$ , whence the following variant of Theorem 4.45 (modified residual estimator) follows: *for discrete coefficients  $(A, c, p)$ , the  $H^1$ -error and  $\mathcal{E}_{\mathcal{T}}^{\text{Rob}}(u_{\mathcal{T}}, \ell)$  are equivalent*

$$C_L \mathcal{E}_{\mathcal{T}}^{\text{Rob}}(u_{\mathcal{T}}, \ell) \leq \|u - u_{\mathcal{T}}\|_{H^1(\Omega)} \leq C_U \mathcal{E}_{\mathcal{T}}^{\text{Rob}}(u_{\mathcal{T}}, \ell).$$

The estimates in Section 4.8 for corrections and estimator reduction extend as well.

#### Pure Neumann problem

Neumann conditions are already covered by the previous section, except for the case of the pure Neumann problem with  $p = 0$  on  $\partial\Omega$  in (2.12) requiring, as key novelty, the solvability constraint  $\ell(1_{\Omega}) = 0$ , i.e. right-hand side applied to the constant function equal to 1 gives 0. For such problems, unique exact and discrete solutions exist provided we choose  $\mathbb{V}$  to be the subspace of  $H^1(\Omega)$  of functions with zero mean value and  $\mathbb{V}_{\mathcal{T}}$  its natural finite element counterpart of degree  $n$ .

The residual  $R_{\mathcal{T}}$  is defined on all  $H^1(\Omega)$  and satisfies  $\langle R_{\mathcal{T}}, 1_{\Omega} \rangle = 0$ . Combining this fact with Lemma 2.3 (second Poincaré inequality) and  $\inf_{c \in \mathbb{R}} \|v - c\|_{L^2(\Omega)} = \|w\|_{L^2(\Omega)}$  with  $w = v - \int_{\Omega} v$ , we derive

$$\|R_{\mathcal{T}}\|_{\mathbb{V}^*} = \sup_{w \in \mathbb{V}} \frac{\langle R_{\mathcal{T}}, w \rangle}{\|\nabla w\|_{L^2(\Omega)}} \approx \sup_{w \in \mathbb{V}} \frac{\langle R_{\mathcal{T}}, w \rangle}{\|w\|_{H^1(\Omega)}} = \sup_{v \in H^1(\Omega)} \frac{\langle R_{\mathcal{T}}, v \rangle}{\|v\|_{H^1(\Omega)}} = \|R_{\mathcal{T}}\|_{H^1(\Omega)^*}.$$

Consequently, localizing  $\|R_{\mathcal{T}}\|_{H^1(\Omega)^*}$  as in the previous section, we can derive a posteriori error estimators with suitable contributions from the boundary  $\partial\Omega$ .

However, the projection  $P_{\mathcal{T}}$  from the previous section cannot be used to generate discrete data in some auxiliary problem because  $\langle \ell, 1_{\Omega} \rangle = 0$  does not imply  $\langle P_{\mathcal{T}}\ell, 1_{\Omega} \rangle = 0$  in general. Further, a simple modification like  $P_{\mathcal{T}}\ell -$

$\langle P_{\mathcal{T}}\ell, 1_{\Omega} \rangle \langle 1_{\Omega}, 1_{\Omega} \rangle^{-1} 1_{\Omega}$  with a global correction destroys the crucial local approximation properties.

To address this issue, we modify the projection  $P_{\mathcal{T}}$  such that the new projection  $\tilde{P}_{\mathcal{T}}$  enforces locally  $\langle \tilde{P}_{\mathcal{T}}\ell, 1_{\Omega} \rangle = \langle \ell, 1_{\Omega} \rangle$  in the spirit of the construction of the Lagrange multiplier in [Fierro and Veese 2003]. To this end, recall that  $P_{\mathcal{T}}$  is now defined on  $H^1(\Omega)$  and its range, the discrete functionals  $\mathbb{F}(\mathcal{T})$ , includes also boundary face Dirac masses, and that the first localization involves the local spaces  $\mathbb{V}_z := \{v \in H^1(\omega_z) \mid v = 0 \text{ on } \partial\omega_z \setminus \partial\Omega\}$ ,  $z \in \mathcal{V}$ . Given  $\ell \in H^1(\Omega)^*$ , set

$$\tilde{P}_{\mathcal{T}}\ell := \sum_{z \in \mathcal{V}} \phi_z \tilde{P}_z \ell \quad \text{with} \quad \tilde{P}_z \ell := P_{\mathcal{T}}\ell - \frac{\langle P_{\mathcal{T}}\ell - \ell, \phi_z \rangle}{\int_{\omega_z} \phi_z} 1_{\omega_z}. \quad (4.92)$$

**Lemma 4.71 (new projection)** *The operator (4.92) is linear, local, and satisfies*

$$\langle \tilde{P}_z \ell, \phi_z \rangle = \langle \ell, \phi_z \rangle \quad \forall z \in \mathcal{V} \quad \text{and} \quad \langle \tilde{P}_{\mathcal{T}}\ell, 1 \rangle = \langle \ell, 1 \rangle.$$

Furthermore,  $\tilde{P}_z$  provides near-best approximation in  $\mathbb{F}(\mathcal{T})|_{\mathbb{V}_z}$  and

$$\|\ell - \tilde{P}_{\mathcal{T}}\ell\|_{H^1(\Omega)^*}^2 \leq C_{\text{loc}} \sum_{z \in \mathcal{V}} \|\ell - \tilde{P}_z \ell\|_{\mathbb{V}_z^*}^2.$$

*Proof.* [1] We start with the algebraic properties. By the definition of  $\tilde{P}_z$ , we have the local relationships  $\langle \ell - \tilde{P}_z \ell, \phi_z \rangle = 0$ , viz.  $\langle \phi_z(\ell - \tilde{P}_z \ell), 1 \rangle = 0$  for all vertices  $z \in \mathcal{V}$ . Summing over all vertices immediately yields the global  $\langle \ell - \tilde{P}_{\mathcal{T}}\ell, 1 \rangle = 0$ .

[2] To show that  $\tilde{P}_z$  is near-best approximating in  $\mathbb{F}(\mathcal{T})|_{\mathbb{V}_z}$ , we bound its error in terms of the one of  $P_{\mathcal{T}}$ . The triangle inequality readily gives

$$\|\ell - \tilde{P}_z \ell\|_{\mathbb{V}_z^*} \leq \|\ell - P_{\mathcal{T}}\ell\|_{\mathbb{V}_z^*} + \left\| \frac{\langle P_{\mathcal{T}}\ell - \ell, \phi_z \rangle}{\int_{\omega_z} \phi_z} 1_{\omega_z} \right\|_{\mathbb{V}_z^*},$$

while a variant of Lemma 2.2 (first Poincaré inequality) and the properties of  $\phi_z$  deliver

$$\begin{aligned} \left\| \frac{\langle P_{\mathcal{T}}\ell - \ell, \phi_z \rangle}{\int_{\omega_z} \phi_z} 1_{\omega_z} \right\|_{\mathbb{V}_z^*} &\lesssim |\omega_z|^{-1} h_z \|\langle P_{\mathcal{T}}\ell - \ell, \phi_z \rangle 1_{\omega_z}\|_{L^2(\omega_z)} \\ &\lesssim |\omega_z|^{-1/2} h_z \|\ell - P_{\mathcal{T}}\ell\|_{\mathbb{V}_z^*} \|\nabla \phi_z\|_{L^2(\omega_z)} \lesssim \|\ell - P_{\mathcal{T}}\ell\|_{\mathbb{V}_z^*}. \end{aligned}$$

Hence, the error of  $\tilde{P}_z$  is dominated by the one of  $P_{\mathcal{T}}$ ,

$$\|\ell - \tilde{P}_z \ell\|_{\mathbb{V}_z^*} \lesssim \|\ell - P_{\mathcal{T}}\ell\|_{\mathbb{V}_z^*}. \quad (4.93)$$

and the near-best approximation of  $\tilde{P}_z$  follows from Corollary 4.31 (local near-best approximation), adapted to the setting at hand.

[3] It remains to prove the claimed inequality. Given  $w \in H^1(\Omega)$ , the definition

of  $\tilde{P}_{\mathcal{T}}$  and the first step yield the following identity

$$\begin{aligned}\langle \ell - \tilde{P}_{\mathcal{T}}\ell, w \rangle &= \sum_{z \in \mathcal{V}} \langle \ell, w\phi_z \rangle - \langle \phi_z \tilde{P}_z \ell, w \rangle \\ &= \sum_{z \in \mathcal{V}} \langle \ell - \tilde{P}_z \ell, w\phi_z \rangle = \sum_{z \in \mathcal{V}} \langle \ell - \tilde{P}_z \ell, (w - c_z)\phi_z \rangle,\end{aligned}$$

with  $c_z = \int_{\omega_z} w$ . Proceeding as in Lemma 4.5 (localization of  $H^{-1}$  norm) establishes the desired inequality and concludes the proof.  $\square$

The operator  $\tilde{P}_{\mathcal{T}}$  possesses additional enhanced global properties, which are not needed here. Lemma 4.71 and (4.93) allow us to solve auxiliary pure Neumann problems with discrete data  $\tilde{P}_{\mathcal{T}}\ell$ , with the option of replacing in the local indicators the restrictions of  $P_{\mathcal{T}}$  with  $\tilde{P}_z$ .

### Non-homogeneous Dirichlet boundary conditions

Let  $\mathbb{V} := H^1(\Omega)$  and  $\mathbb{V}_{\mathcal{T}} := \mathbb{S}_{\mathcal{T}}^{n,0}$  be the subspace of  $\mathbb{V}$  of continuous piecewise polynomials of degree  $\leq n$ . Given Dirichlet boundary data  $g \in H^{1/2}(\Gamma)$ , where  $\Gamma := \partial\Omega$  for simplicity, recall that  $u \in \mathbb{V}(g) = \{v \in \mathbb{V} \mid v = g \text{ on } \Gamma\}$  satisfies (2.10). Let  $g_{\mathcal{T}} \in \mathbb{S}_{\mathcal{T}}^{n,0}$  be a continuous finite element approximation of  $g$  on  $\Gamma$  and  $\mathbb{V}_{\mathcal{T}}(g_{\mathcal{T}})$  be the subspace of  $\mathbb{V}_{\mathcal{T}}$  of discrete functions with trace  $g_{\mathcal{T}}$ . The Galerkin approximation of  $u$  satisfies

$$u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}(g_{\mathcal{T}}) : \quad \mathcal{B}[u_{\mathcal{T}}, v] = \langle f, v \rangle \quad \forall v \in \mathbb{V}_{\mathcal{T}}(0).$$

The error  $e_{\mathcal{T}} = u - u_{\mathcal{T}}$  obviously satisfies Galerkin orthogonality

$$\mathcal{B}[e_{\mathcal{T}}, v] = 0 \quad \forall v \in \mathbb{V}_{\mathcal{T}}(0),$$

but in general  $e_{\mathcal{T}} = g - g_{\mathcal{T}} \neq 0$  on  $\Gamma$ . We follow [Sacchi and Veeseer 2006] to derive a posteriori bounds of  $\|e_{\mathcal{T}}\|_{H^1(\Omega)}$  using minimal regularity  $g \in H^{1/2}(\Gamma)$ .

We start with an orthogonal decomposition of the error  $e_{\mathcal{T}}$  arising from the two equations of the problem. Let  $R_G = R_G(u_{\mathcal{T}}, f) \in H^{-1}(\Omega)$  be the Galerkin residual already introduced in Section 4.1, namely

$$\langle R_G, v \rangle = \langle f, v \rangle - \mathcal{B}[u_{\mathcal{T}}, v] \quad \forall v \in \mathbb{V}(0) = H_0^1(\Omega),$$

and define the Galerkin error  $e_G$  as its representation in  $H_0^1(\Omega)$ :

$$e_G \in H_0^1(\Omega) : \quad \mathcal{B}[e_G, v] = \langle R_G, v \rangle \quad \forall v \in \mathbb{V}(0).$$

Furthermore, let  $R_D = R_D(g) = g - g_{\mathcal{T}} \in H^{1/2}(\Gamma)$  be the *Dirichlet residual*, represented by the *Dirichlet error*  $e_D$  defined by

$$e_D \in \mathbb{V}(g - I_{\mathcal{T}}g) : \quad \mathcal{B}[e_D, v] = 0 \quad \forall v \in \mathbb{V}(0).$$

Then  $e_{\mathcal{T}} = e_G + e_D$  and the orthogonality  $\mathcal{B}[e_D, e_G] = 0$  yields

$$\|e_{\mathcal{T}}\|_{\Omega}^2 = \|e_G\|_{\Omega}^2 + \|e_D\|_{\Omega}^2,$$

while the derivation for homogeneous Dirichlet conditions readily provides

$$\|e_G\|_\Omega \approx \|\nabla e_G\|_{L^2(\Omega)} \approx \mathcal{E}_\mathcal{T}(u_\mathcal{T}, f),$$

where the Galerkin estimator  $\mathcal{E}_\mathcal{T}(u_\mathcal{T}, f)$  is defined by (4.52), or any other estimator from Sect. 4.9. It thus remains to clarify whether  $\|e_\mathcal{T}\|_\Omega$  is definite in the sense  $\|e_\mathcal{T}\|_\Omega = 0 \implies e_\mathcal{T} = 0$  and to derive suitable lower and upper bounds for  $\|e_D\|_\Omega$ .

To this end, we need to be more specific about the choice of  $g_\mathcal{T}$ . Let  $g_\mathcal{T} = I_\mathcal{T}g$  be the Scott-Zhang quasi-interpolant of  $g$ , which is defined locally using boundary values of  $g$  exclusively [Scott and Zhang 1990, Brenner and Scott 2008] and satisfies

$$v \in \mathbb{V}_\mathcal{T}|_\Gamma \implies I_\mathcal{T}v = v \quad \text{on } \Gamma, \quad (\text{invariance}) \quad (4.94a)$$

$$\|I_\mathcal{T}v\|_{L^2(\Gamma)} \lesssim \|v\|_{L^2(\Gamma)} \quad \forall v \in \mathbb{V}. \quad (\text{stability}) \quad (4.94b)$$

These two properties ensure a variant of the equivalence  $\|\cdot\|_{H^1(\Omega)} \approx \|\nabla \cdot\|_{L^2(\Omega)}$  for functions with zero trace on  $\Gamma$ .

**Lemma 4.72 (equivalence for vanishing discretized trace)** *There exists a constant  $C$  depending only on the shape regularity of  $\mathbb{T}$  and  $\Omega$  such that*

$$\|v\|_{H^1(\Omega)} \leq C \|\nabla v\|_{L^2(\Omega)} \quad \forall v \in \mathbb{V} \text{ with } I_\mathcal{T}v = 0 \text{ on } \Gamma.$$

*Proof.* Note that the core of the claimed inequality amounts to a variant of the first Poincaré inequality. In view of the norm equivalence (2.31), it suffices to prove that  $\|v\|_{L^2(\Gamma)} \lesssim \|\nabla v\|_{L^2(\Omega)}$ . Letting  $\bar{v}_\Omega := \oint_\Omega v$ , and using (4.94a) yields  $v = v - I_\mathcal{T}v = (v - \bar{v}_\Omega) - I_\mathcal{T}(v - \bar{v}_\Omega)$  on  $\Gamma$ . Consequently, (4.94b) implies

$$\|v\|_{L^2(\Gamma)} \lesssim \|v - \bar{v}_\Omega\|_{L^2(\Gamma)} \lesssim \|v - \bar{v}_\Omega\|_{H^1(\Omega)} \lesssim \|\nabla v\|_{L^2(\Omega)}, \quad (4.95)$$

because of Lemma 2.4 (traces), and Lemma 2.3 (second Poincaré inequality).  $\square$

Observing  $I_\mathcal{T}e_\mathcal{T} = I_\mathcal{T}g - I_\mathcal{T}^2g = 0$  on  $\Gamma$ , we can apply Lemma 4.72 to get

$$\|e_\mathcal{T}\|_{H^1(\Omega)} \leq C \|\nabla e_\mathcal{T}\|_{L^2(\Omega)} \leq \frac{C}{\alpha_1} \|e_\mathcal{T}\|_\Omega \leq \frac{C \max\{\alpha_2, \|c\|_{L^\infty(\Omega)}\}}{\alpha_1} \|e_\mathcal{T}\|_{H^1(\Omega)},$$

establishing in particular that  $\|e_\mathcal{T}\|_\Omega$  is definite. In the same vein, we derive

$$\|e_D\|_\Omega \approx \|e_D\|_{H^1(\Omega)} \approx \|\nabla e_D\|_{L^2(\Omega)}$$

for the Dirichlet error.

With the intent to achieve directly computable bounds for the Dirichlet error, we next establish the equivalence  $\|e_D\|_{H^1(\Omega)} \approx \|g - I_\mathcal{T}g\|_{H^{1/2}(\Gamma)}$ , where the *intrinsic*  $H^{1/2}$ -norm combines the  $L^2(\Gamma)$ -norm with the seminorm

$$|v|_{H^{1/2}(\Gamma)}^2 = \int_\Gamma \int_\Gamma \frac{|v(x) - v(y)|^2}{|x - y|^d} dx dy.$$

This equivalence follows with the help of the trace and extension theorems for  $H^{1/2}(\Gamma)$ , see, e.g., [Hackbusch 1992, Theorem 6.2.40]. In fact, on the one hand, that trace theorem immediately gives  $\|g - I_\mathcal{T}g\|_{H^{1/2}(\Gamma)} \lesssim \|e_D\|_{H^1(\Omega)}$ . On the other



hand, let  $\chi \in H^1(\Omega)$  denote the extension of  $g - I_{\mathcal{T}}g$  from [Hackbusch 1992, Theorem 6.2.40]. Then

$$\|e_D\|_{H^1(\Omega)} \lesssim \|e_D\|_{\Omega} \leq \|\chi\|_{\Omega} \lesssim \|\chi\|_{H^1(\Omega)} \lesssim \|g - I_{\mathcal{T}}g\|_{H^{1/2}(\Gamma)},$$

where the second inequality is thanks to  $\mathcal{B}[e_D, e_D - \chi] = 0$ .

We are left with the issue that the  $H^{1/2}(\Gamma)$ -seminorm is *nonlocal*. To handle this delicate matter, we invoke its localization due to [Faermann 2000, 2002]

$$|v|_{H^{1/2}(\Gamma)}^2 \leq \sum_{F \in \mathcal{F}_{\Gamma}} \left( \int_F \int_{\omega_F} \frac{|v(x) - v(y)|^2}{|x - y|^d} dx dy + \frac{C}{h_F} \|v\|_{L^2(F)}^2 \right),$$

where  $F \in \mathcal{F}_{\Gamma}$  is a generic face of  $\mathcal{T}$  lying on  $\Gamma$  and  $\omega_F$  is the patch on  $\Gamma$  associated with  $F$ . The last term seems problematic. However, applied to  $v = g - I_{\mathcal{T}}g$ , we can mimic the steps of (4.95) with local variants of (4.94), but using in the last step the second Poincaré inequality in  $H^{1/2}$ , see, e.g., [Sacchi and Veeseer 2006, Lemma 3.2]:

$$\|v\|_{L^2(F)}^2 = \|v - I_{\mathcal{T}}v\|_{L^2(F)}^2 \lesssim \|v - \bar{v}_F\|_{L^2(\omega_F)}^2 \lesssim h_F |v|_{H^{1/2}(\omega_F)}^2$$

where  $\bar{v}_F$  is the mean value of  $v$  on  $\omega_F$ . Note that this bounds also means that the  $L^2$ -part in  $\|g - I_{\mathcal{T}}g\|_{H^{1/2}(\Gamma)}$  is (locally) controlled by its seminorm. Altogether, this leads to defining the *Dirichlet oscillation* with the following local indicators:

$$\begin{aligned} \text{osc}_{\mathcal{T}}(g)_{1/2}^2 &:= \sum_{F \in \mathcal{F}_{\Gamma}} \text{osc}_{\mathcal{T}}(g, F)_{1/2}^2, \\ \text{osc}_{\mathcal{T}}(g, F)_{1/2}^2 &:= \int_{\omega_F} \int_{\omega_F} \frac{|(g - I_{\mathcal{T}}g)(x) - (g - I_{\mathcal{T}}g)(y)|^2}{|x - y|^d} dx dy. \end{aligned} \quad (4.96)$$

We observe that  $\text{osc}_{\mathcal{T}}(g, F)$  is a double singular integral but computationally accessible, for instance, by using suitable quadrature provided  $g$  is continuous [Sacchi and Veeseer 2006, Section 4.1].

**Proposition 4.73 (Dirichlet oscillation)** *There exist constants  $D_1 \geq D_2 > 0$  depending on the shape regularity of  $\mathbb{T}$  and geometry of  $\Gamma$ , such that*

$$D_2 \text{osc}_{\mathcal{T}}(g)_{1/2} \leq \|\nabla e_D\|_{L^2(\Omega)} \leq D_1 \text{osc}_{\mathcal{T}}(g)_{1/2}.$$

*Proof.* The preceding derivation verifies the upper bound. For the lower one, note that for any  $v \in H^1(\Omega)$  such that  $v = g - I_{\mathcal{T}}g$  on  $\Gamma$

$$\begin{aligned} \text{osc}_{\mathcal{T}}(g)_{1/2}^2 &= \sum_{F \in \mathcal{F}_{\Gamma}} \text{osc}_{\mathcal{T}}(g, F)_{1/2}^2 \leq \sum_{F \in \mathcal{F}_{\Gamma}} \int_{\omega_F} \int_{\Gamma} \frac{|v(x) - v(y)|^2}{|x - y|^d} dx dy \\ &\lesssim \int_{\Gamma} \int_{\Gamma} \frac{|v(x) - v(y)|^2}{|x - y|^d} dx dy = |v|_{H^{1/2}(\Gamma)}^2, \end{aligned}$$

because the patches  $\omega_F$ ,  $F \in \mathcal{F}_{\Gamma}$ , possess a uniform overlapping property due to shape regularity of  $\mathbb{T}$ . Applying this to  $v = e_D$  finishes the proof.  $\square$

For suitable settings, local lower a posteriori estimates for the Dirichlet error  $e_D$  can be derived; see [Sacchi and Veiser 2006, Theorem 3.2].

Combining the Dirichlet oscillation with some Galerkin estimator  $\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f)$  by

$$\mathcal{E}_{\mathcal{T}}^{\text{Dir}}(u_{\mathcal{T}}, f, g)^2 := \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f)^2 + \text{osc}_{\mathcal{T}}(g)_{1/2}^2,$$

the preceding discussion is summarized by the following result.

**Theorem 4.74 (estimators for general Dirichlet condition)** *If Assumption 4.44 (discrete coefficients) is valid, then there exist constants  $C_L \leq C_U$  depending on  $(A, c)$ ,  $\Omega$ ,  $\Gamma$ , and the shape regularity of  $\mathbb{T}$  such that*

$$C_L \mathcal{E}_{\mathcal{T}}^{\text{Dir}}(u_{\mathcal{T}}, f, g) \leq \|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)} \leq C_U \mathcal{E}_{\mathcal{T}}^{\text{Dir}}(u_{\mathcal{T}}, f, g).$$

## 5 Convergence of AFEM for Coercive Problems

In this section we consider the coercive problem (2.5) with the intent to design and analyze three AFEMs in increasing order of complexity and applicability, depending on properties of data  $\mathcal{D}$ . Our basic regularity assumption on data reads  $\mathcal{D} = (A, c, f) \in \mathbb{D}$ , where

$$\mathbb{D} := L^\infty(\Omega; \mathbb{R}^{d \times d}) \times L^\infty(\Omega) \times H^{-1}(\Omega). \quad (5.1)$$

We approximate  $\mathcal{D}$  with discrete data  $\widehat{\mathcal{D}} = (\widehat{A}, \widehat{c}, \widehat{f}) \in \mathbb{D}_{\widehat{\mathcal{T}}}$ , where

$$\mathbb{D}_{\widehat{\mathcal{T}}} := [\mathbb{S}_{\widehat{\mathcal{T}}}^{n-1, -1}]^{d \times d} \times \mathbb{S}_{\widehat{\mathcal{T}}}^{n-1, -1} \times \mathbb{F}_{\widehat{\mathcal{T}}} \quad (5.2)$$

is subordinate to a partition  $\widehat{\mathcal{T}} \in \mathbb{T}$ . We will often assume that *data is discrete*, meaning precisely that  $\mathcal{D} = \widehat{\mathcal{D}}$ .

We start with the *one-step AFEM*, hereafter called GALERKIN, which is the standard SEMR loop

$$\text{SOLVE} \longrightarrow \text{ESTIMATE} \longrightarrow \text{MARK} \longrightarrow \text{REFINE}$$

introduced in [Dörfler 1996] and further developed in [Morin, Nochetto and Siebert 2000, 2002, Cascón *et al.* 2008]. This is the simplest algorithm in that it requires data  $\mathcal{D} = (A, c, f)$  to be discrete, but it is a building block for the other two methods. After reviewing a few crucial properties of error and estimator in Section 5.1, we fully discuss GALERKIN in Section 5.2.

The second algorithm is the *one-step AFEM with switch*, which still assumes the coefficients  $(A, c)$  to be discrete but allows for general forcing  $f \in H^{-1}(\Omega)$ . This is a new contribution of this survey that, similarly to [Kreuzer *et al.* n.d.], exploits the structure of the error estimator  $\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f)$  of Section 4

$$\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f)^2 = \eta_{\mathcal{T}}(u_{\mathcal{T}})^2 + \text{osc}_{\mathcal{T}}(f)_{-1}^2,$$

and its equivalence to the energy error. The PDE estimator  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$  relies on the discrete forcing  $P_{\mathcal{T}}f \in \mathbb{F}_{\mathcal{T}}$  and is fully computable, whereas the data oscillation  $\text{osc}_{\mathcal{T}}(f)_{-1}$  encodes the infinite dimensional nature of  $f$  and could be estimated in

important cases of practical interest further discussed in Section 7.3. The quantity  $\text{osc}_{\mathcal{T}}(f)_{-1}$  measures the deviation of  $f$  from being discrete and may dictate the pre-asymptotic regime of AFEM. Therefore,  $\text{osc}_{\mathcal{T}}(f)_{-1}$  must be handled separately from  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$ ; hence the name of the new method, hereafter called AFEM-SW. Assuming that  $\text{osc}_{\mathcal{T}}(f)_{-1}$  is computable, the module

$$[\widehat{\mathcal{T}}] = \text{DATA}(\mathcal{T}, f, \tau)$$

deals with  $\text{osc}_{\mathcal{T}}(f)_{-1}$  whenever it is large relative to  $\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f)$ . In fact, it creates an admissible refinement  $\widehat{\mathcal{T}}$  of the input mesh  $\mathcal{T}$  such that  $\text{osc}_{\widehat{\mathcal{T}}}(f)_{-1}$  is below the desired tolerance  $\tau$ , i.e.  $\text{osc}_{\widehat{\mathcal{T}}}(f)_{-1} \leq \tau$ . We explain the role of data oscillation for error analysis, design AFEM-SW and prove its linear convergence in Section 5.3.

The third algorithm deals with variable data  $\mathcal{D}$  and various degrees of regularity of  $\mathcal{D}$ , and is able to handle discontinuous coefficients  $(A, c)$  not aligned with admissible meshes  $\mathcal{T} \in \mathbb{T}$  emanating from  $\mathcal{T}_0$ . To handle the multiplicative structure of  $(A, c)$  in the model problem (2.5), we consider the following *two-step* AFEM:

**Algorithm 5.1 (AFEM-TS)** Given an initial mesh  $\mathcal{T}_0$ , an initial tolerance  $\varepsilon_0$ , and a parameter  $\omega$  sufficiently small to be determined later, iterate

$$\begin{aligned} & \text{AFEM-TS}(\mathcal{T}_0, \varepsilon_0, \omega) \\ & k = 0 \\ & [\widehat{\mathcal{T}}_k, \widehat{\mathcal{D}}_k] = \text{DATA}(\mathcal{T}_k, \mathcal{D}, \omega \varepsilon_k) \\ & [\mathcal{T}_{k+1}, u_{k+1}] = \text{GALERKIN}(\widehat{\mathcal{T}}_k, \widehat{\mathcal{D}}_k, \varepsilon_k) \\ & \varepsilon_{k+1} = \frac{1}{2} \varepsilon_k; k \leftarrow k + 1 \end{aligned}$$

This structure was first proposed in [Stevenson 2008] and further explored in [Bonito *et al.* 2013b, Cohen *et al.* 2012, Bonito *et al.* 2016, Bonito, Cascón, Morin and Nochetto 2013a, Bonito and Devaud 2015]. The three components of data  $\mathcal{D} = (A, c, f) \in \mathbb{D}$  are first approximated by discrete data  $\widehat{\mathcal{D}} = (\widehat{A}, \widehat{c}, \widehat{f}) \in \mathbb{D}_{\widehat{\mathcal{T}}}$ , as defined in (5.1) and (5.2), within the module

$$[\widehat{\mathcal{T}}, \widehat{\mathcal{D}}] = \text{DATA}(\mathcal{T}, \mathcal{D}, \tau)$$

to accuracy  $\tau = \omega \varepsilon$  significantly smaller than  $\varepsilon$ . This is achieved by an algorithm similar to Algorithm 3.18 (greedy algorithm), which is fully discussed along with applications to  $\mathcal{D}$  in Section 7. The resulting admissible refinement  $\widehat{\mathcal{T}}$  of  $\mathcal{T}$  and discrete data  $\widehat{\mathcal{D}}$  over  $\widehat{\mathcal{T}}$  are next taken by GALERKIN to reduce the PDE error to the desired tolerance  $\varepsilon$ , namely the module

$$[\mathcal{T}, u_{\mathcal{T}}] = \text{GALERKIN}(\widehat{\mathcal{T}}, \widehat{\mathcal{D}}, \varepsilon)$$

constructs a refinement  $\mathcal{T}$  of  $\widehat{\mathcal{T}}$  with discrete data  $\widehat{\mathcal{D}}$  over  $\widehat{\mathcal{T}}$  such that  $\eta_{\mathcal{T}}(u_{\mathcal{T}}) \leq \varepsilon$ . We point out that if data is discrete, i.e.  $\mathcal{D} = \widehat{\mathcal{D}}$ , then DATA is skipped and AFEM-TS reduces to GALERKIN. We tackle AFEM-TS in Section 5.4, where we prove a perturbation estimate with respect to  $\mathcal{D}$  and next discuss convergence

properties of AFEM-TS. We will extend this approach to discontinuous FEMs in Section 9 and to mixed FEMs for (2.5) as well as the Stokes system (2.14) in Section 10.

### 5.1 Properties of error and estimator

We follow Cascón, Kreuzer, Nochetto, and Siebert [Cascón *et al.* 2008] and summarize some basic properties of GALERKIN that emanate from the symmetry of the differential operator (i.e. of  $A$ ) and features of the modules. In doing this, any explicit constant or hidden constant in  $\lesssim$  will only depend on the uniform shape-regularity of  $\mathbb{T}$ , the dimension  $d$ , the polynomial degree  $n$ , and the (global) eigenvalues of  $A$ , but not on a specific grid  $\mathcal{T} \in \mathbb{T}$ , except if explicitly stated.

We recall that the bilinear form  $\mathcal{B}$  in (2.8) with continuous coefficients  $(A, c)$  is symmetric, coercive and continuous in the space  $H_0^1(\Omega)$  (see (2.30)), namely  $\|v\|_\Omega = \mathcal{B}[v, v]^{\frac{1}{2}}$  is a norm equivalent to  $|v|_{H_0^1(\Omega)}$  with equivalence constants  $0 < c_{\mathcal{B}} \leq C_{\mathcal{B}}$

$$c_{\mathcal{B}}|v|_{H_0^1(\Omega)}^2 \leq \|v\|_\Omega^2 \leq C_{\mathcal{B}}|v|_{H_0^1(\Omega)}^2 \quad \forall v \in H_0^1(\Omega). \quad (5.3)$$

The module DATA approximates  $(A, c)$  over a mesh  $\mathcal{T}$  by piecewise polynomial coefficients  $(\widehat{A}, \widehat{c})$  obeying side constraints so that the corresponding perturbed bilinear form  $\widehat{\mathcal{B}}$  still defines a uniform scalar product in  $H_0^1(\Omega)$

$$\|v\|_\Omega = \widehat{\mathcal{B}}[v, v]^{\frac{1}{2}} \quad \forall v \in H_0^1(\Omega), \quad (5.4)$$

which satisfies (5.3) with constants  $0 < c_{\widehat{\mathcal{B}}} \leq C_{\widehat{\mathcal{B}}}$  independent of  $\mathcal{T}$ . We hope this slight abuse of notation will not create confusion because we will always refer to the energy norm in (5.4) when dealing with  $\widehat{\mathcal{B}}$ . We denote by  $\widehat{u} = u(\widehat{\mathcal{D}}) \in H_0^1(\Omega)$  the solution of (2.7) with coefficients  $(\widehat{A}, \widehat{c})$  and forcing function either  $\widehat{f} = f \in H^{-1}(\Omega)$  or its projection  $\widehat{f} = P_{\mathcal{T}}f \in \mathbb{F}_{\mathcal{T}}$  defined in (4.35), namely

$$\widehat{\mathcal{B}}[\widehat{u}, v] = \langle \widehat{f}, v \rangle \quad \forall v \in H_0^1(\Omega), \quad (5.5)$$

In the sequel, we will often compare discrete functions on different meshes. Given  $\mathcal{T} \in \mathbb{T}$ , we denote by  $\mathcal{T}_* \in \mathbb{T}$  an admissible refinement of  $\mathcal{T}$  and write

$$\mathcal{T} \leq \mathcal{T}_* \quad \Leftrightarrow \quad \mathbb{T}(\mathcal{T}) \subset \mathbb{T}(\mathcal{T}_*) \quad (5.6)$$

in the sense that the supporting tree of  $\mathcal{T}$  is contained in that of  $\mathcal{T}_*$ . For any  $\mathcal{T}_* \geq \mathcal{T}$ , we have the following crucial property.

**Lemma 5.2 (Pythagoras)** *Let  $\mathcal{T}_* \geq \mathcal{T} \geq \widehat{\mathcal{T}}$  and let  $\widehat{u} \in H_0^1(\Omega)$  be the solution of (5.5) with discrete coefficients  $(\widehat{A}, \widehat{c})$  over  $\widehat{\mathcal{T}}$ . The corresponding Galerkin solutions  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}$  and  $u_{\mathcal{T}_*} \in \mathbb{V}_{\mathcal{T}_*}$  with coefficients  $(\widehat{A}, \widehat{c})$  and forcing  $f \in H^{-1}(\Omega)$  satisfy the orthogonality property*

$$\|\widehat{u} - v_{\mathcal{T}}\|_\Omega^2 = \|\widehat{u} - u_{\mathcal{T}_*}\|_\Omega^2 + \|u_{\mathcal{T}_*} - v_{\mathcal{T}}\|_\Omega^2 \quad \forall v_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}. \quad (5.7)$$

*Proof.* Exploit the nestedness property  $\mathbb{V}_{\mathcal{T}} \subset \mathbb{V}_{\mathcal{T}_*}$ , and the Galerkin orthogonality property  $\widehat{\mathcal{B}}[\widehat{u} - u_{\mathcal{T}_*}, v_{\mathcal{T}} - u_{\mathcal{T}_*}] = 0$  in  $\mathbb{V}_{\mathcal{T}_*}$  for the scalar product induced by  $\widehat{\mathcal{B}}$ .  $\square$

Property (5.7) is very restrictive: it relies on space nestedness and is valid exclusively for the energy norm. However, it is instrumental for the subsequent analysis in the energy norm or the equivalent norm  $|\cdot|_{H_0^1(\Omega)}$ , but it does not extend to other, perhaps more practical, norms such as the maximum norm. This is an important open problem and a serious limitation of this theory.

We recall that the residual a posteriori error analysis of Section 4 relies on the projection operator  $P_{\mathcal{T}} : H^{-1}(\Omega) \rightarrow \mathbb{F}_{\mathcal{T}}$ , with element and face components  $P_{\mathcal{T}}f|_T = P_T f$  for  $T \in \mathcal{T}$  and  $P_{\mathcal{T}}f|_F = P_F f$ . The *full local error indicator*

$$\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f, T)^2 = \eta_{\mathcal{T}}(u_{\mathcal{T}}, T)^2 + \text{osc}_{\mathcal{T}}(f, T)_{-1}^2$$

splits into a computable *PDE error indicator* with discrete coefficients  $(\widehat{\mathbf{A}}, \widehat{c})$

$$\eta_{\mathcal{T}}(v, T)^2 = h_T^2 \|r(v)\|_T^2 + h_T \|j(v)\|_{\partial T}^2 \quad \forall T \in \mathcal{T}, \quad (5.8)$$

where the *interior* and *jump residuals* are given by

$$\begin{aligned} r(v)|_T &= P_T f + \text{div}(\widehat{\mathbf{A}} \nabla v) - \widehat{c}v \quad \forall T \in \mathcal{T} \\ j(v)|_F &= [\widehat{\mathbf{A}} \nabla v] \cdot \mathbf{n}|_F - P_F f \quad \forall F \in \mathcal{F}, \end{aligned} \quad (5.9)$$

and  $j(v)|_F = 0$  for boundary faces  $F$ , and *data oscillation*

$$\text{osc}_{\mathcal{T}}(f, T)_{-1}^2 = \|f - P_{\mathcal{T}}f\|_{H^{-1}(\omega_T)}^2 \quad \forall T \in \mathcal{T}, \quad (5.10)$$

where  $\omega_T$  is the patch associated with  $T$ . The corresponding global quantities are

$$\begin{aligned} \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f)^2 &= \sum_{T \in \mathcal{T}} \mathcal{E}(u_{\mathcal{T}}, f, T)^2, \\ \eta_{\mathcal{T}}(u_{\mathcal{T}})^2 &= \sum_{T \in \mathcal{T}} \eta_{\mathcal{T}}(u_{\mathcal{T}}, T)^2, \quad \text{osc}_{\mathcal{T}}(f)_{-1}^2 = \sum_{T \in \mathcal{T}} \text{osc}_{\mathcal{T}}(f, T)_{-1}^2, \end{aligned} \quad (5.11)$$

and have the following a posteriori error estimates proved in Theorem 4.45 (modified residual estimator) for the  $H_0^1$ -norm.

**Proposition 5.3 (a posteriori error estimates)** *Let  $\widehat{u} \in H_0^1(\Omega)$  be the solution of (5.5) with discrete coefficients  $(\widehat{\mathbf{A}}, \widehat{c})$  over  $\mathcal{T} \in \mathbb{T}$  but general forcing  $f \in H^{-1}(\Omega)$ . Then, there exist constants  $0 < C_L \leq C_U$ , depending on the shape regularity of  $\mathbb{T}$  such that the Galerkin solution  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}$  satisfies*

$$C_L \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f) \leq |\widehat{u} - u_{\mathcal{T}}|_{H_0^1(\Omega)} \leq C_U \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f), \quad (5.12)$$

Moreover, if  $\|\cdot\|_{\Omega}$  stands for the energy norm in (5.4) with equivalence constants  $c_{\widehat{\mathcal{B}}} \leq C_{\widehat{\mathcal{B}}}$  satisfying (5.3), then (5.12) yields

$$C_2 \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f) \leq \|\widehat{u} - u_{\mathcal{T}}\|_{\Omega} \leq C_1 \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f), \quad (5.13)$$

with  $C_1 = \sqrt{C_{\widehat{\mathcal{B}}}} C_U$  and  $C_2 = \sqrt{c_{\widehat{\mathcal{B}}}} C_L$ .

There is a fundamental difference between (5.12) and earlier versions of a posteriori error estimates, going back to the seminal paper [Babuška and Miller 1987]; see also [Ainsworth and Oden 2000, Braess 2007, Nochetto *et al.* 2009, Verfürth 2013]. It is about the role of data oscillation  $\text{osc}_{\mathcal{T}}(f)_{-1}$ , which is now dominated by the error  $|\widehat{u} - u_{\mathcal{T}}|_{H_0^1(\Omega)}$  and does not spoil the lower bound. This is due to the fact that  $\text{osc}_{\mathcal{T}}(f)_{-1}$  is evaluated in the natural space  $H^{-1}(\Omega)$  and quantifies the discrepancy between  $f$  and a suitable projection  $P_{\mathcal{T}}f$  which gives rise to a quasi-best local approximation of  $f$ . We refer to [Nochetto *et al.* 2009] and [Kreuzer and Veeder 2021] for a discussion of data oscillation.

Suppose now that we have two conforming meshes  $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$  with  $\mathcal{T}_* \geq \mathcal{T}$ . Let

$$\mathcal{R} := \mathcal{R}_{\mathcal{T} \rightarrow \mathcal{T}_*} := \mathcal{T} \setminus \mathcal{T}_* \quad (5.14)$$

be the subset of refined elements of  $\mathcal{T}$ , namely, those elements in  $\mathcal{T}$  that are no longer in  $\mathcal{T}_*$ . We stress that the upper bound in (5.12) cannot be local due to the nonlocal nature of the error  $|\widehat{u} - u_{\mathcal{T}}|_{H_0^1(\Omega)}$ . However, in view of Theorem 4.48 (upper bound for corrections), the following remarkable local upper bound for Galerkin solutions  $u_{\mathcal{T}} \in V_{\mathcal{T}}, u_{\mathcal{T}_*} \in V_{\mathcal{T}_*}$  holds

$$\|u_{\mathcal{T}} - u_{\mathcal{T}_*}\|_{\Omega} \leq C_1 \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f, \mathcal{R}), \quad (5.15)$$

where for  $\mathcal{S} \subset \mathcal{T}$ ,  $\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f, \mathcal{S}) := \left( \sum_{T \in \mathcal{S}} \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f, T)^2 \right)^{1/2}$  is the error estimator restricted to  $\mathcal{S}$ . Consequently, solely the elements where  $\mathcal{T}$  and  $\mathcal{T}_*$  differ, namely the set  $\mathcal{R}$ , account for the discrepancy between  $u_{\mathcal{T}}$  and  $u_{\mathcal{T}_*}$ . This turns out to be consistent with (5.13) because  $\mathcal{T}$  has to be refined everywhere to get to  $\widehat{u}$ , whence  $\mathcal{R} = \mathcal{T}$ .

In contrast to the upper bound in (5.12), the corresponding lower bound is local according to Theorem 4.45 (modified residual estimator). This is due to the local nature of the PDE (2.5). However, when comparing  $u_{\mathcal{T}}$  and  $u_{\mathcal{T}_*}$ , this bound is not valid unless the *interior vertex property* (given in Definition 4.50) is satisfied [Morin *et al.* 2000]; in fact, we present a counterexample later in Example 5.7 taken from [Morin *et al.* 2000].

The interior vertex property is valid upon enforcing a fixed number  $b$  of bisections ( $b = 3, 6$  for  $d = 2, 3$ ). An immediate consequence, proved in Theorem 4.51 (local lower bound for corrections), is the *discrete* lower a posteriori bound for piecewise constant diffusion coefficient  $\mathbf{A}$  and reaction coefficient  $c = 0$  on  $\mathcal{T}_0$ ,

$$C_{L,1} \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f, \mathcal{M}) \leq \|u_{\mathcal{T}} - u_{\mathcal{T}_*}\|_{\Omega} + C_{L,2} \text{osc}_{\mathcal{T}}(f, \omega(\mathcal{M}))_{-1}, \quad (5.16)$$

where  $\omega(\mathcal{M}) := \cup\{\omega_T : T \in \mathcal{M}\}$  is the union of all patches of elements in  $\mathcal{M}$  and  $\text{osc}_{\mathcal{T}}(f, \omega(\mathcal{M}))_{-1}^2 = \sum_{T \in \omega(\mathcal{M})} \text{osc}_{\mathcal{T}}(f, T)_{-1}^2$ ; we refer to [Morin *et al.* 2000, 2002]. We stress that if  $f = P_{\mathcal{T}}f$  is discrete, then  $\text{osc}_{\mathcal{T}}(f)_{-1} = 0$  and (5.16) reduces to

$$C_2 \eta_{\mathcal{T}}(u_{\mathcal{T}}, \mathcal{M}) \leq \|u_{\mathcal{T}} - u_{\mathcal{T}_*}\|_{\Omega}. \quad (5.17)$$

One serious difficulty in dealing with AFEM is that one has access to the energy

error  $\|\widehat{u} - u_{\mathcal{T}}\|_{\Omega}$ , or equivalently to  $|\widehat{u} - u_{\mathcal{T}}|_{H_0^1(\Omega)}$ , only through the full error estimator  $\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f)$ . Lemma 5.2 (Pythagoras) implies monotonicity of the energy error with respect to  $\mathcal{T}$ , namely for  $\mathcal{T}_* \geq \mathcal{T}$

$$\|\widehat{u} - u_{\mathcal{T}_*}\|_{\Omega} \leq \|\widehat{u} - u_{\mathcal{T}}\|_{\Omega}.$$

However, the PDE estimator  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$  fails to be monotone for fixed discrete coefficients  $(\widehat{A}, \widehat{c})$  because it depends on the discrete solution  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}$  that changes with the mesh. The following estimate, proved in Proposition 4.56 (estimator reduction), quantifies the deviation of  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$  from monotonicity: there exists  $\lambda > 0$  such that for any  $\delta > 0$ ,  $v \in \mathbb{V}_{\mathcal{T}}$  and  $v_* \in \mathbb{V}_{\mathcal{T}_*}$ ,

$$\begin{aligned} \eta_{\mathcal{T}_*}(v_*, \mathcal{T}_*)^2 &\leq (1 + \delta) (\eta_{\mathcal{T}}(v, \mathcal{T})^2 - \lambda \eta_{\mathcal{T}}(v, \mathcal{M})^2) \\ &\quad + 2(1 + \delta^{-1}) C_{\text{Lip}} \left( \|v_* - v\|_{\Omega}^2 + \sum_{T \in \mathcal{T}_*} \|P_{\mathcal{T}} f - P_{\mathcal{T}_*} f\|_{H^{-1}(\omega_T)}^2 \right), \end{aligned}$$

where  $C_{\text{Lip}}$  depends on  $(A, c)$  and the shape regularity constant of  $\mathbb{T}$ . We refer to [Cascón *et al.* 2008, Morin *et al.* 2008] for the case  $P_{\mathcal{T}} f = P_{\mathcal{T}_*} f = f \in L^2(\Omega)$ .

## 5.2 Convergence for discrete data: one-Step AFEM

We now present the four basic modules of GALERKIN, the one-step AFEM within Algorithm 5.1 (AFEM-TS), namely

$$\text{SOLVE} \longrightarrow \text{ESTIMATE} \longrightarrow \text{MARK} \longrightarrow \text{REFINE}, \quad (5.18)$$

discuss their main properties, and prove a contraction property between consecutive iterates of GALERKIN. According to Algorithm 5.1, given discrete data  $\widehat{\mathcal{D}}$  over a conforming mesh  $\widehat{\mathcal{T}}$ , created by DATA, and a desired tolerance  $\varepsilon > 0$ , the module

$$[\mathcal{T}, u_{\mathcal{T}}] = \text{GALERKIN}(\widehat{\mathcal{T}}, \widehat{\mathcal{D}}, \varepsilon) \quad (5.19)$$

stops the loop (5.18) as soon as the error tolerance  $\varepsilon$  is reached, i.e. as soon as

$$\eta_{\mathcal{T}}(u_{\mathcal{T}}) \leq \varepsilon. \quad (5.20)$$

Since the data never change within GALERKIN and is always discrete, we assume in this section that  $\mathcal{D} \in \mathbb{D}_{\mathcal{T}}$  and do not use the hat symbol to indicate quantities defined using the (discrete) data.

### 5.2.1 Modules of GALERKIN

*Module SOLVE.* If  $\mathcal{T} \in \mathbb{T}$  is a conforming refinement of  $\mathcal{T}_0$ , and  $\mathbb{V}_{\mathcal{T}}$  is the finite element space of  $C^0$  piecewise polynomials of degree  $\leq n$ , then

$$[u_{\mathcal{T}}] = \text{SOLVE}(\mathcal{T})$$

determines the Galerkin FEM solution *exactly*, namely without algebraic error,

$$u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}} : \quad \mathcal{B}[u_{\mathcal{T}}, v] = \int_{\Omega} \nabla v \cdot A \nabla u_{\mathcal{T}} + c v u = \langle f, v \rangle, \quad (5.21)$$

where  $f \in H^{-1}(\Omega)$ . However, if  $f \in \mathbb{F}_{\mathcal{T}}$  is discrete as defined in (4.35), then

$$\langle f, v \rangle = \sum_{T \in \mathcal{T}} \int_T q_T v + \sum_{F \in \mathcal{F}} \int_F q_F v \quad \forall v \in \mathbb{V}_{\mathcal{T}}.$$

The assumption of exact solvability is made for simplicity. The algebraic error committed in solving (5.21) by iterative solvers can be accommodated within the forthcoming theory. We refer to [Stevenson 2007, Daniel and Vohralik 2023] for details about how to relate the algebraic and PDE errors.

*Module ESTIMATE.* Given a conforming mesh  $\mathcal{T} \in \mathbb{T}$  and the Galerkin solution  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}$ , the output of

$$[\{\eta_{\mathcal{T}}(u_{\mathcal{T}}, T), \text{osc}_{\mathcal{T}}(f, T)_{-1}\}_{T \in \mathcal{T}}] = \text{ESTIMATE}(u_{\mathcal{T}}, \mathcal{T}, \mathcal{D})$$

are the element error indicators  $\eta_{\mathcal{T}}(u_{\mathcal{T}}, T)$  defined in (5.8) with the discrete data  $\mathcal{D}$ , namely

$$\eta_{\mathcal{T}}(u_{\mathcal{T}}, T)^2 = h_T^2 \|r(u_{\mathcal{T}})\|_T^2 + h_T \|j(u_{\mathcal{T}})\|_{\partial T}^2 \quad T \in \mathcal{T},$$

and element data oscillation  $\text{osc}_{\mathcal{T}}(f, T)_{-1}$  defined in (5.10), namely

$$\text{osc}_{\mathcal{T}}(f, T)_{-1} = \|f - P_{\mathcal{T}}f\|_{H^{-1}(\omega_T)}.$$

We observe that for discrete forcing  $f = P_{\mathcal{T}}f$ , global data oscillation vanishes

$$\text{osc}_{\mathcal{T}}(f)_{-1} = \|f - P_{\mathcal{T}}f\|_{H^{-1}(\Omega)} = 0; \quad (5.22)$$

this property is always valid within GALERKIN. In this case, the output of ESTIMATE reduces to just the PDE error indicators. Given  $\mathcal{S} \subset \mathcal{T}$ , we denote

$$\eta_{\mathcal{T}}(v, \mathcal{S})^2 := \sum_{T \in \mathcal{S}} \eta_{\mathcal{T}}(v, T)^2, \quad \eta_{\mathcal{T}}(v) = \eta_{\mathcal{T}}(v, \mathcal{T}) \quad v \in \mathbb{V}_{\mathcal{T}}.$$

*Module MARK.* Given  $\mathcal{T} \in \mathbb{T}$ , the Galerkin solution  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}$ , and element error indicators  $\{\eta_{\mathcal{T}}(u_{\mathcal{T}}, T)\}_{T \in \mathcal{T}}$ , the module MARK selects elements for refinement using *Dörfler Marking* (or bulk chasing) [Dörfler 1996, Morin *et al.* 2000, Nochetto *et al.* 2009, Nochetto and Veeser 2012], i. e., given a parameter  $\theta \in (0, 1]$  the output  $\mathcal{M}$  of

$$[\mathcal{M}] = \text{MARK}(\{\eta_{\mathcal{T}}(u_{\mathcal{T}}, T)\}_{T \in \mathcal{T}}, \mathcal{T}, \theta)$$

satisfies

$$\eta_{\mathcal{T}}(u_{\mathcal{T}}, \mathcal{M}) \geq \theta \eta_{\mathcal{T}}(u_{\mathcal{T}}, \mathcal{T}). \quad (5.23)$$

This marking guarantees that  $\mathcal{M}$  contains a substantial part of the total (or bulk) error, thus its name. The choice of  $\mathcal{M}$  does not have to be minimal at this stage, that is, the marked elements  $T \in \mathcal{M}$  do not necessarily must be those with largest indicators.



*Module REFINE.* Let  $b \in \mathbb{N}$  be the number of desired bisections per marked element. Given  $\mathcal{T} \in \mathbb{T}$  and a subset  $\mathcal{M}$  of marked elements, the output  $\mathcal{T}_* \in \mathbb{T}$  of

$$[\mathcal{T}_*] = \text{REFINE}(\mathcal{T}, \mathcal{M})$$

is the smallest admissible refinement  $\mathcal{T}_*$  of  $\mathcal{T}$  so that all elements of  $\mathcal{M}$  are at least bisected  $b$  times. Therefore, we have  $h_{\mathcal{T}_*} \leq h_{\mathcal{T}}$  and the strict reduction property

$$h_{\mathcal{T}_*}|_T \leq 2^{-b/d} h_{\mathcal{T}}|_T \quad \forall T \in \mathcal{M}, \quad (5.24)$$

where  $h_{\mathcal{T}} : \Omega \rightarrow \mathbb{R}^+$  is a piecewise constant meshsize function that coincides with  $h_T = |T|^{1/d}$  on every  $T \in \mathcal{T}$ . We finally let

$$\mathcal{R} := \mathcal{R}_{\mathcal{T} \rightarrow \mathcal{T}_*} := \mathcal{T} \setminus \mathcal{T}_*$$

be the subset of refined elements of  $\mathcal{T}$  and note that  $\mathcal{M} \subseteq \mathcal{R}$ .

Concatenating these four modules we get the standard SEMR one-step AFEM.

**Algorithm 5.4 (GALERKIN)** Let  $\mathcal{T} \geq \mathcal{T}_0$  be a conforming refinement of a suitable initial mesh  $\mathcal{T}_0$ . Let data  $\mathcal{D} = (A, c, f) \in D_{\mathcal{T}}$  be discrete on  $\mathcal{T}$  and  $\varepsilon > 0$  be a stopping tolerance. The following one-step AFEM creates a conforming refinement  $\mathcal{T}_* \geq \mathcal{T}$  and Galerkin solution  $u_{\mathcal{T}_*} \in \mathbb{V}_{\mathcal{T}_*}$  for data  $\mathcal{D}$  such that  $\eta_{\mathcal{T}_*}(u_{\mathcal{T}_*}) \leq \varepsilon$ :

```

 $[\mathcal{T}_*, u_{\mathcal{T}_*}] = \text{GALERKIN}(\mathcal{T}, \mathcal{D}, \varepsilon)$ 
  set  $j = 0, \mathcal{T}_0 = \mathcal{T}$ 
  do
     $[u_j] = \text{SOLVE}(\mathcal{T}_j)$ 
     $[\{\eta_j(u_j, T)\}_{T \in \mathcal{T}_j}] = \text{ESTIMATE}(u_j, \mathcal{T}_j, \mathcal{D})$ 
    if  $\eta_j(u_j) \leq \varepsilon$ 
      return  $\mathcal{T}_j, u_j$ 
     $[\mathcal{M}_j] = \text{MARK}(\{\eta_j(u_j, T)\}_{T \in \mathcal{T}_j}, \mathcal{T}_j, \theta)$ 
     $[\mathcal{T}_{j+1}] = \text{REFINE}(\mathcal{T}_j, \mathcal{M}_j)$ 
     $j \leftarrow j + 1$ 
  while true

```

### 5.2.2 Contraction property of GALERKIN

A key question to ask is what is (are) the quantity(ies) that GALERKIN may contract. In light of (5.7), an obvious candidate is the energy error  $\|u - u_j\|_{\Omega}$ , where  $u_j \in \mathbb{V}_j = \mathbb{V}_{\mathcal{T}_j}$  solves the problem

$$\mathcal{B}[u_j, w] = \langle f, w \rangle \quad \forall w \in \mathbb{V}_j. \quad (5.25)$$

We now show that this is in fact the case for discrete data  $\mathcal{D} \in \mathbb{D}_{\mathcal{T}}$  provided the *discrete local estimate* (5.17) holds. The latter is a consequence of the *interior vertex property* of Definition 4.50 whenever  $A$  is piecewise constant and  $c = 0$  in  $\mathcal{T}$  and data oscillation vanishes  $\text{osc}_{\mathcal{T}}(f)_{-1} = 0$  [Morin *et al.* 2000, 2002].

**Lemma 5.5 (contraction property with discrete lower bound)** *Let data  $\mathcal{D} \in \mathbb{D}_{\mathcal{T}}$*

be discrete and  $u = u(\mathcal{D}) \in H_0^1(\Omega)$  be the corresponding exact solution. If the subset  $\mathcal{M}_j \subset \mathcal{T}_j$  of elements marked by MARK satisfies the discrete local estimate (5.17) with respect to  $\mathcal{T}_{j+1} \geq \mathcal{T}_j$ , then for  $\alpha := \left(1 - \left(\theta \frac{C_2}{C_1}\right)^2\right)^{1/2} < 1$  the Galerkin solutions  $u_j \in \mathbb{V}_j, u_{j+1} \in \mathbb{V}_{j+1}$  of (5.25) satisfy

$$\|u - u_{j+1}\|_\Omega \leq \alpha \|u - u_j\|_\Omega, \quad (5.26)$$

where  $0 < \theta < 1$  is the parameter in (5.23) and  $C_1 \geq C_2$  are the constants in (5.13) and (5.17) respectively.

*Proof.* For convenience, we use the notation

$$e_j = \|u - u_j\|_\Omega, \quad E_j = \|u_{j+1} - u_j\|_\Omega, \quad \eta_j = \eta_j(u_j, \mathcal{T}_j), \quad \eta_j(\mathcal{M}_j) = \eta_j(u_j, \mathcal{M}_j)$$

and recall that  $\mathcal{E}_{\mathcal{T}_j}(u_j, f) = \eta_j$  because  $\text{osc}_{\mathcal{T}_j}(f)_{-1} = 0$ . The key idea is to use the Pythagoras equality (5.7), namely  $e_{j+1}^2 = e_j^2 - E_j^2$ , and show that  $E_j$  is a significant portion of  $e_j$ . Since (5.17) implies

$$C_2 \eta_j(\mathcal{M}_j) \leq E_j,$$

applying Dörfler marking (5.23) and the upper bound in (5.13), we deduce

$$E_j^2 \geq C_2^2 \theta^2 \eta_j^2 \geq \left(\theta \frac{C_2}{C_1}\right)^2 e_j^2.$$

This is the desired property of  $E_j$  and leads to (5.26).  $\square$

The contraction property (5.26) is very special and only valid for the energy norm. For the  $H_0^1$ -norm we have the following simple but useful consequence.

**Corollary 5.6 (linear convergence)** *If  $c_{\mathcal{B}} \leq C_{\mathcal{B}}$  are the constants in (5.3), then*

$$|u - u_k|_{H_0^1(\Omega)} \leq \sqrt{\frac{C_{\mathcal{B}}}{c_{\mathcal{B}}}} \alpha^{k-j} |u - u_j|_{H_0^1(\Omega)} \quad k \geq j \geq 0.$$

We wonder whether or not the interior vertex property is necessary for (5.17), and thus for (5.26). We present an example, introduced in [Morin *et al.* 2000, 2002] to justify such a property for constant data and  $n = 1$ .

**Example 5.7 (lack of strict error monotonicity)** Let  $\Omega = (0, 1)^2$ ,  $\mathbf{A} = \mathbf{I}$ ,  $c = 0$ ,  $f = 1$  (constant data), and consider the sequences of meshes depicted in Figure 5.1. If  $\phi_0$  denotes the basis function associated with the only interior vertex of the initial mesh  $\mathcal{T}_0$ , then  $u_0 = u_1 = \frac{1}{12} \phi_0$  and  $u_2 \neq u_1$ .

The mesh  $\mathcal{T}_1 \geq \mathcal{T}_0$  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  $\|u - u_0\|_\Omega = \|u - u_1\|_\Omega$ , whence (5.17) fails, between two consecutive steps of GALERKIN for  $b = d = 2$ . This is no longer true provided an interior vertex in each marked element is created, because then Lemma 5.5 (contraction property with discrete lower bound) holds.

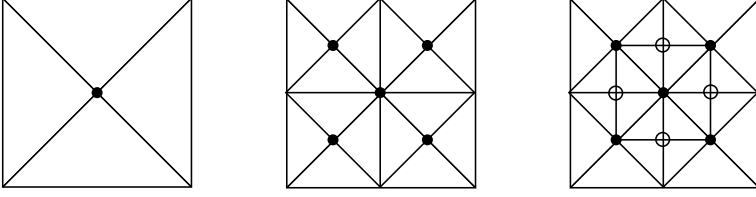


Figure 5.1. Grids  $\mathcal{T}_0$ ,  $\mathcal{T}_1$ , and  $\mathcal{T}_2$  of Example 5.7. The mesh  $\mathcal{T}_1$  has nodes in the middle of edges of  $\mathcal{T}_0$ , but only  $\mathcal{T}_2$  has nodes in the interior of elements of  $\mathcal{T}_0$ . Hence,  $\mathcal{T}_2$  satisfies the interior vertex property of Definition 4.50 with respect to  $\mathcal{T}_0$  whereas  $\mathcal{T}_1$  does not.

*Circumventing the discrete lower bound.* Enforcing (5.17) requires a minimal number  $b_*$  of bisections, say  $b_* = 3, 6$  for  $d = 2, 3$ , to guarantee the interior vertex property. This can be quite taxing, especially for  $d = 3$ , and relies on the strong assumption of  $\mathbf{A}$  being piecewise constant and  $c = 0$  on  $\mathcal{T}$ . It is clear from the preceding discussion that the energy error alone cannot be expected to contract between consecutive iterates. We explore next what quantity to monitor instead of the energy error in the analysis, with the aim of avoiding (5.17) and building a theory applicable to general discrete coefficients  $(\mathbf{A}, c)$ . This exploits the special structure of residual estimators and does not directly extend to non-residual estimators.

*Heuristics.* According to (5.7), the energy error is monotone  $\|u - u_{j+1}\|_\Omega \leq \|u - u_j\|_\Omega$ , but the previous Example shows that strict inequality may fail. However, if  $u_{j+1} = u_j$ , estimate (4.67) reveals a strict estimator reduction  $\eta_{j+1}(u_{j+1}) < \eta_j(u_j)$ . We thus expect that, for a suitable scaling factor  $\gamma > 0$ , the so-called *quasi error*

$$\zeta_j^2(u_j) := \|u - u_j\|_\Omega^2 + \gamma \eta_j^2(u_j) \quad (5.27)$$

may contract. This heuristics illustrates a distinct aspect of AFEM theory, the interplay between continuous quantities such the energy error  $\|u - u_j\|_\Omega$  and discrete ones such as the estimator  $\eta_j(u_j)$ : no one alone has the requisite properties to yield a contraction between consecutive adaptive steps. This result was originally proven in [Cascón et al. 2008].

**Theorem 5.8 (general contraction property)** *Let  $\mathcal{D} \in \mathbb{D}_\mathcal{T}$  be discrete data. Let  $\theta \in (0, 1]$  be the Dörfler marking parameter, and  $\{\mathcal{T}_j, \mathbb{V}_j, u_j\}_{j=0}^\infty$  be a sequence of conforming meshes, finite element spaces and discrete solutions  $u_j \in \mathbb{V}_j$  created by GALERKIN for the model problem (5.25). If  $u = u(\mathcal{D}) \in H_0^1(\Omega)$  is the exact solution of (5.5), then there exist constants  $\gamma > 0$  and  $0 < \alpha < 1$ , additionally depending on the number  $b \geq 1$  of bisections and  $\theta$ , such that for all  $j \geq 0$*

$$\|u - u_{j+1}\|_\Omega^2 + \gamma \eta_{j+1}^2(u_{j+1}) \leq \alpha^2 \left( \|u - u_j\|_\Omega^2 + \gamma \eta_j^2(u_j) \right). \quad (5.28)$$

*Proof.* We split the proof into four steps and use the notation in Lemma 5.5 (contraction property with discrete lower bound)

[1] The error orthogonality (5.7) reads

$$e_{j+1}^2 = e_j^2 - E_j^2. \quad (5.29)$$

Employing Proposition 4.56 (estimator reduction) with  $\mathcal{T} = \mathcal{T}_j$ ,  $\mathcal{T}_* = \mathcal{T}_{j+1}$ ,  $v = u_j$ ,  $v_* = u_{j+1}$  and  $f = f_* \in \mathbb{F}_j$  gives

$$\eta_{j+1}^2 \leq (1 + \delta)(\eta_j^2 - \lambda \eta_j^2(\mathcal{M}_j)) + (1 + \delta^{-1}) C_{\text{Lip}}^2 E_j^2. \quad (5.30)$$

After multiplying (5.30) by  $\gamma > 0$ , to be determined later, we add (5.29) and (5.30) to obtain

$$e_{j+1}^2 + \gamma \eta_{j+1}^2 \leq e_j^2 + (\gamma(1 + \delta^{-1}) C_{\text{Lip}}^2 - 1) E_j^2 + \gamma(1 + \delta)(\eta_j^2 - \lambda \eta_j^2(\mathcal{M}_j)).$$

[2] We now choose the parameters  $\delta, \gamma$ : let  $\delta$  satisfy

$$(1 + \delta)(1 - \lambda \theta^2) = 1 - \frac{\lambda \theta^2}{2},$$

and  $\gamma$  verify

$$\gamma(1 + \delta^{-1}) C_{\text{Lip}}^2 = 1.$$

Note that this choice of  $\gamma$  yields

$$e_{j+1}^2 + \gamma \eta_{j+1}^2 \leq e_j^2 + \gamma(1 + \delta)(\eta_j^2 - \lambda \eta_j^2(\mathcal{M}_j)). \quad (5.31)$$

[3] We next employ Dörfler Marking (5.23), namely  $\eta_j(\mathcal{M}_j) \geq \theta \eta_j$ , to deduce

$$e_{j+1}^2 + \gamma \eta_{j+1}^2 \leq e_j^2 + \gamma(1 + \delta)(1 - \lambda \theta^2) \eta_j^2.$$

This, in conjunction with the choice of  $\delta$ , gives

$$e_{j+1}^2 + \gamma \eta_{j+1}^2 \leq e_j^2 + \gamma \left(1 - \frac{\lambda \theta^2}{2}\right) \eta_j^2 \quad (5.32)$$

which we write

$$e_{j+1}^2 + \gamma \eta_{j+1}^2 \leq e_j^2 - \frac{\gamma \lambda \theta^2}{4} \eta_j^2 + \gamma \left(1 - \frac{\lambda \theta^2}{4}\right) \eta_j^2.$$

[4] Finally, the upper bound in (5.13), namely  $e_j \leq C_1 \eta_j$ , implies that

$$e_{j+1}^2 + \gamma \eta_{j+1}^2 \leq \left(1 - \frac{\gamma \lambda \theta^2}{4 C_1^2}\right) e_j^2 + \gamma \left(1 - \frac{\lambda \theta^2}{4}\right) \eta_j^2.$$

This in turn leads to

$$e_{j+1}^2 + \gamma \eta_{j+1}^2 \leq \alpha^2 (e_j^2 + \gamma \eta_j^2),$$

with  $\alpha^2 := \max \left\{1 - \frac{\gamma \lambda \theta^2}{4 C_1^2}, 1 - \frac{\lambda \theta^2}{4}\right\} < 1$ , and thus concludes the proof of theorem.  $\square$

**Remark 5.9 (basic ingredients)** This proof solely uses Dörfler marking (5.23); Pythagoras identity (5.7); the a posteriori upper bound in (5.13); and Proposition 4.56 (estimator reduction). The proof circumvents the use altogether of the lower bound in (5.13) and the discrete lower bound (5.17).

The contraction property (5.28) is valid for a suitable combination of the energy norm  $\|u - u_j\|_\Omega$  and the PDE estimator  $\eta_j(u_j)$ . We cannot expect this type of result for the underlying space norm  $|u - u_j|_{H_0^1(\Omega)}$ . We have instead the following statement, whose structure reflects the possible stagnation of  $|u - u_j|_{H_0^1(\Omega)}$  during the refinement process, as documented in Example 5.7.

**Corollary 5.10 (linear convergence of error)** *If the assumptions of Theorem 5.8 are valid, and  $0 < \alpha < 1$ ,  $\gamma > 0$  are the constants in (5.28), then there holds*

$$|u - u_k|_{H_0^1(\Omega)} \leq C_* \alpha^{k-j} |u - u_j|_{H_0^1(\Omega)} \quad \forall k \geq j \geq 0, \quad (5.33)$$

with  $C_* = \left(\frac{C_{\mathcal{B}}}{c_{\mathcal{B}}} \left(1 + \frac{\gamma}{C_2^2}\right)\right)^{1/2} > 1$  and constants  $C_{\mathcal{B}} \geq c_{\mathcal{B}} > 0$  and  $C_2 > 0$  given in (5.3) and (5.13) respectively.

*Proof.* Simply concatenate (5.3), (5.28) and (5.13) to obtain

$$\begin{aligned} c_{\mathcal{B}} |u - u_k|_{H_0^1(\Omega)}^2 &\leq \|u - u_k\|_\Omega^2 + \gamma \eta_k(u_k)^2 \\ &\leq \alpha^{2(k-j)} \left( \|u - u_j\|_\Omega^2 + \gamma \eta_j(u_j)^2 \right) \\ &\leq \alpha^{2(k-j)} \left( C_{\mathcal{B}} \left(1 + \frac{\gamma}{C_2^2}\right) \right) |u - u_j|_{H_0^1(\Omega)}^2. \end{aligned}$$

This implies (5.33) and concludes the proof.  $\square$

We stress that, in contrast to (5.28), (5.33) does rely on the lower bound in (5.13). This is not the case if we express linear convergence in terms of the PDE estimator. The proof is similar to the preceding one and is omitted.

**Corollary 5.11 (linear convergence of estimator)** *If the assumptions of Theorem 5.8 are valid, and  $0 < \alpha < 1$ ,  $\gamma > 0$  are the constants in (5.28), then there holds*

$$\eta_k(u_k) \leq C_\# \alpha^{k-j} \eta_j(u_j) \quad \forall k \geq j \geq 0, \quad (5.34)$$

with  $C_\# = \left(1 + \frac{C_1^2}{\gamma}\right)^{1/2} > 1$  and  $C_1$  given in (5.13).

**Remark 5.12 (stopping)** In view of (5.34), (5.12), we realize that GALERKIN requires  $j \leq J$  iterations until the stopping criteria  $\eta_j \leq \varepsilon$  is satisfied and delivers the error  $|u - u_j|_{H_0^1(\Omega)} \leq C_U \varepsilon$ , where

$$J \leq 1 + \frac{\log \frac{\varepsilon}{C_\# \eta_0}}{\log \alpha}.$$

### 5.2.3 Discontinuous coefficients: Kellogg's example

We examine a simple yet quite demanding example with piecewise constant coefficients in checkerboard pattern for  $d = 2$  due to [Kellogg 1974/75], and used in [Morin *et al.* 2000, 2002] as a benchmark for GALERKIN. We consider  $\Omega = (-1, 1)^2$ ,  $A = a_1 I$  in the first and third quadrants, and  $A = a_2 I$  in the second and fourth quadrants. This checkerboard pattern is the worst for the regularity of the solution  $u$  at the origin. For  $f = c = 0$ , a function of the form  $u(r, \theta) = r^\gamma \mu(\theta)$  in polar coordinates solves (2.5) with nonvanishing Dirichlet condition for suitable  $0 < \gamma < 2$  and  $\mu$  [Morin *et al.* 2000, 2002, Nochetto *et al.* 2009]. We choose  $\gamma = 0.1$ , which leads to  $u \in H^s(\Omega)$  for  $1 \leq s < 1.1$  and piecewise in  $W_p^2$  for some  $p > 1$ . This corresponds to diffusion coefficients  $a_1 \approx 161.44$  and  $a_2 = 1$ , which can be computed via Newton's method; the closer  $\gamma$  is to 0, the larger is the ratio  $a_1/a_2$ . The solution  $u$  and a sample mesh are depicted in Figure 5.2(left).

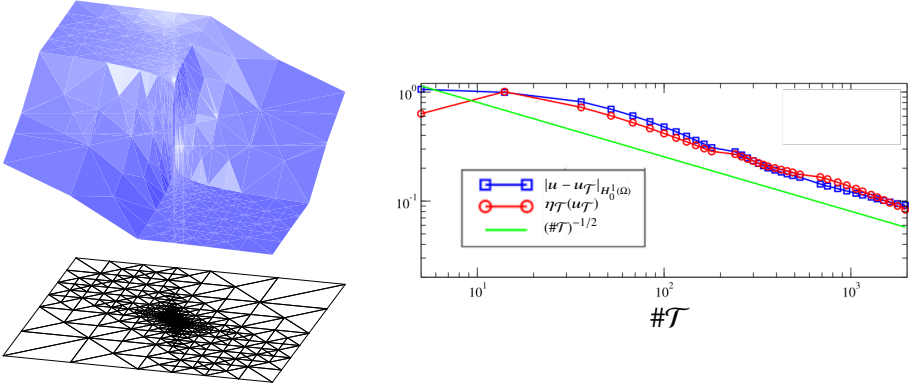


Figure 5.2. Discontinuous coefficients in checkerboard pattern: (left) graph of the discrete solution  $u$ , which is  $u \approx r^{0.1}$ , and underlying strongly graded grid  $\mathcal{T}$  towards the origin (notice the steep gradient of  $u$  at the origin); (right) estimate and true error in terms of  $\#\mathcal{T}$  (the optimal decay for piecewise linear elements in  $2d$  is indicated by the green line with slope  $-1/2$ ).

Figure 5.2 (right) documents the optimal performance of GALERKIN: both the energy error and estimator exhibit optimal decay  $(\#\mathcal{T})^{-1/2}$  in terms of the cardinality  $\#\mathcal{T}$  of the underlying mesh  $\mathcal{T}$  for piecewise linear finite elements. On the other hand, Figure 5.3 displays a strongly graded mesh  $\mathcal{T}$  towards the origin generated by GALERKIN using bisection, and three zooms which reveal a selfsimilar structure. It is worth stressing that the meshsize is of order  $10^{-10}$  at the origin and that  $\#\mathcal{T} \approx 2 \times 10^3$ , whereas to reach a similar resolution with a uniform mesh  $\mathcal{T}$  we would need  $\#\mathcal{T} \approx 10^{20}$ . This example clearly reveals that adaptivity can restore optimal performance even with modest computational resources.

Classical FEMs with quasi-uniform meshes  $\mathcal{T}$  require regularity  $u \in H^2(\Omega)$  to deliver an optimal convergence rate  $(\#\mathcal{T})^{-1/2}$  with polynomial degree  $n = 1$ . Since

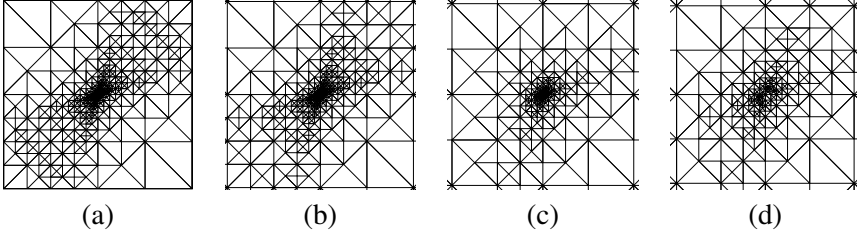


Figure 5.3. Discontinuous coefficients in checkerboard pattern: (a) final grid  $\mathcal{T}$  highly graded towards the origin with cardinality  $\#\mathcal{T} \approx 2000$ ; (b) zoom to  $(-10^{-3}, 10^{-3})^2$ ; (c) zoom to  $(-10^{-6}, 10^{-6})^2$ ; (d) zoom to  $(-10^{-9}, 10^{-9})^2$ . For a similar resolution, a uniform grid  $\mathcal{T}$  would require cardinality  $\#\mathcal{T} \approx 10^{20}$ .

$u \notin H^s(\Omega)$  for any  $s > 1.1$ , this is not possible for the example above. However, the problem is not quite the lack of second derivatives, but rather the fact that they are not square integrable. In fact, the function  $u$  is in  $W_p^2$  for  $p > 1$  in each quadrant, and so over the initial mesh  $\mathcal{T}_0$ , namely  $u \in W_p^2(\Omega; \mathcal{T}_0)$ . The computational rate of convergence  $(\#\mathcal{T})^{-1/2}$  is consistent with Corollary 3.20. We will prove that GALERKIN delivers this rate in Section 6.

### 5.3 Data oscillation: one-Step AFEM with switch

In Section 5.2 we assumed that the full data  $\mathcal{D} = (\mathbf{A}, c, f) \in \mathbb{D}_{\mathcal{T}}$  is discrete, and in particular  $f = P_{\mathcal{T}}f \in \mathbb{F}_{\mathcal{T}}$ . The finite dimensional nature of  $\widehat{\mathcal{D}}$  allowed us to develop a rather simple theory of convergence for GALERKIN, the one-step AFEM, that hinges exclusively on the PDE local error indicator  $\eta_{\mathcal{T}}(u_{\mathcal{T}}, T)$  defined in (5.8). We now keep  $(\mathbf{A}, c)$  discrete, whence the elliptic operator in (2.5) includes the Laplacian, but explore the role of a general forcing  $f \neq P_{\mathcal{T}}f$ . Therefore, in contrast with (5.22), we now investigate the effect of data oscillation (5.11)

$$\text{osc}_{\mathcal{T}}(f)_{-1}^2 = \sum_{T \in \mathcal{T}} \|f - P_{\mathcal{T}}f\|_{H^{-1}(\omega_T)}^2$$

for any  $\mathcal{T} \in \mathbb{T}$ , and present a linear convergence theory. We recall from Theorem 4.45 (modified residual estimator) that the total error estimator  $\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f)^2 = \eta_{\mathcal{T}}(u_{\mathcal{T}})^2 + \text{osc}_{\mathcal{T}}(f)_{-1}^2$  is equivalent to the  $H^1$ -error, namely

$$C_L \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f) \leq \|\nabla(u - u_{\mathcal{T}})\|_{L^2(\Omega)} \leq C_U \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f). \quad (5.35)$$

As in the previous section, to simplify notation we do not use the hat symbol to indicate quantities defined with the discrete data  $(\mathbf{A}, c)$ .

#### 5.3.1 Role of data oscillation

At a first sight, it might seem that Example 5.7 (lack of strict error monotonicity) is very special and can only occur at the beginning of the refinement process. We now show that this situation can happen at any stage and that even an interior vertex property may not guarantee error or data oscillation decrease.

**Example 5.13 (interior vertex)** Let the polynomial degree be  $n = 1$ , fix  $m \in \mathbb{N}$  and consider (5.21) with  $A = I$  the identity matrix,  $c = 0$ ,  $\Omega = (0, 1)^2$  and checkerboard  $f$  given by the following expression and depicted in Figure 5.4 (left)

$$f(x) = \begin{cases} 1 & \text{if } x \in (i 2^{-m}, (i+1) 2^{-m}) \times (j 2^{-m}, (j+1) 2^{-m}) \text{ and } i+j \text{ odd} \\ -1 & \text{otherwise.} \end{cases}$$

We start with the same mesh  $\mathcal{T}_0$  with four elements as in Example 5.7, and construct recursively grids  $\mathcal{T}_{k+1} \in \mathbb{T}$ ,  $k \geq 0$ , as a conforming refinement of  $\mathcal{T}_k \in \mathbb{T}$  via two newest-vertex bisections of every triangle of  $\mathcal{T}_k$ ; see Figure 5.4 (right). Since  $f$  is  $L^2$ -orthogonal to every piecewise linear basis function of the space  $\mathbb{V}_{\mathcal{T}_k} = \mathbb{S}_{\mathcal{T}_k}^{1,0}$  for  $0 \leq k \leq m-1$ , we deduce that  $u_{\mathcal{T}_k} = 0$  and the energy error does not change

$$\|u - u_{\mathcal{T}_k}\|_{\Omega} = \|u - u_{\mathcal{T}_0}\|_{\Omega} \quad 0 \leq k \leq m-1. \quad (5.36)$$

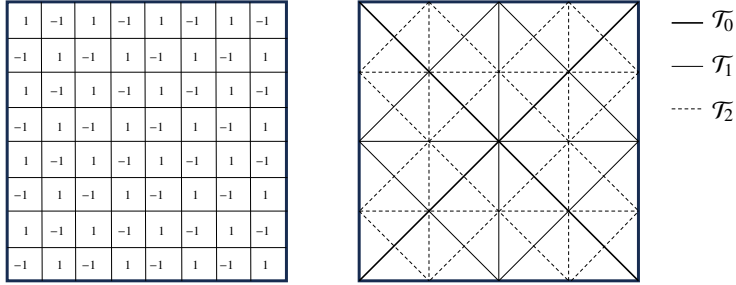


Figure 5.4. Representation of the checkerboard function  $f$  of Example 5.13 for  $m = 3$  (left), and grids  $\mathcal{T}_k$  for  $k = 0, 1, 2$  (right).

We see that this procedure creates three interior vertices in every triangle of  $\mathcal{T}_k$  after two refinement steps, namely in  $\mathcal{T}_{k+2}$  as long as  $k+2 \leq m$ . Since the error does not change, we conclude that the interior vertex property is necessary for error reduction but is not sufficient in the presence of data oscillation  $\text{osc}_{\mathcal{T}_k}(f)_{-1} \neq 0$ . We conclude

*Data oscillation  $\text{osc}_{\mathcal{T}}(f)_{-1}$  is not generally of higher order than the error, especially in the early stages of the adaptive process.* (5.37)

On the other hand, for  $k = m$  the discrete solution  $u_{\mathcal{T}_m}$  does no longer vanish globally, but is still zero along the lines where  $f$  changes sign due to the symmetry of the problem, and the same happens with  $u_{\mathcal{T}_{m+1}}$ . Therefore, the behavior of  $u_{\mathcal{T}_m}$  and  $u_{\mathcal{T}_{m+1}}$  in a fixed square, where  $f$  is constant, is exactly the same as in Example 5.7. This implies that  $u_{\mathcal{T}_m} = u_{\mathcal{T}_{m+1}}$ , and illustrates that the rather special situation of Example 5.7 can occur at any stage of the refinement process.



**Example 5.14 (vanishing of  $P_{\mathcal{T}}f$  for  $n = 1$ )** Since  $P_{\mathcal{T}}f$  is constructed locally upon testing  $f$  against cubic and quadratic bubbles (see Remark 4.26 (local computation)), and  $f$  of Example 5.13 is highly oscillatory, we realize that  $P_{\mathcal{T}_k}f$  is rather small relative to  $f$  in  $H^{-1}(\Omega)$ , but it is not zero. This is due to the lack of complete symmetry of the checkerboard pattern and the triangular grid. Suppose that each square of Fig. 5.4, where  $f = \pm 1$ , is further split across the diagonals into four triangles, and that  $f$  is assigned the alternating values  $\pm 1$  and  $\mp 1$  in each triangle depending on whether  $f$  was originally 1 or  $-1$  in that square; this configuration is displayed in Fig. 5.5. Suppose further that the coefficients  $(A, c)$  of the operator (2.5) are piecewise constant, as it happens for the Laplacian, the polynomial degree is  $n = 1$ , and the definition (4.39) of  $P_T$  over a triangle  $T \in \mathcal{T}$  uses  $q \in \mathbb{P}_0$  rather than  $\mathbb{P}_1$ . In light of (4.39) and (4.40), symmetry yields for all  $T \in \mathcal{T}$  and  $F \in \mathcal{F}$

$$\int_T f \phi_T = 0 \quad \Rightarrow \quad P_T f = 0, \quad \int_F f \phi_F = 0 \quad \Rightarrow \quad P_F f = 0, \quad (5.38)$$

whence  $P_{\mathcal{T}}f = 0$ . Since also  $u_{\mathcal{T}} = 0$  because  $f$  is orthogonal to all basis functions of  $\mathbb{V}_{\mathcal{T}}$ , we deduce  $\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f) = 0$  and all the information about the error  $\|u - u_{\mathcal{T}}\|_{\Omega} \neq 0$  resides in the data oscillation  $\text{osc}_{\mathcal{T}}(f)_{-1} \neq 0$ . Moreover, the fact that  $P_{\mathcal{T}}f = 0$  for several iterations reveals the important property that  $\text{osc}_{\mathcal{T}}(f)_{-1}$  may not change upon refinement because

$$\text{osc}_{\mathcal{T}}(f)_{-1}^2 = \sum_{T \in \mathcal{T}} \|f\|_{H^{-1}(\omega_T)}^2. \quad (5.39)$$

Since  $\|u - u_{\mathcal{T}}\|_{\Omega} \approx \text{osc}_{\mathcal{T}}(f)_{-1}$ , according to (4.45), a special care must be exercised to reduce data oscillation when it dominates. This justifies the structure of the algorithm one-step AFEM with switch below.

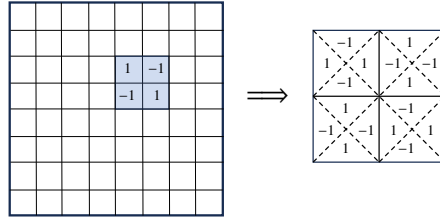


Figure 5.5. Refinement of the shaded according to the process described in Example 5.14.

**Example 5.15 (vanishing of  $P_{\mathcal{T}}f$  for  $n > 1$ )** Given  $n \geq 1$  a polynomial degree and  $\mathcal{T}_k$ ,  $k = 1, \dots, m$ , uniform refinements of  $\mathcal{T}_0$ , there are finitely many conditions to verify for  $f \in H^{-1}(\Omega)$  to be orthogonal to  $\mathbb{V}_{\mathcal{T}_k}$  and to  $\mathbb{F}_{\mathcal{T}_k}$ . Since  $\dim H^{-1}(\Omega) = \dim L^2(\Omega) = \infty$  there are infinitely many loads  $f \in H^{-1}(\Omega)$  as well as in  $L^2(\Omega)$

that yield  $u_{\mathcal{T}_k} = P_{\mathcal{T}_k} f = 0$ , which implies (5.36). Moreover,  $\eta_{\mathcal{T}_k}(u_k) = 0$  and  $\mathcal{E}_{\mathcal{T}_k}(u_{\mathcal{T}_k}) = \text{osc}_{\mathcal{T}_k}(f)_{-1}$  satisfies (5.39). One explicit example is as follows.

Given an initial mesh  $\mathcal{T}_0$ , suppose that  $f$  consists of line Dirac masses supported on the skeleton of  $\mathcal{T}_0$  with densities  $g_F$  on  $F \in \mathcal{F}_{\mathcal{T}_0}$  made of piecewise polynomials of degree  $2n + 1$ . We further assume that the  $g_F$  are orthogonal to  $\mathbb{P}_{2n}$  over  $F$  as well as over all sub-faces obtained from  $m \geq 1$  uniform refinements of  $\mathcal{T}_0$ ; see Fig. 5.4. In such a situation, (5.38) applies and  $u_{\mathcal{T}_k} = P_{\mathcal{T}_k} f = 0$ , whence (5.36) and (5.39) are valid for  $0 \leq k \leq m$ .

These three examples reveal the following crucial and novel feature about the interplay of the energy error  $\|u - u_{\mathcal{T}}\|_{\Omega}$  and data oscillation  $\text{osc}_{\mathcal{T}}(f)_{-1}$ :

*Data oscillation  $\text{osc}_{\mathcal{T}}(f)_{-1}$  may be responsible for the energy error  $\|u - u_{\mathcal{T}}\|_{\Omega}$  to stagnate, even with the interior vertex property, and may entirely dominate it relative to the error estimator  $\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f)$  over many mesh refinements unless it is reduced.* (5.40)

### 5.3.2 Reducing data oscillation

The PDE error estimator  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$  in (5.8) is fully discrete and thus computable. In contrast, the computation, or rather estimation, of  $\text{osc}_{\mathcal{T}}(f)_{-1}$  hinges on a priori knowledge of  $f$  and cannot be assessed in general. Assuming that the local indicators introduced in Lemma 4.8 (localization re-indexing)

$$\text{osc}_{\mathcal{T}}(f, T)_{-1} = \|f - P_{\mathcal{T}} f\|_{H^{-1}(\omega_T)} \quad T \in \mathcal{T}. \quad (5.41)$$

are computable without further regularity than  $f \in H^{-1}(\Omega)$ , it is natural to think of *tree approximation* as the algorithm of choice to reduce  $\text{osc}_{\mathcal{T}}(f)_{-1}$  [Binev and DeVore 2004, Binev, Fierro and Veeseer 2023, Binev 2018]. However, this optimal algorithm is not readily applicable because of the lack of a suitable sub-additivity property.

On the other hand, greedy algorithms, such as that in Section 3.6 (constructive approximation), do not work under minimal regularity. In Section 7.3, we present practical examples of rough  $f$  for which  $\text{osc}_{\mathcal{T}}(f)_{-1}$  can be replaced by a larger computable surrogate estimator  $\widehat{\text{osc}}_{\mathcal{T}}(f)_{-1}$ . The latter splits into element contributions and is amenable to a greedy strategy. Since this is specialized and technical, we prefer to postpone the full discussion to Section 7.3 and assume now the existence of a module DATA with the following property: given a tolerance  $\tau > 0$  and a conforming mesh  $\mathcal{T} \in \mathbb{T}$ , DATA constructs a conforming refinement  $\mathcal{T}_* \in \mathbb{T}$

$$[\mathcal{T}_*] = \text{DATA}(\mathcal{T}, f, \tau),$$

such that  $\text{osc}_{\mathcal{T}_*}(f)_{-1} \leq \tau$ . The complexity of DATA depends on the decay rate of the best approximation error  $\min_{\mathcal{T} \in \mathbb{T}_N} \widehat{\text{osc}}_{\mathcal{T}}(f)_{-1}$  of  $f$  with  $N$  degrees of freedom. We address this important issue in Section 7.3 for each example separately.

### 5.3.3 Linear convergence

The following algorithm, AFEM-SW, a one-step AFEM with switch, is a minor, but essential, modification of GALERKIN in that the call to the modules MARK and REFINE is conditional to the size of  $\text{osc}_{\mathcal{T}}(f)_{-1}$  relative to  $\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f)$ . This structure is consistent with the heuristic discussion in [Cascón *et al.* 2008, Section 6] to avoid separate marking. A similar algorithm is being developed in [Kreuzer *et al.* n.d.].

**Algorithm 5.16 (AFEM-SW: one-step AFEM with switch)** Let  $\mathcal{T}_0$  be a suitable initial mesh, the coefficients  $(A, c)$  be discrete over  $\mathcal{T}_0$ , and  $\varepsilon > 0$  be a stopping tolerance. Given parameters  $0 < \theta, \omega, \xi < 1$ , AFEM-SW iterates the following loop until  $\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f) \leq \varepsilon$ :

```

 $[\mathcal{T}, u_{\mathcal{T}}] = \text{AFEM-SW}(\mathcal{T}_0, \mathcal{D}, \varepsilon)$ 
  set  $j = 0$ 
  do
     $[u_{\mathcal{T}_j}] = \text{SOLVE}(\mathcal{T}_j)$ 
     $[\eta_{\mathcal{T}_j}(u_{\mathcal{T}_j}), \text{osc}_{\mathcal{T}_j}(f)_{-1}] = \text{ESTIMATE}(u_{\mathcal{T}_j}, \mathcal{T}_j, \mathcal{D})$ 
    if  $\mathcal{E}_{\mathcal{T}_j}(u_{\mathcal{T}_j}, f) \leq \varepsilon$ 
      return  $\mathcal{T}_j, u_{\mathcal{T}_j}$ 
    else if  $\text{osc}_{\mathcal{T}_j}(f)_{-1} \leq \sigma_j := \omega \mathcal{E}_{\mathcal{T}_j}(u_{\mathcal{T}_j}, f)$ 
       $[\mathcal{M}_j] = \text{MARK}(\{\eta_{\mathcal{T}_j}(u_{\mathcal{T}_j}, T)\}_{T \in \mathcal{T}_j}, \mathcal{T}_j, \theta)$ 
       $[\mathcal{T}_{j+1}] = \text{REFINE}(\mathcal{T}_j, \mathcal{M}_j)$ 
    else
       $[\mathcal{T}_{j+1}] = \text{DATA}(\mathcal{T}_j, f, \xi \sigma_j)$ 
     $j \leftarrow j + 1$ 
  while true

```

Note that SOLVE computes the Galerkin approximation using the exact right-hand side  $f \in H^{-1}(\Omega)$  (not necessarily in  $\mathbb{F}_{\mathcal{T}_j}$ ), thereby preserving the Galerkin orthogonality property. Moreover, ESTIMATE is now responsible for computing the PDE estimator

$$\eta_{\mathcal{T}_j}(u_{\mathcal{T}_j}) = \eta_{\mathcal{T}_j}(u_{\mathcal{T}_j}, f, \mathcal{T}_j)$$

using  $P_{\mathcal{T}_j} f \in \mathbb{F}_{\mathcal{T}_j}$ , as well as data oscillation  $\text{osc}_{\mathcal{T}_j}(f)_{-1}$ , which together give

$$\mathcal{E}_{\mathcal{T}_j}(u_{\mathcal{T}_j}, f) = \left( \eta_{\mathcal{T}_j}(u_{\mathcal{T}_j})^2 + \text{osc}_{\mathcal{T}_j}(f)_{-1}^2 \right)^{1/2},$$

and MARK consists of Dörfler marking (5.23) with parameter  $\theta$ .

We proceed as in Section 5.2.2 to prove linear convergence of AFEM-SW. We first show a contraction property for the *quasi-error*, which instead of (5.27) reads

$$\zeta_{\mathcal{T}_j}(u_{\mathcal{T}_j}, f)^2 := \|u - u_{\mathcal{T}_j}\|_{\Omega}^2 + \gamma \eta_{\mathcal{T}_j}(u_{\mathcal{T}_j})^2 + \text{osc}_{\mathcal{T}_j}(f)_{-1}^2, \quad (5.42)$$

where  $u = u(A, c, f)$  is the Galerkin solution with  $(A, c)$  discrete but  $f$  exact and the scaling parameter satisfies  $0 < \gamma \leq 1$ .

**Theorem 5.17 (contraction property of AFEM-SW)** *Let  $(A, c)$  be discrete coefficients over  $\mathcal{T}_0$  and let  $f \in H^{-1}(\Omega)$ . Let  $\theta \in (0, 1]$  be the Dörfler parameter and  $(\mathcal{T}_j, \mathbb{V}_j, u_j)$  be the sequence of conforming meshes  $\mathcal{T}_j$ , finite element spaces  $\mathbb{V}_j$ , and Galerkin solutions  $u_j \in \mathbb{V}_j$  produced by AFEM-SW. There exist parameters  $0 < \omega_0 < 1$  sufficiently small and  $0 < \gamma \leq 1$  and  $0 < \alpha < 1$  such that for any  $\omega \leq \omega_0$  and  $\xi \leq 1/2$  the quasi-error  $\zeta_{\mathcal{T}_j}$  in (5.42) contracts*

$$\zeta_{\mathcal{T}_{j+1}}(u_{\mathcal{T}_{j+1}}, f) \leq \alpha \zeta_{\mathcal{T}_j}(u_{\mathcal{T}_j}, f) \quad j \geq 0. \quad (5.43)$$

*Proof.* We argue as in Theorem 5.8 (general contraction property) upon distinguishing the two possible cases within Algorithm 5.16. But first, we must account for a crucial difference: the discrete forcing function  $P_{\mathcal{T}_j} f$  used in the definition of the estimator  $\mathcal{E}_{\mathcal{T}_j}(u_{\mathcal{T}_j}, f)$  changes in each iteration. We use the same notation as in Theorem 5.8 along with  $\text{osc}_j := \text{osc}_{\mathcal{T}_j}(f)_{-1}$ ,  $\mathcal{E}_j^2 := \eta_j^2 + \text{osc}_j^2$ , and  $P_j := P_{\mathcal{T}_j}$ .

□ *Estimator reduction property.* In view of Proposition 4.56 (estimator reduction), we need to estimate the discrepancy between discrete forcing functions

$$\sum_{T \in \mathcal{T}_{j+1}} \|P_{j+1}f - P_j f\|_{H^{-1}(\omega_T)}^2 \leq 2 \sum_{T \in \mathcal{T}_{j+1}} \left( \|f - P_{j+1}f\|_{H^{-1}(\omega_T)}^2 + \|f - P_j f\|_{H^{-1}(\omega_T)}^2 \right).$$

For the first term we recall Lemma 4.57 (quasi-monotonicity of oscillation) to write

$$\sum_{T \in \mathcal{T}_{j+1}} \|f - P_{j+1}f\|_{H^{-1}(\omega_T)}^2 = \text{osc}_{j+1}^2 \leq C_{\text{osc}}^2 \text{osc}_j^2.$$

For the second term, instead, we combine the projection property  $P_{j+1}(P_j f) = P_j f$  with Lemma 4.5 (localization of  $H^{-1}$ -norm) and Corollary 4.31 (local near-best approximation), and the fact that  $\mathcal{T}_{j+1}$  is a refinement of  $\mathcal{T}_j$ , to see that

$$\begin{aligned} \sum_{T' \subset \omega_T} \|f - P_{j+1}(P_j f)\|_{H^{-1}(\omega_{T'})}^2 &\leq C_{\text{ISb}}^2 \sum_{T' \subset \omega_T} \|f - P_j f\|_{H^{-1}(\omega_{T'})_{-1}}^2 \\ &\leq C_{\text{ISb}}^2 C_{\text{ovrl}}^2 \text{osc}_{\mathcal{T}_j}(f, \omega_T)^2 \quad \forall T \in \mathcal{T}_j. \end{aligned}$$

Adding over  $T$  and recalling Proposition 4.56 we end up with

$$\begin{aligned} \eta_{\mathcal{T}_{j+1}}(u_{j+1}, f, \mathcal{T}_{j+1})^2 &\leq (1 + \delta) (\eta_{\mathcal{T}_j}(u_j, f, \mathcal{T}_j)^2 - \lambda \eta_{\mathcal{T}_j}(u_j, f, \mathcal{M}_j)^2) \\ &\quad + (1 + \delta^{-1}) C_{\text{Lip}}^2 (|u_j - u_{j+1}|_{H_0^1(\Omega)}^2 + \text{osc}_j^2), \end{aligned} \quad (5.44)$$

for a constant  $C_{\text{Lip}}$  large enough to absorb all preceding constants, and any  $\delta > 0$ .

□ *Case  $\text{osc}_j \leq \omega \mathcal{E}_j$ .* We first observe that  $\eta_j^2 \geq (1 - \omega^2) \mathcal{E}_j^2$  and  $\text{osc}_j^2 \leq \frac{\omega^2}{1 - \omega^2} \eta_j^2$ . We then proceed as in Theorem 5.8 with the quantity  $e_j^2 + \gamma \eta_j^2$ , and observe that

the choices of  $\delta$  and  $\gamma$

$$\delta \leq -1 + \frac{1 - \frac{\lambda\theta^2}{2}}{1 - \lambda\theta^2} = \frac{\lambda\theta^2}{2(1 - \lambda\theta^2)}, \quad \gamma \leq \frac{\delta}{4C_{\text{Lip}}^2} \leq \frac{1}{2(1 + \delta^{-1})C_{\text{Lip}}^2}, \quad (5.45)$$

imply  $\gamma(1 + \delta^{-1})C_{\text{Lip}}^2 \leq 1/2$ . This, together with (5.7) and (5.44), leads to

$$e_{j+1}^2 + \gamma\eta_{j+1}^2 \leq e_j^2 + \gamma\left(1 - \frac{\lambda\theta^2}{2}\right)\eta_j^2 + \frac{1}{2}\text{osc}_j^2;$$

compare with (5.32). We invoke the upper bound in (5.13) to write

$$\eta_j^2 \geq (1 - \omega^2)\mathcal{E}_j^2 \geq (1 - \omega^2)\frac{e_j^2}{C_1^2} \geq \frac{e_j^2}{2C_1^2}$$

provided  $\omega^2 \leq 1/2$ , whence

$$e_{j+1}^2 + \gamma\eta_{j+1}^2 \leq \left(1 - \frac{\gamma\lambda\theta^2}{8C_1^2}\right)e_j^2 + \gamma\left(1 - \frac{\lambda\theta^2}{8}\right)\eta_j^2 - \frac{\gamma\lambda\theta^2}{8}\eta_j^2 + \frac{1}{2}\text{osc}_j^2.$$

We next consider the data oscillation, for which we invoke Lemma 4.57 (quasi-monotonicity of oscillation)

$$\text{osc}_{j+1} \leq C_{\text{osc}} \text{osc}_j, \quad \text{osc}_j^2 \leq C_{\text{osc}}^2 \frac{\omega^2}{1 - \omega^2} \eta_j^2 \leq 2C_{\text{osc}}^2 \omega^2 \eta_j^2.$$

Adding the two preceding inequalities yields

$$\begin{aligned} \zeta_{j+1}^2 &= e_{j+1}^2 + \gamma\eta_{j+1}^2 + \text{osc}_{j+1}^2 \leq \left(1 - \frac{\gamma\lambda\theta^2}{8C_1^2}\right)e_j^2 \\ &\quad + \left(1 - \frac{\lambda\theta^2}{8}\right)(\gamma\eta_j^2 + \text{osc}_j^2) \\ &\quad + \left[-\gamma\frac{\lambda\theta^2}{8} + 2(C_{\text{osc}}^2 - 1 + \frac{\lambda\theta^2}{8} + \frac{1}{2})\omega^2\right]\eta_j^2. \end{aligned}$$

We drop the term  $-\frac{1}{2} + \frac{\lambda\theta^2}{8} \leq 0$  and let  $\gamma = \frac{\delta}{4C_{\text{Lip}}^2}$ , which is consistent with (5.45).

We seek conditions on  $\omega$  that make the factor of  $\eta_j^2$  non-positive. Imposing

$$\omega^2 \leq \frac{\gamma\lambda\theta^2}{16C_{\text{osc}}^2} = \frac{\lambda\theta^2}{64C_{\text{osc}}^2C_{\text{Lip}}^2}\delta \quad (5.46)$$

yields

$$\zeta_{j+1}^2 \leq \alpha_1^2 \zeta_j^2$$

with

$$\alpha_1^2 := \max \left\{ 1 - \frac{\delta\lambda\theta^2}{32C_1^2C_{\text{Lip}}^2}, 1 - \frac{\lambda\theta^2}{8} \right\} < 1.$$

□<sup>3</sup> *Case*  $\text{osc}_j > \omega \mathcal{E}_j$ . The module DATA with input parameter  $\xi \leq 1/2$  gives

$$\text{osc}_{j+1} \leq \xi \omega \mathcal{E}_j < \xi \text{osc}_j.$$

We now exploit the contraction of  $\text{osc}_j$  to compensate the moderate increase of  $\eta_j^2$  and presence of  $\text{osc}_j^2$ , both governed by (5.44). In fact,  $\gamma(1 + \delta^{-1})C_{\text{Lip}}^2 \leq \frac{1}{2}$  yields

$$e_{j+1}^2 + \gamma \eta_{j+1}^2 \leq e_j^2 + \gamma(1 + \delta) \eta_j^2 + \frac{1}{2} \text{osc}_j^2.$$

We add  $\text{osc}_{j+1}^2$  to both sides and rewrite the right-hand side to arrive at

$$\begin{aligned} \zeta_{j+1}^2 = e_{j+1}^2 + \gamma \eta_{j+1}^2 + \text{osc}_{j+1}^2 &\leq e_j^2 - \frac{1 - 2\xi^2}{8} \text{osc}_j^2 \\ &\quad + (1 - \delta) \gamma \eta_j^2 + \left( \frac{1 + 2\xi^2}{4} + \frac{1}{2} \right) \text{osc}_j^2 \\ &\quad + 2\delta \gamma \eta_j^2 - \frac{1 - 2\xi^2}{8} \text{osc}_j^2. \end{aligned}$$

Our next task is to find conditions on  $\omega$  for the last line to be non-positive. To this end, we resort to the upper bound  $\eta_j^2 < \omega^{-2} \text{osc}_j^2$  and  $\xi \leq 1/2$  to obtain

$$2\delta \gamma \eta_j^2 - \frac{1 - 2\xi^2}{8} \text{osc}_j^2 < \left( \frac{2\delta \gamma}{\omega^2} - \frac{1}{16} \right) \text{osc}_j^2 \leq 0$$

provided we impose the relation

$$\omega^2 \geq 32\delta \gamma = \frac{8}{C_{\text{Lip}}^2} \delta^2. \quad (5.47)$$

We next use the upper bound  $e_j \leq C_1 \mathcal{E}_j \leq C_1 \omega^{-1} \text{osc}_j$  to write

$$e_j^2 - \frac{1 - 2\xi^2}{8} \text{osc}_j^2 \leq \left( 1 - \frac{\omega^2}{16C_1^2} \right) e_j^2,$$

whence we end up with

$$\zeta_{j+1}^2 \leq \alpha_2^2 \zeta_j^2$$

provided we define

$$\alpha_2 := \max \left\{ 1 - \frac{\omega^2}{16C_1^2}, 1 - \delta, \frac{3 + 2\xi^2}{4} \right\} < 1.$$

□<sup>4</sup> *Choosing the parameters.* We see that the asserted estimate (5.43) is valid with  $\alpha = \max\{\alpha_1, \alpha_2\} < 1$  provided the constraints (5.46) and (5.47) are compatible, i.e.,

$$\frac{8}{C_{\text{Lip}}^2} \delta^2 \leq \omega^2 \leq \frac{\lambda \theta^2}{64 C_{\text{osc}}^2 C_{\text{Lip}}^2} \delta.$$

We choose  $\delta_0 = \frac{\lambda\theta^2}{512C_{\text{osc}}^2C_{\text{Lip}}}$  and  $\omega_0 = \frac{\lambda\theta^2}{128C_{\text{osc}}^2C_{\text{Lip}}\sqrt{2C_{\text{Lip}}}}$ . Then, for all  $\omega \leq \omega_0$ , there exists  $\delta \leq \delta_0$  that satisfies the previous inequalities as well as  $\gamma = \frac{\delta}{4C_{\text{Lip}}^2} \leq 1$ , perhaps upon reducing  $\delta_0$ . This completes the proof of Theorem 5.43.  $\square$

Note that we could replace the conditional  $\text{osc}_{\mathcal{T}_j}(f)_{-1} \leq \omega \mathcal{E}_{\mathcal{T}_j}(u_{\mathcal{T}_j}, f)$  by  $\text{osc}_{\mathcal{T}_j}(f)_{-1} \leq \omega \eta_{\mathcal{T}_j}(u_{\mathcal{T}_j})$ , but the tolerance  $\tau$  of DATA cannot be

$$\tau = \xi \omega \eta_{\mathcal{T}_j}(u_{\mathcal{T}_j})$$

because the algorithm may not terminate when  $\eta_{\mathcal{T}_j}(u_{\mathcal{T}_j}) = 0$ , see e.g. Examples 5.13 - 5.15. In fact, the tolerance  $\tau = \xi \omega \mathcal{E}_{\mathcal{T}_j}(u_{\mathcal{T}_j}, f)$  is *dynamic* and relative to  $\mathcal{E}_{\mathcal{T}_j}(u_{\mathcal{T}_j}, f)$ . This avoids separate marking, which was shown in [Cascón *et al.* 2008, Section 6] to give non-optimal convergence rates. In contrast, we will prove in Section 6 that Algorithm 5.16 is rate optimal.

It turns out that Theorem 5.17 yields linear convergence of error and estimator.

**Corollary 5.18 (linear convergence of error)** *For  $0 < \alpha < 1$  and  $0 < \omega \leq \omega_0$ ,  $\xi \leq \frac{1}{2}$  as in Theorem 5.17, and  $C_* = (1 + C_2^{-1})^{1/2}$  with  $C_2$  as in (5.13), there holds*

$$|u - u_{\mathcal{T}_k}|_{H_0^1(\Omega)} \leq C_* \alpha^{k-j} |u - u_{\mathcal{T}_j}|_{H_0^1(\Omega)} \quad \forall k \geq j \geq 0.$$

*Proof.* We again use the same notation as in Lemma 5.5 and Theorem 5.17. In view of the definition (5.42) of quasi-error  $\zeta_j := \zeta_{\mathcal{T}_j}(u_{\mathcal{T}_j}, f)$ , we thus have  $e_j \leq \zeta_j$  and

$$\zeta_j^2 \leq e_j^2 + \eta_j^2 + \text{osc}_j^2 \leq (1 + C_2^{-1}) e_j^2$$

because  $C_2 \mathcal{E}_j \leq e_j$  from (5.13). This implies

$$e_j \leq \zeta_j \leq C_* e_j \quad \forall j \geq 0,$$

and invoking Theorem 5.17 (contraction property for AFEM-SW)

$$e_k^2 \leq \zeta_k^2 \leq \alpha^{2(k-j)} \zeta_j^2 \leq \alpha^{2(k-j)} C_*^2 e_j$$

gives the desired estimate.  $\square$

We stress that Corollary 5.18 relies on the lower bound in (5.13) whereas Corollary 5.19 uses only the upper bound. Its proof is similar and thus omitted.

**Corollary 5.19 (linear convergence of estimator)** *For  $0 < \alpha < 1$  and  $0 < \omega \leq \omega_0$ ,  $\xi \leq \frac{1}{2}$  as in Theorem 5.17, and  $C_\# = ((1 + C_1^2)\gamma^{-1})^{1/2}$  with  $C_1$  as in (5.13), there holds*

$$\mathcal{E}_{\mathcal{T}_k}(u_{\mathcal{T}_k}, f) \leq C_\# \alpha^{k-j} \mathcal{E}_{\mathcal{T}_j}(u_{\mathcal{T}_j}, f) \quad \forall k \geq j \geq 0.$$

#### 5.4 Convergence for general data: two-step AFEM

We now remove the restriction of Sections 5.2 and 5.3 of discrete data and allow for general data  $\mathcal{D} = (A, c, f) \in \mathbb{D}$  as defined in (5.2). The current goal is to study

Algorithm 5.1 (AFEM-TS), which concatenates the modules DATA and GALERKIN. We start with the study of continuous dependence with respect to data  $\mathcal{D}$ . We next discuss the approximation of  $\mathcal{D}$  within the module DATA, the computational cost of GALERKIN, and eventually the convergence of Algorithm 5.1.

#### 5.4.1 Perturbation theory

We start with a brief discussion of data perturbation. Given constants  $0 < \alpha_1 \leq \alpha_2$  and  $0 \leq c_1 \leq c_2$ , we define the constrained spaces for the diffusion and reaction coefficients by

$$M(\alpha_1, \alpha_2) := \left\{ \mathbf{A} \in L^\infty(\Omega; \mathbb{R}_{\text{sym}}^{d \times d}) : 0 < \alpha_1 \leq \lambda_j(\mathbf{A}(x)) \leq \alpha_2 \right. \\ \left. \text{for a.e. } x \in \Omega, 1 \leq j \leq d \right\}, \quad (5.48)$$

where  $\lambda_j(\mathbf{A}(x))$  denotes the  $j$ -th eigenvalue of  $\mathbf{A}$  at  $x \in \Omega$  and

$$R(c_1, c_2) := \left\{ c \in L^\infty(\Omega) : c_1 \leq c(x) \leq c_2 \text{ for a.e. } x \in \Omega \right\}. \quad (5.49)$$

The coefficients  $(\mathbf{A}, c)$  are assumed to satisfy the structural assumption

$$\mathbf{A} \in M(\alpha_1, \alpha_2), \quad c \in R(c_1, c_2); \quad (5.50)$$

see (2.6). This guarantees coercivity and continuity of the bilinear form  $\mathcal{B}$  in (2.8), and thus unique solvability of (2.7).

Regarding the discrete coefficients,  $(\widehat{\mathbf{A}}, \widehat{c})$  will ultimately be piecewise polynomials in a grid  $\widehat{\mathcal{T}} \in \mathbb{T}$ . The side constraints in (5.48) and (5.49) are generally violated by any linear projection onto piecewise polynomials of degree  $n - 1 \geq 1$ , e.g. the  $L^2$ -projection, and require a nonlinear correction maintaining high order accuracy. This is a crucial but delicate matter addressed later in Section 7.2. For the moment, we simply assume that the discrete coefficients  $(\widehat{\mathbf{A}}, \widehat{c})$  satisfy

$$\widehat{\mathbf{A}} \in M(\widehat{\alpha}_1, \widehat{\alpha}_2), \quad \widehat{c} \in R(\widehat{c}_1, \widehat{c}_2), \quad (5.51)$$

with

$$\frac{\alpha_1}{2} \leq \widehat{\alpha}_1 \leq \widehat{\alpha}_2 \leq C_{\text{ctr}} \alpha_2, \quad -\frac{\alpha_1}{4C_P^2} \leq \widehat{c}_1 \leq \widehat{c}_2 \leq C_{\text{ctr}}(\alpha_1 + c_2) \quad (5.52)$$

where  $C_P > 0$  is the Poincaré constant in (2.2) and  $C_{\text{ctr}} \geq 1$  is a constant; see (7.21) and (7.23). This implies coercivity and continuity of the perturbed bilinear form

$$\widehat{\mathcal{B}}[v, w] := \int_{\Omega} \nabla v \cdot \widehat{\mathbf{A}} \nabla w + \widehat{c} v w, \quad \forall v, w \in H_0^1(\Omega), \quad (5.53)$$

because for all  $v, w \in H_0^1(\Omega)$

$$\widehat{\mathcal{B}}[v, v] \geq \widehat{\alpha}_1 \int_{\Omega} |\nabla v|^2 - \frac{\alpha_1}{4C_P^2} \int_{\Omega} |v|^2 \geq \frac{\alpha_1}{4} |v|_{H_0^1(\Omega)}^2$$



and

$$|\widehat{\mathcal{B}}[v, w]| \leq \int_{\Omega} \widehat{\alpha}_2 |\nabla v| |\nabla w| + \widehat{c}_2 |v| |w| \leq (\widehat{\alpha}_2 + \widehat{c}_2 C_P^2) |v|_{H_0^1(\Omega)} |w|_{H_0^1(\Omega)}.$$

Therefore, the energy norm  $\|v\|_{\Omega}^2 = \widehat{\mathcal{B}}[v, v]$  is equivalent to the  $H_0^1$ -seminorm

$$c_{\widehat{\mathcal{B}}} |v|_{H_0^1(\Omega)}^2 \leq \|v\|_{\Omega}^2 \leq C_{\widehat{\mathcal{B}}} |v|_{H_0^1(\Omega)}^2, \quad (5.54)$$

where  $c_{\widehat{\mathcal{B}}} = \frac{\alpha_1}{4}$  and  $C_{\widehat{\mathcal{B}}} = \widehat{\alpha}_2 + \widehat{c}_2 C_P^2$ . Hence, the Lax-Milgram Theorem guarantees the existence of a unique solution  $\widehat{u} = u(\widehat{\mathcal{D}}) \in H_0^1(\Omega)$  of the perturbed problem (5.5) defined using the discrete data  $\widehat{\mathcal{D}} = (\widehat{A}, \widehat{c}, \widehat{f})$ .

We now quantify the effect of perturbing data from  $\mathcal{D}$  to  $\widehat{\mathcal{D}}$  in the space

$$\widehat{D}(\Omega) := L^r(\Omega; \mathbb{R}^{d \times d}) \times W_q^{-s}(\Omega) \times H^{-1}(\Omega), \quad (5.55)$$

where  $2 \leq r \leq \infty$  and  $0 \leq s \leq 1$ ,  $\frac{d}{2-s} < q \leq \infty$ ;  $W_q^{-s}(\Omega)$  is the dual of  $W_{q^*}^s(\Omega)$  with  $q^* = \frac{q}{q-1}$ . The use of  $r = \infty$  for  $A$  entails the further assumption

$$A \text{ is piecewise uniformly continuous over a generic mesh } \mathcal{T} \in \mathbb{T}, \quad (5.56)$$

which turns out to be rather restrictive but customary in the theory of AFEM. Our present approach allows for  $r < \infty$  and thus for discontinuous coefficients  $(A, c)$  non aligned with  $\mathcal{T}$ , which is important in practice. However, it requires the following slightly stronger regularity property of the solution  $u \in H_0^1(\Omega)$  of (2.7):

$$\|\nabla u\|_{L^p(\Omega)} \leq C_p \|f\|_{W_p^{-1}(\Omega)} \quad 2 < p \leq p_0. \quad (5.57)$$

We refer to Lemma 2.13 ( $W_p^1$ -regularity) that shows the existence of  $C_p > 0$  and  $p_0 > 2$  that only depend on  $\Omega, \alpha_1, \alpha_2$  and  $c_2$ .

**Lemma 5.20 (continuous dependence on data)** *Let  $\mathcal{D} = (A, c, f) \in \mathbb{D}$  be such that  $A \in M(\alpha_1, \alpha_2)$  and  $c \in R(c_1, c_2)$ . Let  $\widehat{\mathcal{D}} = (\widehat{A}, \widehat{c}, \widehat{f}) \in \mathbb{D}$  be an approximation of  $\mathcal{D}$  such that  $\widehat{A} \in M(\widehat{\alpha}_1, \widehat{\alpha}_2)$  and  $\widehat{c} \in R(\widehat{c}_1, \widehat{c}_2)$ . Let  $2 \leq r \leq \infty$ ,  $2 \leq r_* = \frac{2r}{r-2} \leq p_0$  be so that  $f \in W_{r_*}^{-1}(\Omega)$ . If  $u = u(\mathcal{D}), \widehat{u} = u(\widehat{\mathcal{D}}) \in H_0^1(\Omega)$  are the solutions of (2.7) and (5.5) with data  $\mathcal{D}, \widehat{\mathcal{D}}$ , respectively, and  $u$  satisfies (5.57) with  $p = r_*$  in case  $r < \infty$ , then for any  $0 \leq s \leq 1$  and  $\frac{d}{2-s} < q \leq \infty$  there holds*

$$\|\nabla(u - \widehat{u})\|_{L^2(\Omega)} \leq C(\mathcal{D}, \Omega) \|\mathcal{D} - \widehat{\mathcal{D}}\|_{\widehat{D}(\Omega)}, \quad (5.58)$$

where the constant  $C(\mathcal{D}, \Omega)$  depends on  $\mathcal{D}, \Omega, p_0, q$  and  $s$ , and blows up as  $q \rightarrow \frac{d}{2-s}$  for  $d = 2$  while it remains bounded for  $d > 2$ .

*Proof.* Subtracting the weak formulations (2.7) for  $u$  and (5.5) for  $\widehat{u}$ , and reordering, we easily obtain for any  $v \in H_0^1(\Omega)$

$$\int_{\Omega} \nabla v \cdot \widehat{A} \nabla(u - \widehat{u}) + \widehat{c} v(u - \widehat{u}) = \int_{\Omega} \nabla v \cdot (\widehat{A} - A) \nabla u + (\widehat{c} - c) v u + \langle f - \widehat{f}, v \rangle.$$

We choose  $v = u - \widehat{u} \in H_0^1(\Omega)$  and invoke (5.54) to deduce

$$c_{\widehat{B}} \|\nabla v\|_{L^2(\Omega)}^2 \leq \int_{\Omega} \nabla v \cdot (\widehat{A} - A) \nabla u + (\widehat{c} - c)vu + \langle f - \widehat{f}, v \rangle.$$

We estimate each term separately, starting with the first and last terms

$$\int_{\Omega} \nabla v \cdot (\widehat{A} - A) \nabla u \leq \|\widehat{A} - A\|_{L^r(\Omega)} \|\nabla u\|_{L^{r^*}(\Omega)} \|\nabla v\|_{L^2(\Omega)} \quad (5.59)$$

with  $2 \leq r_* = \frac{2r}{r-2} \leq p_0$ , as well as

$$\langle f - \widehat{f}, v \rangle \leq \|f - \widehat{f}\|_{H^{-1}(\Omega)} \|\nabla v\|_{H^1(\Omega)}.$$

For the reaction term, which is more delicate, we invoke the duality pairing  $W_q^s - W_q^{-s}$  for any  $0 \leq s \leq 1$  and  $q' = \frac{q}{q-1} \geq 1$  to obtain

$$\int_{\Omega} (\widehat{c} - c)vu \leq \|\widehat{c} - c\|_{W_q^{-s}(\Omega)} |vu|_{W_q^s(\Omega)}.$$

We now estimate  $|vu|_{W_q^s(\Omega)} \lesssim |vu|_{W_{p'}^1(\Omega)}$  where  $1/p' = \min\{1, (1-s)/d + 1/q'\}$  guarantees that  $W_{p'}^1(\Omega) \subset W_{q'}^s(\Omega)$  [Leoni 2009, Theorem 14.32]. Recalling that  $q > d/(2-s)$ , we deduce

$$\frac{1-s}{d} + \frac{1}{q'} = \frac{1-s}{d} + 1 - \frac{1}{q} > \frac{d-1}{d} \geq \frac{1}{2},$$

whence  $1/p' > 1/2$  and there exists  $t < \infty$  satisfying  $1/t + 1/2 = 1/p'$  and

$$|vu|_{W_q^s(\Omega)} \lesssim \|\nabla v\|_{L^2(\Omega)} \|u\|_{L^t(\Omega)} + \|v\|_{L^t(\Omega)} \|\nabla u\|_{L^2(\Omega)}.$$

Using the definition of  $p'$ , we obtain the explicit expression  $t = \max\{2, t_0\}$ , where

$$t_0 = \frac{2dq}{q(2(1-s) + d) - 2d}.$$

Moreover, for the Sobolev embedding  $H^1(\Omega) \hookrightarrow L^t(\Omega)$  we require

$$1 - d\left(\frac{1}{2} - \frac{1}{t}\right) > 0 \quad \Rightarrow \quad q > \frac{d}{2-s},$$

which is our assumption on  $q$ . Therefore, as  $q \rightarrow \frac{d}{2-s}$ , we see that  $t_0 \rightarrow \frac{2d}{d-2}$  and the limit is infinite for  $d = 2$  but finite and larger than 2 for  $d > 2$ . Sobolev embedding together with the first Poincaré inequality (2.2) gives the estimate

$$\|vu\|_{W_{q'}^s(\Omega)} \leq C(\Omega, t) \|\nabla v\|_{L^2(\Omega)} \|\nabla u\|_{L^2(\Omega)},$$

where  $C(\Omega, t)$  is proportional to  $t$  for  $d = 2$ .

We finally observe that the factors  $\|\nabla u\|_{L^{r^*}(\Omega)}$  and  $\|\nabla u\|_{L^2(\Omega)}$  appear in the estimates of the coefficients  $A$  and  $c$ , thereby reflecting the multiplicative nature of these terms. Since  $2 \leq r_* \leq p_0$ , they can be further bounded in terms of  $\|f\|_{W_{r_*}^{-1}(\Omega)}$

according to (5.57). This in conjunction with the preceding estimates yields the assertion (5.58).  $\square$

A natural and rather popular choice of parameters  $(r, q, s)$  in Lemma 5.20 (continuous dependence on data) is  $r = q = \infty$  and  $s = 0$ , but this would prevent the coefficients  $(A, c)$  from being discontinuous within elements; see (5.56). We will explore this matter further in Section 7 (data approximation).

**Remark 5.21 ( $L^2$ -approximation of  $A$ )** It is appealing to estimate the distortion  $A - \widehat{A}$  in  $L^2(\Omega)$  rather than in  $L^r(\Omega)$  because it is a simpler norm to deal with. Since  $\|A\|_{L^\infty(\Omega)} \leq \alpha_2$ ,  $\|\widehat{A}\|_{L^\infty(\Omega)} \leq \widehat{\alpha}_2$ , and  $2 \leq r \leq \infty$ , we deduce

$$\|A - \widehat{A}\|_{L^r(\Omega)} \leq \|A - \widehat{A}\|_{L^\infty(\Omega)}^{1-\frac{2}{r}} \|A - \widehat{A}\|_{L^2(\Omega)}^{\frac{2}{r}} \lesssim \|A - \widehat{A}\|_{L^2(\Omega)}^{\frac{2}{r}}.$$

However, this may be sub-optimal in general. One important situation where this is sharp corresponds to  $A$  being piecewise constant with jump discontinuities across a Lipschitz hypersurface  $\gamma$  and  $\widehat{A} = A$  on every element  $T \in \mathcal{T}$  not intersecting  $\gamma$ . In that case, the equivalence

$$\|A - \widehat{A}\|_{L^p(\Omega)} \approx |\{x \in \Omega : A(x) \neq \widehat{A}(x)\}|^{\frac{1}{p}},$$

is valid for  $1 \leq p \leq \infty$ , whence

$$\|A - \widehat{A}\|_{L^r(\Omega)} \approx \|A - \widehat{A}\|_{L^2(\Omega)}^{\frac{2}{r}}.$$

#### 5.4.2 Approximation of $\mathcal{D}$ : module DATA

In this section we briefly discuss the structure of DATA, which is the module of Algorithm 5.1 (two-step AFEM) responsible for data approximation.

In the sequel we will no longer rely on the Banach space  $\widehat{D}(\Omega)$  defined in (5.55) and used in Lemma 5.20 (continuous dependence on data). We rather restrict the error notion to the following stronger Banach space

$$D(\Omega) := L^r(\Omega; \mathbb{R}^{d \times d}) \times L^q(\Omega) \times H^{-1}(\Omega), \quad (5.60)$$

where  $q = 2$  for  $d < 4$  or  $q > \frac{d}{2}$  for  $d \geq 4$ ; we justify the choice of  $q$  below. Let  $\mathbb{D}$  and  $\mathbb{D}_{\mathcal{T}}$  be the spaces defined in (5.1) and (5.2) for a conforming mesh  $\mathcal{T} \in \mathbb{T}$ . Given  $\mathcal{D} = (A, c, f) \in \mathbb{D}$ , let  $\delta_{\mathcal{T}}(\mathcal{D})$  be the best approximation error of  $\mathcal{D}$  within  $\mathbb{D}_{\mathcal{T}}$  measured in the space  $D(\Omega)$ , namely

$$\delta_{\mathcal{T}}(\mathcal{D}) := \inf_{\mathcal{D}_{\mathcal{T}} \in \mathbb{D}_{\mathcal{T}}} \|\mathcal{D} - \mathcal{D}_{\mathcal{T}}\|_{D(\Omega)}. \quad (5.61)$$

This quantity characterizes the approximation quality of  $\mathbb{D}_{\mathcal{T}}$ , thereby having theoretical value. Since  $\delta_{\mathcal{T}}(\mathcal{D})$  is hard to access in view of the norms involved in the definition of  $D(\Omega)$ , the module DATA computes the surrogate quantity

$$\text{osc}_{\mathcal{T}}(\mathcal{D}) := \|\mathcal{D} - \widehat{\mathcal{D}}\|_{D(\Omega)}$$

for some approximation  $\widehat{\mathcal{D}} \in \mathbb{D}_{\mathcal{T}}$  to be specified below.

**Assumption 5.22 (properties of DATA)** Given a conforming mesh  $\mathcal{T} \in \mathbb{T}$  and a tolerance  $\tau > 0$ , the call

$$[\widehat{\mathcal{T}}, \widehat{\mathcal{D}}] = \text{DATA}(\mathcal{T}, \mathcal{D}, \tau) \quad (5.62)$$

creates an admissible refinement  $\widehat{\mathcal{T}}$  of  $\mathcal{T}$  and discrete data  $\widehat{\mathcal{D}} = \mathcal{D}_{\widehat{\mathcal{T}}} \in \mathbb{D}_{\widehat{\mathcal{T}}}$  such that for a constant  $C_{\text{data}}$

$$\text{osc}_{\widehat{\mathcal{T}}}(\mathcal{D}) := \|\mathcal{D} - \widehat{\mathcal{D}}\|_{D(\Omega)} \leq C_{\text{data}} \tau \quad (5.63)$$

as well as the structural conditions (5.51) are achieved in a finite number of iterations that depends on the regularity of  $\mathcal{D}$ , and such that

$$\text{osc}_{\widehat{\mathcal{T}}}(\mathcal{D}) \leq \Lambda_{\text{data}} \delta_{\widehat{\mathcal{T}}}(\mathcal{D}) \quad (5.64)$$

with  $\Lambda_{\text{data}} \geq 1$  depending only on the shape regularity of  $\mathbb{T}$ , the polynomial degree  $n$  and the Lebesgue exponents in the space  $D(\Omega)$ .

In view of Lemma 5.20 (continuous dependence on data), there exists a constant  $C_D > 0$  depending on  $\mathcal{D}, \Omega$ , and the shape regularity of  $\mathbb{T}$ , such that the exact solutions  $u = u(\mathcal{D})$  and  $\widehat{u} = u(\widehat{\mathcal{D}})$  of (2.5) and (5.5), corresponding to data  $\mathcal{D}$  and  $\widehat{\mathcal{D}}$  respectively, satisfy the error estimate

$$|u - \widehat{u}|_{H_0^1(\Omega)} \leq C_D \tau. \quad (5.65)$$

A brief discussion follows about computing  $\text{osc}_{\widehat{\mathcal{T}}}(\mathcal{D})$ , where  $\widehat{\mathcal{T}}$  remains fixed and is replaced by  $\mathcal{T}$  to simplify the notation. Specific details are given later in Assumptions 6.10 and 6.11 of Section 6.10 and especially in Section 7.

*Approximating the coefficients.* We now construct approximations  $(\widehat{A}, \widehat{c})$  using local  $L^2$ -projections and emphasize that this does not enforce the side constraints in the structural assumption (5.51). We propose in Section 7 a nonlinear correction satisfying the side constraints without sacrificing the accuracy.

Given  $T \in \mathcal{T}$ , and  $v \in L^p(T)$  with  $1 \leq p \leq \infty$ , we denote by  $\Pi_T v := \Pi_T^{n-1} v$  the  $L^2$ -projection of  $v$  onto the space  $\mathbb{P}_{n-1}$  of polynomials of degree  $\leq n-1$ , namely

$$\Pi_T v \in \mathbb{P}_{n-1} : \quad \int_T \Pi_T v w = \int_T v w \quad \forall w \in \mathbb{P}_{n-1}. \quad (5.66)$$

**Lemma 5.23 ( $L^p$ -stability of  $\Pi_T$ )** For every  $1 \leq p \leq \infty$  and  $v \in L^p(T)$ , there exists a constant  $C$  depending on  $p, n$  and the shape regularity of  $\mathbb{T}$  such that

$$\|\Pi_T v\|_{L^p(T)} \leq C \|v\|_{L^p(T)} \quad \forall T \in \mathcal{T}. \quad (5.67)$$

*Proof.* It is trivial to see that  $\|\Pi_T v\|_{L^2(T)} \leq \|v\|_{L^2(T)}$ . Let  $2 < p \leq \infty$  and combine an inverse estimate with a Hölder inequality to write

$$\|\Pi_T v\|_{L^p(T)} \leq C h_T^{\frac{d}{p} - \frac{d}{2}} \|\Pi_T v\|_{L^2(T)} \leq C h_T^{\frac{d}{p} - \frac{d}{2}} \|v\|_{L^2(T)} \leq C \|v\|_{L^p(T)}.$$

For  $1 \leq p < 2$  we proceed by duality. Let  $\varphi \in L^q(T)$  with  $q = \frac{p}{p-1}$ . Then

$$\int_T \Pi_T v \varphi = \int_T v \Pi_T \varphi \leq \|v\|_{L^p(T)} \|\Pi_T \varphi\|_{L^q(T)} \leq C \|v\|_{L^p(T)} \|\varphi\|_{L^q(T)},$$

which implies (5.67) and concludes the proof.  $\square$

We immediately have the following simple consequence of Lemma 5.23.

**Corollary 5.24 (best approximation of  $\Pi_T$ )** *For every  $1 \leq p \leq \infty$  and  $v \in L^p(T)$ , there exists a constant  $C_{BA} \geq 1$  depending on  $p, n$  and the shape regularity of  $\mathbb{T}$  such that*

$$\|v - \Pi_T v\|_{L^p(T)} \leq C_{BA} \inf_{w \in \mathbb{P}_{n-1}} \|v - w\|_{L^p(T)}. \quad (5.68)$$

*Proof.* We combine the invariance of  $\Pi_T$  on  $\mathbb{P}_{n-1}$ , i.e.  $\Pi_T w = w$  for  $w \in \mathbb{P}_{n-1}$ , with (5.67) to see that

$$\|v - \Pi_T v\|_{L^p(T)} = \|(v - w) - \Pi_T(v - w)\|_{L^p(T)} \leq C \|v - w\|_{L^p(T)}.$$

This implies (5.68) as asserted.  $\square$

The  $L^2$ -projection is easily computable because it entails solving the linear system (5.66). However, this flexibility comes at the expense of a best approximation constant  $C_{BA} > 1$  in (5.68) for  $p \neq 2$ . The best  $L^p$ -approximation of  $v$  in  $T$  is also computable, because it boils down to a convex minimization problem, and would result in  $C_{BA} = 1$ . This excellent property is superseded by the simplicity of (5.66), which makes  $\Pi_T v$  the approximation of choice.

**Corollary 5.25 (quasi-monotonicity of  $\Pi_T$ )** *Let  $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$  be so that  $\mathcal{T} \leq \mathcal{T}_*$ , and let  $T \in \mathcal{T}, T_* \in \mathcal{T}_*$  satisfy  $T_* \subset T$ . If  $C_{BA}$  is the constant in (5.68), then*

$$\|v - \Pi_{T_*} v\|_{L^p(T_*)} \leq C_{BA} \|v - \Pi_T v\|_{L^p(T)} \quad (5.69)$$

for all  $1 \leq p \leq \infty$ , and  $C_{BA} = 1$  for  $p = 2$ .

*Proof.* Simply use (5.68) to write

$$\|v - \Pi_{T_*} v\|_{L^p(T_*)} \leq C_{BA} \|v - \Pi_T v\|_{L^p(T_*)} \leq C_{BA} \|v - \Pi_T v\|_{L^p(T)}.$$

This is the desired bound.  $\square$

We are now ready to define the discontinuous  $\mathbb{P}_{n-1}$ -approximation  $\widehat{v}$  of  $v \in L^p(\Omega)$ . Inequality (5.69) with  $C_{BA} > 1$  is fine for most instances except Lemma 7.5 below. Therefore, we introduce a nonlinear modification of the obvious choice  $\widehat{v}$  for  $T \in \mathcal{T}$ , namely  $\widehat{v} = \Pi_T v$ . We give a recursive (and computable) definition as follows: If  $T \in \mathcal{T}_0$ , then  $\widehat{v}|_T := \Pi_T v$ ; if  $T \in \mathcal{T}$  let  $\widehat{v}|_{P(T)} \in \mathbb{P}_{n-1}$  be the approximation of  $v$  in the parent element  $P(T)$  of  $T$ , and set

$$\widehat{v}|_T := \begin{cases} \Pi_T v & \text{if } \|v - \Pi_T v\|_{L^p(T)} \leq \|v - \widehat{v}|_{P(T)}\|_{L^p(T)}, \\ \widehat{v}|_{P(T)} & \text{if } \|v - \Pi_T v\|_{L^p(T)} > \|v - \widehat{v}|_{P(T)}\|_{L^p(T)}. \end{cases} \quad (5.70)$$

We then define

$$\text{osc}_{\mathcal{T}}(v, T)_p := \|v - \widehat{v}\|_{L^p(T)} \quad \forall T \in \mathcal{T}. \quad (5.71)$$

Since the chain of elements emanating from  $\mathcal{T}_0$  and culminating with  $T$  is unique, the notion  $\text{osc}_{\mathcal{T}}(v, T)_p$  is well defined and independent of  $\mathcal{T}$ . The following result is an immediate consequence of (5.70).

**Lemma 5.26 (monotonicity of oscillation)** *For all  $1 \leq p \leq \infty$ ,  $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$  with  $\mathcal{T} \leq \mathcal{T}_*$ , and  $T_* \in \mathcal{T}_*$ ,  $T \in \mathcal{T}$  so that  $T_* \subset T$ , there holds*

$$\text{osc}_{\mathcal{T}_*}(v, T_*)_p \leq \text{osc}_{\mathcal{T}}(v, T)_p. \quad (5.72)$$

Consequently, for any  $n \geq 1$  and  $T \in \mathcal{T}$ , let  $\widehat{\mathbf{A}} \in [\mathbb{S}_{\mathcal{T}}^{n-1, -1}]^{d \times d}$ ,  $\widehat{c} \in \mathbb{S}_{\mathcal{T}}^{n-1, -1}$  be defined locally via (5.70), and let the surrogate element error indicators of  $(\mathbf{A}, c)$  be given by

$$\text{osc}_{\mathcal{T}}(\mathbf{A}, T)_r := \|\mathbf{A} - \widehat{\mathbf{A}}\|_{L^r(T)}, \quad \text{osc}_{\mathcal{T}}(c, T)_q := \|c - \widehat{c}\|_{L^q(T)}, \quad (5.73)$$

for some  $2 \leq r \leq \infty$  and  $\frac{d}{2} < q \leq \infty$  according to (5.58) for  $s = 0$ . The simplest choice  $q = 2$  yields  $\widehat{c}_T = \Pi_T c$  in (5.70), but requires the restriction  $d < 4$ , which is fine in practice.

For  $n = 1$  the situation is a bit special on two counts. First,  $\Pi_T v$  reduces to meanvalues of  $v$ , namely

$$\Pi_T \mathbf{A} := \frac{1}{|T|} \int_T \mathbf{A}, \quad \Pi_T c := \frac{1}{|T|} \int_T c, \quad \forall T \in \mathcal{T}. \quad (5.74)$$

for  $\mathbf{A} \in M(\alpha_1, \alpha_2)$ ,  $c \in R(c_1, c_2)$  defined in (5.50). Hence,  $\widehat{\mathbf{A}} \in M(\widehat{\alpha}_1, \widehat{\alpha}_2)$  with  $\widehat{\alpha}_1 = \alpha_1$ ,  $\widehat{\alpha}_2 = \alpha_2$  and  $\widehat{c} \in R(\widehat{c}_1, \widehat{c}_2)$  with  $\widehat{c}_1 = c_1$ ,  $\widehat{c}_2 = c_2$ , i.e., the  $L^2$ -projections (5.74) on piecewise constants over  $\mathcal{T}$  as well as  $\widehat{\mathbf{A}}$  and  $\widehat{c}$  satisfy the side conditions in (5.51) without changing the original range of parameters. In addition, instead of (5.73), we can exploit *superconvergence* in  $W_q^{-1}(\Omega)$  with  $q > \frac{d}{2-s} = d$  in (5.58). In fact, we utilize the orthogonality of  $\Pi_T$  in conjunction with (5.68) and (3.16) to obtain for an arbitrary function  $w \in W_{q^*}^1(\Omega)$  and  $q^* = \frac{q}{q-1}$

$$\int_T (c - \Pi_T c)w = \int_T (c - \Pi_T c)(w - \Pi_T w) \lesssim h_T^t \|c - \Pi_T c\|_{L^r(T)} |w|_{W_{q^*}^1(\omega_T)}$$

where  $t = 1 - \frac{d}{q^*} + \frac{d}{r^*} = 1 + \frac{d}{q} - \frac{d}{r} > 0$  and  $r^* = \frac{r}{r-1}$ . We consider two cases:  $r = 2, \infty$ . If  $r = 2$  and  $s = 1$ , then  $q > d$  results in  $0 < t < 2 - \frac{d}{2}$  and entails the restriction  $d < 4$ . This implies  $\|c - \widehat{c}\|_{W_q^{-1}(\Omega)} \lesssim \text{osc}_{\mathcal{T}}(c)_2$  where

$$\text{osc}_{\mathcal{T}}(c, T)_2 := h_T^t \|c - \Pi_T c\|_{L^2(T)}. \quad (5.75)$$

If  $r = \infty$  and  $s = 1$ , then  $q = \infty$  yields  $t = 1$  and  $\|c - \widehat{c}\|_{W_q^{-1}(\Omega)} \lesssim \text{osc}_{\mathcal{T}}(c)_\infty$  where

$$\text{osc}_{\mathcal{T}}(c, T)_\infty := h_T \|c - \Pi_T c\|_{L^\infty(T)}. \quad (5.76)$$

*Approximating the load.* Dealing with  $f \in H^{-1}(\Omega)$  is trickier for several reasons. First the norm in  $H^{-1}(\Omega)$  is nonlocal, so its localization is non-obvious. We recall the definition (4.52) of local oscillation  $\text{osc}_{\mathcal{T}}(f, T)_{-1}$  for  $T \in \mathcal{T}$  and Corollary 4.31 (local near-best approximation) to deduce

$$\text{osc}_{\mathcal{T}}(f, T)_{-1} := \|f - P_{\mathcal{T}}f\|_{H^{-1}(\omega_T)} \leq C_{\text{lstb}} \inf_{\chi \in \mathbb{P}_{\mathcal{T}}\omega_T} \|f - \chi\|_{H^{-1}(\omega_T)} \quad (5.77)$$

where  $C_{\text{lstb}}$  is the constant in Lemma 4.28 (local  $H^{-1}$ -stability); equivalently,  $\text{osc}_{\mathcal{T}}(f, T)_{-1}$  delivers a near best approximation of  $f$  in  $H^{-1}(\omega_T)$ . The second issue at stake is that without further assumptions on  $f$ , it is not possible to evaluate or bound the left hand side of (5.77). In Section 7 we will consider several classes of loads amenable to computation and yet relevant in practice.

A popular variant of this approach for  $f \in L^2(\Omega)$  replaces  $\chi$  in (5.77) by the  $L^2$ -projection  $\Pi_{\mathcal{T}}$  onto discontinuous piecewise polynomials of degree  $n - 1$ , and sets  $\widehat{f} = \Pi_{\mathcal{T}}f$ . This leads to the standard local weighted  $L^2$ -element error indicator

$$\widetilde{\text{osc}}_{\mathcal{T}}(f, T)_{-1} := h_T \|f - \widehat{f}\|_{L^2(T)} \quad \forall T \in \mathcal{T}. \quad (5.78)$$

*Data error estimators.* They are the following quantities for the coefficients  $(\mathbf{A}, c)$

$$\begin{aligned} \text{osc}_{\mathcal{T}}(\mathbf{A})_r &:= \left( \sum_{T \in \mathcal{T}} \text{osc}_{\mathcal{T}}(\mathbf{A}, T)_r^r \right)^{\frac{1}{r}}, \\ \text{osc}_{\mathcal{T}}(c)_q &:= \left( \sum_{T \in \mathcal{T}} \text{osc}_{\mathcal{T}}(c, T)_q^q \right)^{\frac{1}{q}}, \end{aligned} \quad (5.79)$$

which accumulate in  $\ell^r$  and  $\ell^q$  for  $2 \leq r \leq \infty$  and  $\frac{d}{2} < q$ ; recall that  $q = 2$  is an admissible choice provided  $d < 4$ . In contrast, the global error estimator for  $f$

$$\text{osc}_{\mathcal{T}}(f)_{-1} := \left( \sum_{T \in \mathcal{T}} \text{osc}_{\mathcal{T}}(f, T)_{-1}^2 \right)^{\frac{1}{2}} \quad (5.80)$$

accumulates in  $\ell^2$ . The total data error estimator satisfies (5.64) and reads

$$\text{osc}_{\mathcal{T}}(\mathcal{D}) := \text{osc}_{\mathcal{T}}(\mathbf{A})_r + \text{osc}_{\mathcal{T}}(c)_q + \text{osc}_{\mathcal{T}}(f)_{-1}. \quad (5.81)$$

*The module DATA.* This module reduces the oscillation of data  $\mathcal{D} = (\mathbf{A}, c, f)$  sequentially. It consists of a linear approximation followed by a nonlinear correction.

Given a coefficient  $v = \mathbf{A}, c$ , a mesh  $\mathcal{T} \in \mathbb{T}$ , a tolerance  $\tau$ , an accumulation index  $1 \leq p \leq \infty$ , and a number of bisections  $b \geq 1$  per marked element, the call

$$[\widetilde{\mathcal{T}}, \widetilde{v}] = \text{GREEDY}(v, \mathcal{T}, \tau, p, b)$$

returns a conforming refinement  $\widetilde{\mathcal{T}}$  of  $\mathcal{T}$  and a piecewise polynomial approximation  $\widetilde{v}$  of  $v$  over  $\widetilde{\mathcal{T}}$  such that the oscillation computed with  $v - \widetilde{v}$  satisfies

$$\text{osc}_{\widetilde{\mathcal{T}}}(v)_p \leq \tau.$$

For the load function  $f$ , since the computation of  $\text{osc}_{\mathcal{T}}(f)_{-1}$  is impossible without further assumptions on  $f$ , we will consider three surrogate estimators  $\widetilde{\text{osc}}_{\mathcal{T}}(f)_{-1}$  in Section 7.3 that also accumulate in  $\ell^p$  such that, for all  $\mathcal{T} \in \mathbb{T}$ ,

$$\text{osc}_{\mathcal{T}}(f)_{-1} \leq C_{\text{data}} \widetilde{\text{osc}}_{\mathcal{T}}(f)_{-1},$$

where  $C_{\text{data}} \geq 1$ . GREEDY applied to the surrogate estimator constructs  $\widehat{\mathcal{T}} \geq \mathcal{T}$  satisfying

$$\widetilde{\text{osc}}_{\widehat{\mathcal{T}}}(f)_{-1} \leq \tau \quad \Rightarrow \quad \text{osc}_{\widehat{\mathcal{T}}}(f)_{-1} \leq C_{\text{data}} \tau. \quad (5.82)$$

In all cases, the routine GREEDY is similar to that in Algorithm 3.18 (greedy algorithm) with several important distinctions: it accumulates the local error indicators in the  $\ell^p$ -norm and starts from any mesh  $\mathcal{T} \geq \mathcal{T}_0$  to save computational work.

Finally, the structure of the module DATA is as follows: it concatenates GREEDY with CONSTRAINT-A and CONSTRAINT-c in order to satisfy Assumption 5.22 (properties of DATA). The routine GREEDY deals with pure approximation without constraints: called with tolerance  $\tau/3$ , it sequentially reduces the oscillation for  $A, c, f$  with the most recent updated mesh to reduce their errors so that

$$\text{osc}_{\widehat{\mathcal{T}}}(A)_r \leq \tau/3, \quad \text{osc}_{\widehat{\mathcal{T}}}(c)_q \leq \tau/3, \quad \widetilde{\text{osc}}_{\widehat{\mathcal{T}}}(f)_{-1} \leq \tau/3$$

on a conforming refinement  $\widehat{\mathcal{T}} \geq \mathcal{T}$ . This is discussed in detail in Section 7.1.

From (5.82), we get  $\text{osc}_{\widehat{\mathcal{T}}}(f)_{-1} \leq C_{\text{data}} \tau/3$ . On the other hand, the resulting coefficients  $(\widehat{A}, \widehat{c})$  most likely do not satisfy the constraints (5.51) for  $n > 1$ . This requires a further nonlinear correction

$$[\widehat{A}] = \text{CONSTRAINT-A}(\widehat{\mathcal{T}}, \widehat{A}), \quad [\widehat{c}] = \text{CONSTRAINT-c}(\widehat{\mathcal{T}}, \widehat{c}),$$

that enforces (5.51) on the same grid  $\widehat{\mathcal{T}}$  without compromising the accuracy gain produced by GREEDY: there exists a constant  $\geq 1$ , still denoted  $C_{\text{data}}$  for simplicity, such that

$$\text{osc}_{\widehat{\mathcal{T}}}(\widehat{A})_r \leq C_{\text{data}} \tau/3, \quad \text{osc}_{\widehat{\mathcal{T}}}(\widehat{c})_q \leq C_{\text{data}} \tau/3 \quad \Rightarrow \quad \text{osc}_{\widehat{\mathcal{T}}}(\mathcal{D}) \leq C_{\text{data}} \tau.$$

For instance, for a fixed parameter  $L \geq 2$ , we get  $\widehat{\alpha}_1 = \frac{1}{2}\alpha_1$  and  $\widehat{\alpha}_2 = (1 + 4L)\frac{\alpha_2}{2}$  for the parameters in (5.51). We give details in Sections 7.2, 7.3, and 7.4.

The optimality properties of DATA hinge on the performance of GREEDY and the regularity of  $\mathcal{D}$ . Since this is not necessary for the present convergence assessment, we discuss it later in Section 7.

#### 5.4.3 Computational cost of GALERKIN

The output pair  $(\widehat{\mathcal{T}}, \widehat{\mathcal{D}})$  of DATA is next taken by GALERKIN, the one-step AFEM of Algorithm 5.4 in Section 5.2.1, to run an inner loop of the form (5.18) with fixed discrete data  $\widehat{\mathcal{D}}$  and initial mesh  $\widehat{\mathcal{T}}$ . The call (5.19) of GALERKIN stops as soon as the error tolerance  $\varepsilon$  is reached, which takes a finite number of iterations



because GALERKIN is a contraction between consecutive iterates, and creates the next mesh-solution pair  $(\mathcal{T}, u_{\mathcal{T}})$ . It is worth noticing that, in the absence of this stopping test, the Galerkin solution  $u_{\mathcal{T}}$  would converge to the solution  $\widehat{u} = u(\widehat{\mathcal{D}})$  of (5.5), which is not the desired solution  $u = u(\mathcal{D})$  of (2.5).

We stress that, in view of (5.63) and (5.65), the relative resolution of the modules DATA and GALERKIN is critical for the discrepancy between the exact and perturbed solutions  $u$  and  $\widehat{u}$ . This is ultimately responsible for the performance of AFEM-TS and is studied in Section 6.

We now investigate the number of iterations within GALERKIN, which dictate its computational cost. We point out that at iteration  $k - 1 \geq 0$  of AFEM-TS, the output  $(\mathcal{T}_k, u_k)$  of GALERKIN, and thus of AFEM-TS, satisfies

$$\eta_k(u_k) = \eta_{\mathcal{T}_k}(u_k) \leq \varepsilon_{k-1} \quad \Rightarrow \quad |u_k - \widehat{u}_{k-1}|_{H_0^1(\Omega)} \leq C_U \varepsilon_{k-1} \quad (5.83)$$

according to (5.12). We recall that  $\widehat{u}_{k-1} = \widehat{u}_{k-1}(\widehat{\mathcal{D}}_{k-1}) \in H_0^1(\Omega)$  is the exact solution with discrete data  $\widehat{\mathcal{D}}_{k-1}$ , and that  $\mathcal{E}_{\mathcal{T}_k}(u_k, f)$  is defined with discrete data  $\widehat{\mathcal{D}}_{k-1}$  and satisfies  $\mathcal{E}_{\mathcal{T}_k}(u_k, f) = \eta_{\mathcal{T}_k}(u_k)$  because data oscillation  $\text{osc}_{\mathcal{T}_k}(f)_{-1} = 0$ . The next iteration  $k$  of AFEM-TS calls DATA, which in turn refines the mesh  $\mathcal{T}_k$  to  $\widehat{\mathcal{T}}_k$  and updates the data approximation from  $\widehat{\mathcal{D}}_{k-1}$  to  $\widehat{\mathcal{D}}_k$  over  $\widehat{\mathcal{T}}_k$ . The pair  $(\widehat{\mathcal{T}}_k, \widehat{\mathcal{D}}_k)$  determines the first Galerkin solution  $u_{k,0} \in \mathbb{V}_{k,0} = \mathbb{V}_{\widehat{\mathcal{T}}_k}$  of GALERKIN and corresponding estimator  $\eta_{k,0}(u_{k,0})$  with  $\mathcal{T}_{k,0} = \widehat{\mathcal{T}}_k$ , which must satisfy

$$\eta_{k,0}(u_{k,0}) > \varepsilon_k \quad (5.84)$$

for GALERKIN to be executed. The reduction of  $\eta_{\mathcal{T}_{k,j}}(u_{k,j})$  for  $j \geq 1$  dictates the number of iterations of GALERKIN. We examine this next.

**Proposition 5.27 (computational cost of GALERKIN)** *If the assumptions of Theorem 5.8 are valid, then for any  $k \in \mathbb{N}$ , the number of subiterations  $J_k$  inside a call to GALERKIN at iteration  $k$  of AFEM-TS is bounded independently of  $k$ .*

*Proof.* The  $j$ -th error  $e_{k,j} := |\widehat{u}_k - u_{k,j}|_{H_0^1(\Omega)}$  within GALERKIN converges linearly in view of Corollary 5.10 (linear convergence of error) because the discrete data  $\widehat{\mathcal{D}}_k$  is fixed in these inner iterations. Exploiting the lower bound  $C_L \eta_{k,j}(u_{k,j}) \leq e_{k,j}$  stated in (5.12), we thus deduce

$$\eta_{k,j}(u_{k,j}) \leq C_L^{-1} e_{k,j} \leq C_L^{-1} C_* \alpha^{j-i} e_{k,i}, \quad j \geq i \geq 0,$$

whence  $\eta_{k,j}(u_{k,j}) \leq C_{\#} \alpha^j e_{k,0}$  with  $C_{\#} := C_L^{-1} C_*$ . The number of iterations of GALERKIN depends on the size of  $\eta_{k,0}(u_{k,0})$  relative to  $\varepsilon_k$ . We assume that  $\eta_{k,0}(u_{k,0}) > \varepsilon_k$  according to (5.84). We first prove that  $\eta_{k,0}(u_{k,0}) \lesssim \varepsilon_k$  and next argue that  $J_k$  is bounded uniformly in  $k$ . We proceed in two steps.

**[1] Bound on  $|\widehat{u}_k - u_{k,0}|_{H_0^1(\Omega)}$ :** Since  $u_k \in \mathbb{V}_k \subset \mathbb{V}_{k,0} = \mathbb{V}_{\widehat{\mathcal{T}}_k}$ , and the Galerkin solution  $u_{k,0} \in \mathbb{V}_{k,0}$  minimizes the error  $\|u_{k,0} - \widehat{u}_k\|_{\Omega}$  in  $\mathbb{V}_{k,0}$ , relative to the

energy norm induced by the bilinear form  $\widehat{\mathcal{B}}$  with discrete data  $\widehat{\mathcal{D}}_k$ , we deduce

$$\|u_{k,0} - \widehat{u}_k\|_{\Omega} \leq \|u_k - \widehat{u}_k\|_{\Omega} \leq \sqrt{C_{\widehat{\mathcal{B}}}} \left( |u_k - \widehat{u}_{k-1}|_{H_0^1(\Omega)} + |\widehat{u}_{k-1} - \widehat{u}_k|_{H_0^1(\Omega)} \right),$$

where the last inequality uses (5.3) for  $\widehat{\mathcal{B}}$ . Invoking the a posteriori upper bound (5.13) and the termination condition of GALERKIN at step  $k - 1$ , we obtain

$$|u_k - \widehat{u}_{k-1}|_{H_0^1(\Omega)} \leq C_U \mathcal{E}_{\mathcal{T}_k}(u_k, f) = C_U \eta_k(u_k) \leq C_U \varepsilon_{k-1} = 2C_U \varepsilon_k.$$

On the other hand, using (5.65) with  $\tau = \omega \varepsilon_k$  and  $0 < \omega \leq 1$ , we arrive at

$$|u - \widehat{u}_k|_{H_0^1(\Omega)} \leq C_1 \varepsilon_k,$$

with  $C_1 = \omega C_D$ . The triangle inequality thus yields

$$|\widehat{u}_{k-1} - \widehat{u}_k|_{H_0^1(\Omega)} \leq |u - \widehat{u}_{k-1}|_{H_0^1(\Omega)} + |u - \widehat{u}_k|_{H_0^1(\Omega)} \leq C_1(\varepsilon_{k-1} + \varepsilon_k) = 3C_1 \varepsilon_k,$$

whence

$$e_{k,0} = |u_{k,0} - \widehat{u}_k|_{H_0^1(\Omega)} \leq \sqrt{\frac{C_{\widehat{\mathcal{B}}}}{c_{\widehat{\mathcal{B}}}}} (2C_U + 3C_1) \varepsilon_k =: C_2 \varepsilon_k.$$

$\square$  *Bound on  $J_k$ :* We observe that GALERKIN stops once  $\eta_{k,j}(u_{k,j}) \leq \varepsilon_k$ . Since the smallest such  $j$  is  $J_k$ , we see that

$$\varepsilon_k < \eta_{k,J_k-1}(u_{k,J_k-1}) \leq C_{\#} \alpha^{J_k-1} e_{k,0} \leq C_{\#} C_2 \varepsilon_k \alpha^{J_k-1}.$$

This implies the asserted bound  $J_k \leq 1 + \frac{\log(C_{\#} C_2)}{\log \alpha^{-1}}$  uniform in  $k$ .  $\square$

#### 5.4.4 Realization of AFEM-TS

We now make the two step AFEM algorithm precise.

**Algorithm 5.28 (AFEM-TS)** Given an initial tolerance  $\varepsilon_0 > 0$ , a target tolerance  $\text{tol}$  and initial mesh  $\mathcal{T}_0$ , as well as a safety parameter  $\omega \in (0, 1]$ , AFEM consists of the two-step algorithm:

```

 $[\mathcal{T}, u_{\mathcal{T}}] = \text{AFEM-TS}(\mathcal{T}_0, \varepsilon_0, \omega, \text{tol})$ 
  set  $k = 0$  and do
     $[\widehat{\mathcal{T}}_k, \widehat{\mathcal{D}}_k] = \text{DATA}(\mathcal{T}_k, \mathcal{D}, \omega \varepsilon_k)$ 
     $[\mathcal{T}_{k+1}, u_{k+1}] = \text{GALERKIN}(\widehat{\mathcal{T}}_k, \widehat{\mathcal{D}}_k, \varepsilon_k)$ 
     $\varepsilon_{k+1} = \frac{1}{2} \varepsilon_k$ 
     $k \leftarrow k + 1$ 
  while  $\varepsilon_{k-1} > \text{tol}$ 
  return  $\mathcal{T}_k, u_k$ 

```

**Proposition 5.29 (convergence of AFEM-TS)** *For each  $k \geq 0$  the modules DATA and GALERKIN converge in a finite number of iterations, the latter independent of  $k$ . Moreover, there exists a constant  $C_*$  depending on  $\mathcal{T}_0, \Omega, d, n$ , the Lebesgue*

exponents  $r, q$  in  $D(\Omega)$ , the parameters  $\alpha_1, \alpha_2, c_1, c_2$  in (5.48) and (5.49), and the shape regularity constant of  $\mathbb{T}$ , such that the output of the  $(k+1)$ -th iteration  $[\mathcal{T}_{k+1}, u_{k+1}] = \text{GALERKIN}(\widehat{\mathcal{T}}_k, \widehat{\mathcal{D}}_k, \varepsilon_k)$  satisfies  $|u - u_{k+1}|_{H_0^1(\Omega)} \leq C_* \varepsilon_k$  for all  $k \geq 0$ . Therefore, AFEM-TS stops after  $K < 2 + \frac{\log \frac{\varepsilon_0}{\text{tol}}}{\log 2}$  iterations and delivers

$$|u - u_K|_{H_0^1(\Omega)} \leq C_* \text{tol}.$$

*Proof.* In view of Assumption 5.22 (properties of DATA) the module DATA iterates a finite number of steps to reach tolerance  $\tau = \omega \varepsilon_k$  for every  $k \geq 0$ . Moreover, the number of iterations of GALERKIN is independent of  $k$  due to Proposition 5.27 (computational cost of GALERKIN), whence we deduce that each loop of AFEM-TS requires finite iterations. Thus, the output  $u_{k+1}$  of the  $(k+1)$ -th loop satisfies

$$|u - u_{k+1}|_{H_0^1(\Omega)} \leq |u - \widehat{u}_k|_{H_0^1(\Omega)} + |\widehat{u}_k - u_{k+1}|_{H_0^1(\Omega)} \leq (\omega C_D + C_U) \varepsilon_k = C_* \varepsilon_k,$$

according to (5.65) with  $\tau \leq \omega \varepsilon_k$  and (5.83) for all  $k \geq 0$ . Finally, AFEM-TS terminates after  $K$  loops, where  $K$  satisfies  $\frac{1}{2} \text{tol} < \varepsilon_{K-1} \leq \text{tol}$ , and the asserted estimate holds.  $\square$

This elementary proof gives no insight whether the  $H_0^1$ -error decays optimally in terms of degrees of freedom. We assess this fundamental question in Sections 6 and 7, but investigate it computationally in Section 5.4.5.

A two-step algorithm similar to AFEM-TS was first proposed in [Stevenson 2008], and further explored in [Bonito et al. 2013b, Cohen et al. 2012]. Note that other quantities, such as the number of degrees of freedom, could be employed to stop AFEM-TS instead. It is also worth realizing that the structure of the algorithm is independent of the size of tolerance  $\text{tol}$ . In this vein, a user could take  $\varepsilon_0 = \text{tol}$ , provided  $\text{tol}$  is affordable by the computational resources at hand. With such a choice, the modules DATA and GALERKIN run only once, in sequence: data are approximated to the desired accuracy in one shot, then fed to the PDE solver which produces the approximate solution. Since the quasi-optimality theory in Section 6 would also hold for this choice of  $\varepsilon_0$ , one might wonder why not using this simpler strategy. We stress that iterating over  $\varepsilon_k$  has the following advantages:

- *Restarts:* Dynamical shrinking of  $\text{tol}$ , for instance to account for the user decision to improve the accuracy, does not entail a restart of AFEM-TS but rather a continuation from the previous computed solution. In this sense, the resulting iteration would be similar to the proposed structure of AFEM-TS.
- *Computational resources:* AFEM-TS allows for “balanced investment” of computational resources between the modules DATA and GALERKIN. If the stopping criterion, either accuracy or number of degrees of freedom, is unrealistic for the problem at hand, AFEM-TS would still produce a discrete solution with equilibrated data and solution errors.
- *Nonlinear problems:* The interleaving approach of AFEM-TS appears to be

better suited for treating nonlinear problems for which data  $\mathcal{D}$  may depend on the solution. Therefore a call to GALKERIN, and corresponding solution update, must precede a call to DATA.

- *Iterative solvers:* If an efficient iterative solver is adopted within SOLVE, then the previous discrete solution of GALKERIN could be taken as initial iterate, thereby making SOLVE fast because  $\frac{\epsilon_{k+1}}{\epsilon_k} = \frac{1}{2}$ . If instead one computes with DATA alone until the fixed tolerance `tol` is reached, then GALKERIN would work directly on fine meshes, which are not adapted to the geometric domain singularities, and without good initial guess. This would lead to fewer but heavier iterations of GALKERIN, which is detrimental from a linear algebra perspective.

#### 5.4.5 Computational assessment of AFEM-TS

In this section we explore computationally the relative performance of GALKERIN and DATA, for the two-step AFEM, and elucidate the behavior of data and coefficient oscillations within DATA. Our observations motivate the rigorous study of Section 6, which provides theoretical support to our experiments. The numerical computations are made with the help of [Funken, Praetorius and Wissgott 2011].

We consider problem (2.5) in the L-shaped domain  $\Omega = (-1, 1)^2 \setminus ([0, 1] \times [-1, 0])$ , with diffusion term  $A = aI$ , where

$$a(x, y) = 1 + \exp(-50((x + 0.5)^2 + (y + 0.5)^2)) + \exp(-50((x + 0.5)^2 + (y - 0.5)^2)),$$

and reaction term

$$c(x, y) = 1 + \exp(-50((x + 0.5)^2 + y^2)) + \exp(-50(x^2 + (y - 0.5)^2));$$

note that the Gaussians in the definition of  $a$  and  $c$  have the same intensity but are located in different places within  $\Omega$ . The load term  $f$  and the Dirichlet boundary conditions are chosen in accordance with the analytical solution

$$u(x, y) = r^{\frac{2}{3}} \sin(2\alpha/3) + \exp(-1000((x - 0.5)^2 + (y - 0.5)^2)),$$

where  $(r, \alpha)$  are the polar coordinates around the origin. Notice that the exact solution  $u$  is singular at the reentrant corner: it belongs to the Sobolev spaces  $H(\Omega)^{\frac{5}{3}-\varepsilon}$  with  $\varepsilon > 0$  and  $W_p^2(\Omega)$  with  $p > 1$ . It also exhibits a rapid transition of order  $10^{-3/2}$  around the point  $(0.5, 0.5)$  due to the presence of a very narrow Gaussian. The Gaussians are meant to test the performance of the module DATA, while in addition the corner singularity of the solution tests the execution of the module GALKERIN.

We utilize the following parameters in the numerical test

$$\theta = 0.5, \quad \omega = 1, \quad \text{tol} = 2^{-4}, \quad h_0 = 0.125, \quad \varepsilon_0 = 1.$$

Notice that the number of iterations of the algorithm AFEM is  $K = \log_2(\epsilon_0/\text{tol}) = 4$ . We compute the relative  $H^1$ -error between the exact solution  $u$  and the FEM solution  $u_{\mathcal{T}}$  and notice that its decay rate is  $(\#\mathcal{T})^{-1/2}$  in Fig. 5.6 (left). This rate

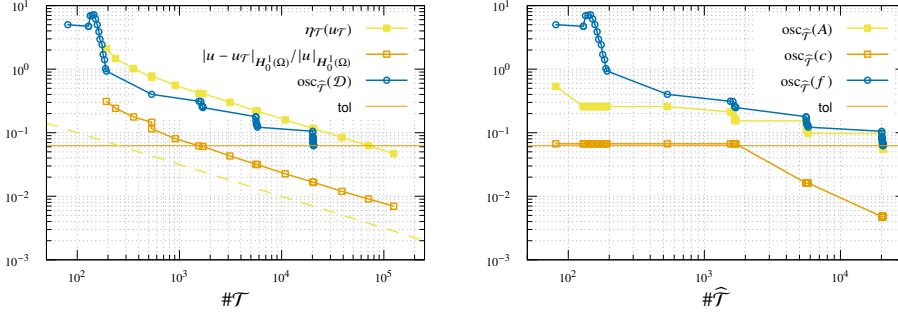


Figure 5.6. Left: estimator  $\eta_T(u_T)$ , data error  $\text{osc}_{\hat{\mathcal{T}}}(\mathcal{D})$ , and relative  $H^1$ -error obtained with the algorithm AFEM performing  $b = 3$  bisections per marked element. The optimal decay is indicated by the dashed line with slope  $-0.5$ . Right: Diffusion error  $\text{osc}_{\hat{\mathcal{T}}}(A)$ , reaction error  $\text{osc}_{\hat{\mathcal{T}}}(c)$ , load error  $\text{osc}_{\hat{\mathcal{T}}}(f)_{-1}$ , obtained with the algorithm AFEM.

is consistent with that of the PDE estimator  $\eta_T(u_T)$  and data estimator  $\text{osc}_T(\mathcal{D})$ . In Fig. 5.6 (right) we display the component of the data error  $\text{osc}_{\hat{\mathcal{T}}}(A)$ ,  $\text{osc}_{\hat{\mathcal{T}}}(c)$ ,  $\text{osc}_{\hat{\mathcal{T}}}(f)_{-1}$  defined in (5.79) and (5.80) with local contributions defined in (5.73) for  $A$  with  $r = \infty$ , in (5.76) for  $c$  with  $t = 1$  and (5.78) for  $f$ . Recall that at each iteration  $k$ , DATA circles through  $\text{osc}_{\hat{\mathcal{T}}}(A)$ ,  $\text{osc}_{\hat{\mathcal{T}}}(c)$ , and  $\text{osc}_{\hat{\mathcal{T}}}(f)_{-1}$  reducing each of these oscillations to  $1/3$  of the iteration tolerance  $\varepsilon_k = 2^{-k}$ . The presence of the weight  $h^t$  in  $\text{osc}_{\hat{\mathcal{T}}}(c)$  considerably reduces the influence of the approximation of  $c$ , which is below threshold from the start and thus never generates any refinement (see Table 5.1). The local oscillation for  $f$  also includes a weight vanishing as  $h \rightarrow 0$  but  $\text{osc}_{\hat{\mathcal{T}}}(f)_{-1}$  is above the desired tolerance, which would in principle generate refinements. However, since at each iteration DATA considers  $\text{osc}_{\hat{\mathcal{T}}}(A)$  first and the regions refined to reduce  $\text{osc}_{\hat{\mathcal{T}}}(f)_{-1}$  are included in the regions needed to be refined to reduce  $\eta_T(u_T)$  and  $\text{osc}_{\hat{\mathcal{T}}}(A)$ , the GREEDY routine applied to  $f$  does not refine any element except during the first iteration when the Galerkin error has not yet been reduced by the algorithm. Overall, the reduction of the Galerkin error is driving most of the refinements. The number of marked element to reduce the approximation errors of  $A$ ,  $c$ ,  $f$ , and the residual estimator are reported in Table 5.1 along with those when  $b = 1$  refinement is used per marked element. In Figure 5.7, we provide the resulting meshes after the first iteration of DATA and GALERKIN.

### 5.5 Convergence for other boundary conditions

We consider first the variational problem (2.13) with *Robin* boundary condition. We approximate data  $\mathcal{D} = (A, c, p, f, g)$  by piecewise polynomials  $\hat{\mathcal{D}} = (\hat{A}, \hat{c}, \hat{p}, \hat{f}, \hat{g})$ . The only difference with respect to (5.2) is that the new functions  $(p, g)$  are approximated on  $\partial\Omega$  by discontinuous polynomials  $(\hat{p}, \hat{g})$  of degree  $n - 1$  and  $2n - 1$ . The projection operator  $P_T$  approximates  $g\delta_{\partial\Omega}$  by  $\hat{g}\delta_{\partial\Omega} = P_T(g\delta_{\partial\Omega})$  without compon-

$k$	$\text{osc}_{\widehat{\tau}}(\mathbf{A})$		$\text{osc}_{\widehat{\tau}}(c)$		$\text{osc}_{\widehat{\tau}}(f)_{-1}$		$\eta_{\tau}(u_{\tau})$	
	b=1	b=3	b=1	b=3	b=1	b=3	b=1	b=3
1	32	16	0	0	26	13	363	308
2	16	16	0	0	0	0	1636	1138
3	120	43	0	0	0	0	7447	4227
4	123	62	0	0	0	0	42792	15268
5	82	138	0	0	0	0	144345	102350

Table 5.1. Number of marked elements to reduce the data and Galerkin errors at each iterations  $k = 1, 2, 3, 4$  of AFEM-TS when using  $b = 1$  and  $b = 3$  refinements per marked element. Regardless of the value used for  $b$ , the reduction of the Galerkin error is driving most of the refinements followed by the error in the approximation of the diffusion coefficient  $\mathbf{A}$ . The approximation of  $f$  is subordinate to the approximation of  $u$  and  $\mathbf{A}$  arising earlier in the adaptive loop and thus does not generate any refinement except during the first iteration when the Galerkin error has not yet been tackled by the algorithm. The approximation of  $c$  is below the final tolerance from the start and does not generate any refinement.

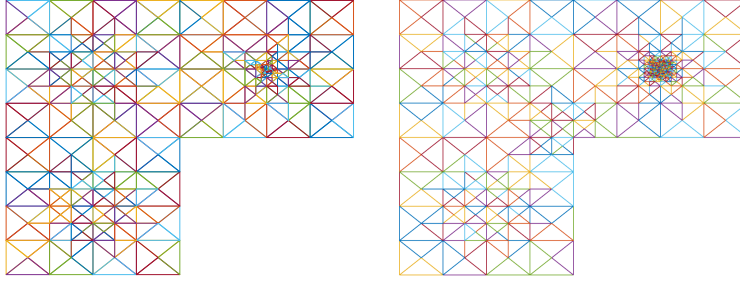


Figure 5.7. Resulting meshes after the first iteration of DATA (left) and after the first iteration of GALERKIN (right). DATA marked 29 elements for refinement while GALERKIN marked 308 elements. Refer to Table 5.1 for more details.

ent in the bulk because  $g\delta_{\partial\Omega}$  is a line Dirac mass aligned with the mesh. Discrete functions  $(\widehat{p}, \widehat{g})$  must be produced by DATA, subject to a sign constraint on  $p$ . The approximate bilinear form  $\widehat{\mathcal{B}}$  and linear functional  $\widehat{\ell}$  read

$$\widehat{\mathcal{B}}[w, v] := \int_{\Omega} \nabla v \cdot \widehat{\mathbf{A}} \nabla w + \widehat{c}vw + \int_{\partial\Omega} \widehat{p}vw, \quad \widehat{\ell}(v) := \langle \widehat{f}, v \rangle + \int_{\partial\Omega} \widehat{g}v. \quad (5.85)$$

The a posteriori error estimates of Section 4 extend to this pair  $(\widehat{\mathcal{B}}, \widehat{\ell})$ . The algorithms GALERKIN, AFEM-SW, and AFEM-TS are similar to those above and possess a similar supporting convergence theory. The *Neumann* boundary condition is a particular case with  $p = 0$ . We do not pursue this any further.

However, the *pure Neumann* boundary condition is special because of the global compatibility condition  $\widehat{\ell}(1) = \langle \widehat{\ell}, 1 \rangle = 0$ . In Section 4.10 we introduce a new

projection operator  $\tilde{P}_\mathcal{T}$ , a modification of  $P_\mathcal{T}$ , with the requisite properties of local approximation and global compatibility  $\langle \tilde{P}_\mathcal{T}\ell, 1 \rangle = 0$  provided  $\ell \in H^1(\Omega)^*$  satisfies  $\langle \ell, 1 \rangle = 0$ . We thus set  $\tilde{\ell} = \tilde{P}_\mathcal{T}\ell$  to solve the Galerkin problems and use  $\tilde{P}_z$  in the local indicators. We do not explore this matter further.

For a *non-homogeneous Dirichlet* boundary data  $g \in H^{1/2}(\partial\Omega)$ , DATA must produce a continuous piecewise polynomial approximation  $\hat{g}$  of degree  $n$ , thereby consistent with the Galerkin solution  $u_\mathcal{T}$ . The Dirichlet oscillation  $\text{osc}_\mathcal{T}(g)_{1/2}$  is defined in (4.96) and is locally computable. Data oscillation now becomes

$$\text{osc}_\mathcal{T}(\ell) = \text{osc}_\mathcal{T}(f)_{-1} + \text{osc}_\mathcal{T}(g)_{1/2}$$

and added to the PDE estimator  $\eta_\mathcal{T}(u_\mathcal{T})$  for  $g = 0$  gives a full estimator equivalent to the error, according to Theorem 4.74 (estimators for general Dirichlet conditions). With these minor modifications, the convergence theory for GALERKIN, AFEM-SW, and AFEM-TS extends to this case. We do not provide any further details.

### 5.6 Convergence for alternative estimators

We have so far developed a convergence theory for the residual estimator  $\mathcal{E}_\mathcal{T}(u_\mathcal{T}, f)$ . The purpose of this section is to extend this theory to the three alternative estimators discussed in Section 4.9, namely

- $\mathcal{E}_\mathcal{T}^{\text{lpb}}(u_\mathcal{T}, f)^2 = \eta_\mathcal{T}^{\text{lpb}}(u_\mathcal{T})^2 + \text{osc}_\mathcal{T}(f)_{-1}^2$ : estimator based on *local problems*;
- $\mathcal{E}_\mathcal{T}^{\text{hier}}(u_\mathcal{T}, f)^2 = \eta_\mathcal{T}^{\text{hier}}(u_\mathcal{T})^2 + \text{osc}_\mathcal{T}(f)_{-1}^2$ : *hierarchical* estimator
- $\mathcal{E}_\mathcal{T}^{\text{feq}}(u_\mathcal{T}, f)^2 = \eta_\mathcal{T}^{\text{feq}}(u_\mathcal{T})^2 + \text{osc}_\mathcal{T}(f)_{-1}^2$ : estimator based on *flux equilibration*.

They are all computed on stars  $\omega_z$  with  $z \in \mathcal{V}$  and possess a similar structure. The first term is the PDE estimator, from now on called  $\zeta_\mathcal{T}(u_\mathcal{T})$  to refer to any of them, and is locally equivalent to the discrete residual  $P_\mathcal{T}R_\mathcal{T}$

$$\zeta_\mathcal{T}(u_\mathcal{T}, z) \approx \|P_\mathcal{T}R_\mathcal{T}\|_{H^{-1}(\omega_z)} \quad \forall z \in \mathcal{V}; \quad (5.86)$$

see Theorems 4.59, 4.64, and 4.69. In fact, they are all different mechanisms to extract information from  $P_\mathcal{T}R_\mathcal{T}$ . Since the vertex-indexed residual PDE indicator  $\eta_\mathcal{T}(u_\mathcal{T}, z) := \eta_\mathcal{T}^{\text{res}}(u_\mathcal{T}, z)$ , defined in (4.70a), is also proven to be equivalent to  $\|P_\mathcal{T}R_\mathcal{T}\|_{H^{-1}(\omega_z)}$  in Theorem 4.58 (vertex-indexed modified residual estimator), we deduce the existence of two equivalence constants  $C_L^{\text{eq}} \leq C_U^{\text{eq}}$  such that

$$C_L^{\text{eq}} \eta_\mathcal{T}(u_\mathcal{T}, z) \leq \zeta_\mathcal{T}(u_\mathcal{T}, z) \leq C_U^{\text{eq}} \eta_\mathcal{T}(u_\mathcal{T}, z) \quad \forall z \in \mathcal{V}. \quad (5.87)$$

Following [Kreuzer and Siebert 2011], we will exploit this property to prove convergence of AFEM driven by  $\zeta_\mathcal{T}(u_\mathcal{T})$ . An obstruction for a direct convergence theory is that our preceding results rely heavily on Lemma 4.53 (reduction property of the estimator), which is not necessarily valid for any of the alternative estimators. We refer to [Cascón and Nochetto 2012] who present a direct approach based on the local lower bound for discrete solutions of Theorem 4.51 (local lower bound for corrections). The latter is guaranteed by Definition 4.50 (interior vertex property)



for operators with coefficients  $A$  piecewise constant and  $c = 0$ , and any polynomial degree  $n \geq 1$ , but we do not know its validity for more general coefficients  $(A, c)$ .

The key for convergence is imposing a Dörfler marking. We say that a set of vertices  $\mathcal{M}_{\mathcal{T}}$  satisfies a Dörfler property with parameter  $\theta \leq 1$  if

$$\zeta_{\mathcal{T}}(u_{\mathcal{T}}, \mathcal{M}_{\mathcal{T}})^2 := \sum_{z \in \mathcal{M}_{\mathcal{T}}} \zeta_{\mathcal{T}}(u_{\mathcal{T}}, z)^2 \geq \theta^2 \sum_{z \in \mathcal{V}} \zeta_{\mathcal{T}}(u_{\mathcal{T}}, z)^2 =: \zeta_{\mathcal{T}}(u_{\mathcal{T}})^2. \quad (5.88)$$

Let  $\mathcal{M}$  be the collection of elements contained in the stars  $\omega_z$  with  $z \in \mathcal{M}_{\mathcal{T}}$ . Then MARK marks all elements in  $\mathcal{M}$ , and REFINE bisects them  $b \geq 1$  times. This gives rise to a star-driven GALERKIN procedure.

**Lemma 5.30 (Dörfler property)** *If the set of vertices  $\mathcal{M}_{\mathcal{T}}$  satisfies a Dörfler property with parameter  $\theta$  for  $\zeta_{\mathcal{T}}(u_{\mathcal{T}})$ , then  $\mathcal{M}$  satisfies a Dörfler property with parameter  $\bar{\theta} = \frac{C_L^{eq}}{C_U^{eq}} \theta$  for  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$ .*

*Proof.* Simply use (5.87) to derive (5.88) for  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$  with parameter  $\bar{\theta}$ .  $\square$

Hence, star-driven procedures for  $\zeta_{\mathcal{T}}(u_{\mathcal{T}})$  lead to the corresponding counterparts for  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$ . It turns out that algorithms GALERKIN, AFEM-SW, and AFEM-TS can be reformulated for vertex-indexed indicators  $\{\eta_{\mathcal{T}}(u_{\mathcal{T}}, z)\}_{z \in \mathcal{V}}$  as defined in (4.70a), without changing their essential properties. We may thus wonder about them driven by  $\{\zeta_{\mathcal{T}}(u_{\mathcal{T}}, z)\}_{z \in \mathcal{V}}$  instead. Since these algorithms hinge on the Dörfler property (5.88), Lemma 5.30 gives rise to similar convergence properties for  $\zeta_{\mathcal{T}}(u_{\mathcal{T}})$ -driven algorithms provided (5.88) is enforced. We state this next without proof.

**Corollary 5.31 (convergence of GALERKIN)** *If the coefficients  $(A, c, f) \in \mathbb{D}_{\mathcal{T}}$ , then there exist  $0 < \bar{\alpha} < 1$  and  $C_*, C_{\#} > 0$  such that the solution-estimator pairs  $(u_j, \zeta_j(u_j))$  of GALERKIN converge linearly, namely for all  $k \geq j \geq 0$*

$$|u - u_k|_{H_0^1(\Omega)} \leq C_* \bar{\alpha}^{k-j} |u - u_j|_{H_0^1(\Omega)}, \quad \zeta_k(u_k) \leq C_{\#} \bar{\alpha}^{k-j} \zeta_j(u_j).$$

**Corollary 5.32 (convergence of AFEM-SW)** *If the coefficients  $(A, c)$  are discrete and  $f \in H^{-1}(\Omega)$ , then for  $0 < \omega \leq \omega_0$ ,  $\xi \leq \frac{1}{2}$  as in Theorem 5.17, there exist  $0 < \bar{\alpha} < 1$  and  $C_*, C_{\#} > 0$  such that the solution - estimator pairs  $(u_j, \mathcal{E}_j(u_j, f))$  of AFEM-SW, where  $\mathcal{E}_j(u_j, f)^2 = \zeta_j(u_j)^2 + \text{osc}_j(f)_{-1}^2$ , converge linearly: for all  $k \geq j \geq 0$*

$$|u - u_k|_{H_0^1(\Omega)} \leq C_* \bar{\alpha}^{k-j} |u - u_j|_{H_0^1(\Omega)}, \quad \mathcal{E}_k(u_k, f) \leq C_{\#} \bar{\alpha}^{k-j} \mathcal{E}_j(u_j, f).$$

Both GALERKIN and AFEM-SW converge under restrictions on data  $\mathcal{D} = (A, c, f)$ . For arbitrary data  $\mathcal{D}$ , AFEM-TS concatenates GALERKIN and DATA, the later being unrelated to  $\zeta_{\mathcal{T}}(u_{\mathcal{T}})$ . Therefore, Corollary 5.31 and Proposition 5.27 (computational cost of GALERKIN) yield the following extension of Proposition 5.29 (convergence of AFEM-TS).

**Corollary 5.33 (convergence of AFEM-TS)** *The algorithm AFEM-TS driven by*



$\zeta_{\mathcal{T}}(u_{\mathcal{T}})$  stops after  $K < 2 + \frac{\log \frac{\varepsilon_0}{\text{tol}}}{\log 2}$  iterations and delivers the error

$$|u - u_K|_{H_0^1(\Omega)} \leq C_* \text{tol}.$$

The number of iterations of GALERKIN is bounded uniformly for all outer loops.

## 6 Convergence Rates of AFEM for Coercive Problems

The ultimate goal of AFEM is to produce a *quasi-best* approximation  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}$  to the solution  $u \in \mathbb{V}$  of (2.7) with error measured in  $\mathbb{V} = H_0^1(\Omega)$ . The performance of AFEM is measured by the size of the error  $|u - u_{\mathcal{T}}|_{H_0^1(\Omega)}$  relative to the cardinality  $\#\mathcal{T}$  of  $\mathcal{T}$ . The latter usually reflects the total computational cost of implementing AFEM. As a benchmark, it is useful to compare the performance of AFEM with the best approximation of  $u \in \mathbb{V}$  and  $\mathcal{D} = (A, c, f) \in \mathbb{D}$ , provided we have full knowledge of them. This is the main purpose of this section.

Under suitable assumptions on the solution  $u$  and data  $\mathcal{D}$ , we prove the existence of constants  $C(u, \mathcal{D}) > 0$  and  $s \in (0, \frac{n}{d}]$  such that

$$|u - u_{\mathcal{T}_k}|_{H_0^1(\Omega)} \leq C(u, \mathcal{D}) (\#\mathcal{T}_k)^{-s}, \quad (6.1)$$

provided  $s$  is the best decay rate with meshes in  $\mathbb{T}$  with a comparable number of degrees of freedom. The upper bound  $\frac{n}{d}$  of  $s$  is dictated by the best decay rate with polynomials of degree  $n \geq 1$  in dimension  $d$  unless  $u$  is degenerate (for instance,  $u$  belongs to a finite element space  $\mathbb{V}_{\mathcal{T}}$  with  $\mathcal{T} \in \mathbb{T}$ ). The dependence on  $\mathcal{D}$  of the constant  $C(u, \mathcal{D})$  accounts for the multiplicative structure of the interaction between the coefficients  $(A, c)$  and  $u$ , and cannot be avoided in general.

A crucial insight for the simplest scenario, the Laplacian and piecewise constant forcing  $f$ , is due to [Stevenson 2007]. It has been extended to operators with variable coefficients by [Cascón *et al.* 2008] and later expressed in terms of the estimator by [Carstensen *et al.* 2014]. It reads as follows:

*if a marking strategy reduces the PDE estimator  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$  to a fraction of its current value, then the refined set of elements  $\mathcal{R}$  inherits an error indicator  $\eta_{\mathcal{T}}(u_{\mathcal{T}}, \mathcal{R})$  comparable to  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$ , hence a Dörfler marking.* (6.2)

This allows to compare meshes produced by AFEM with optimal ones and to conclude a quasi-optimal error decay. To this end, we introduce in Section 6.1 approximation classes for functions in  $\mathbb{V}$  and  $\mathbb{D}$ , tailored to the decomposition of  $\Omega$  into conforming refinements of an initial conforming partition  $\mathcal{T}_0$ , the root of  $\mathbb{T}$ . We will assume that  $u = u(\mathcal{D}) \in \mathbb{V}$  and  $\mathcal{D} = (A, c, f) \in \mathbb{D}$  belong to these classes which, however, are not characterized in terms of regularity of  $u$  and  $\mathcal{D}$ . In Section 6.2, we investigate the approximability properties of perturbations  $\widehat{u} = u(\widehat{\mathcal{D}})$  of the exact solution  $u$ , namely exact solutions of (5.5) with perturbed data  $\widehat{\mathcal{D}}$ . Next, in Section 6.3, we consider a conforming refinement  $\mathcal{T}_* \in \mathbb{T}$  of a partition  $\mathcal{T} \in \mathbb{T}$ , and give conditions under which an optimal Dörfler marking property holds. We first apply this in Section 6.4 to study and derive rate-optimality of GALERKIN and

AFEM-SW, the one-step AFEMs. We then combine the quasi-optimal performances of GALERKIN and DATA to prove rate-optimality of the two-step AFEM in Section 6.5. We conclude in Section 6.8 upon bridging the gap between approximation and regularity classes. In particular, we give sufficient conditions for functions in Besov, Sobolev and Lipschitz spaces to belong to the approximation classes.

### 6.1 Nonlinear approximation classes

In Section 6.1.1 we discuss approximation classes for functions in  $\mathbb{V}$ , which are applicable to the solution  $u$  of (2.7). In Section 6.1.2 we turn our attention to approximation classes for functions in  $\mathbb{D}$ , which are in turn applicable to data  $\mathcal{D}$ . We refer to [ DeVore 1998 ], as well as [ DeVore and Lorentz 1993, Binev et al. 2002 ] for a discussion within nonlinear approximation theory.

#### 6.1.1 Nonlinear approximation classes for functions in $\mathbb{V}$

For any  $N \in \mathbb{N}$ ,  $N \geq \#\mathcal{T}_0$ , we define the following collection of partitions within  $\mathbb{T}$

$$\mathbb{T}_N = \{ \mathcal{T} : \mathcal{T} \in \mathbb{T} \text{ satisfies } \#\mathcal{T} \leq N \} .$$

This is the set of *conforming* meshes generated from  $\mathcal{T}_0$  with at most  $N - \#\mathcal{T}_0$  bisections. Given  $v \in \mathbb{V}$  we let  $\sigma_N(v)$  be the smallest approximation  $H_0^1$ -error incurred on  $v$  with *continuous* piecewise polynomial functions of degree  $\leq n$  over meshes  $\mathbb{T}_N$ :

$$\sigma_N(v) := \inf_{\mathcal{T} \in \mathbb{T}_N} \inf_{v_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}} |v - v_{\mathcal{T}}|_{H_0^1(\Omega)}. \quad (6.3)$$

This is a theoretical measure of performance in that finding a mesh  $\mathcal{T} \in \mathbb{T}_N$  that realizes  $\sigma_N(v)$  has exponential complexity. Proving a bound  $|v - v_{\mathcal{T}}|_{1,\Omega} \leq C_1 \sigma_{C_2 N}(v)$  for  $\mathcal{T} \in \mathbb{T}_N$  with  $C_2 \leq 1 \leq C_1$  independent of  $N$ , the so-called *instance optimality*, is rather difficult and beyond the scope of this survey. In fact, a function  $v \in \mathbb{V}_{\mathcal{T}}$  with  $\mathcal{T} \in \mathbb{T}_N$  could be the solution of our model problem (2.7), because we allow forcing  $f \in H^{-1}(\Omega)$ . Hence, we see that  $\sigma_N(v) = 0$  and AFEM should then capture  $v$  exactly on a finer mesh  $\mathcal{T} \in \mathbb{T}_{C_2^{-1}N}$ . We refer to [ Diening et al. 2016 ] for a proof of instance optimality for a forcing  $f \in L^2(\Omega)$  and the Laplace operator, namely for coefficients  $\mathbf{A} = \mathbf{I}$  and  $c = 0$ .

We will instead be able to prove that the error  $|v - v_{\mathcal{T}}|_{H_0^1(\Omega)}$  for the Galerkin solution  $v_{\mathcal{T}}$  for  $\mathcal{T} \in \mathbb{T}_N$  decays in terms of  $N$  with the same rate  $N^{-s}$  as  $\sigma_N(v)$ ; we thus say that AFEM is *rate-optimal*. We first note that for  $v \in H^{n+1}(\Omega)$  and  $\mathcal{T} \in \mathbb{T}_N$  quasi-uniform, we expect to have

$$\inf_{v_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}} |v - v_{\mathcal{T}}|_{H_0^1(\Omega)} \lesssim N^{-\frac{n}{d}} |v|_{H^{n+1}(\Omega)} \quad (6.4)$$

because the global meshsize  $h$  and  $N$  satisfy  $h \approx N^{-1/d}$ . This error estimate within the linear Sobolev scale provides the largest possible decay rate  $-n/d$ .

**Definition 6.1 (approximation class of  $u$ )** Given  $0 < s \leq n/d$ , the class  $\mathbb{A}_s :=$

$\mathbb{A}_s(H_0^1(\Omega); \mathcal{T}_0)$ , relative to the partition  $\mathcal{T}_0$  and approximation in the  $H_0^1$ -norm by continuous piecewise polynomials of degree  $\leq n$  on the forest  $\mathbb{T}$  emanating from  $\mathcal{T}_0$ , is the set of functions  $v \in \mathbb{V} = H_0^1(\Omega)$  such that

$$|v|_{\mathbb{A}_s} := \sup_{N \geq \#\mathcal{T}_0} (N^s \sigma_N(v)) < \infty \quad (6.5a)$$

whence

$$\sigma_N(v) \leq |v|_{\mathbb{A}_s} N^{-s} \quad \forall N \geq \#\mathcal{T}_0. \quad (6.5b)$$

We also write  $\mathbb{A}_s = \mathbb{A}_s^0$  to emphasize continuity of the discrete functions in  $\mathbb{V}_{\mathcal{T}} = \mathbb{S}_{\mathcal{T}}^{n,0} \cap \mathbb{V}$  with  $\mathcal{T} \in \mathbb{T}$ . The quantity  $|v|_{\mathbb{A}_s}$  is a quasi seminorm in  $\mathbb{A}_s$ , which is not a linear space but rather a nonlinear class of functions. Notice that as  $s$  increases, the cost of membership to be in  $\mathbb{A}_s$  increases, namely  $\mathbb{A}_{s_1} \subset \mathbb{A}_{s_2}$  for  $s_1 \geq s_2$ .

We may as well consider approximating  $v \in \mathbb{V}$  with *discontinuous* piecewise polynomials  $\mathbb{S}_{\mathcal{T}}^{n,-1}$  of degree  $\leq n$ , which is a richer space than  $\mathbb{S}_{\mathcal{T}}^{n,0}$ . We can likewise define the corresponding modulus of approximation

$$\sigma_N^{(-1)}(v) := \inf_{\mathcal{T} \in \mathbb{T}_N} \inf_{v_{\mathcal{T}} \in \mathbb{S}_{\mathcal{T}}^{n,-1}} |v - v_{\mathcal{T}}|_{H_0^1(\Omega; \mathcal{T})} \quad (6.6)$$

and approximation class  $\mathbb{A}_s^{-1} := \mathbb{A}_s^{-1}(H_0^1(\Omega); \mathcal{T}_0)$  of functions  $v \in H_0^1(\Omega)$  such that

$$|v|_{\mathbb{A}_s^{-1}} := \sup_{N \geq \#\mathcal{T}_0} (N^s \sigma_N^{(-1)}(v)) < \infty \quad \Rightarrow \quad \sigma_N^{(-1)}(v) \leq |v|_{\mathbb{A}_s^{-1}} N^{-s}. \quad (6.7)$$

It is obvious that  $\sigma_N^{(-1)}(v) \leq \sigma_N(v)$  for all  $v \in H_0^1(\Omega)$  because  $\mathbb{S}_{\mathcal{T}}^{n,0} \subset \mathbb{S}_{\mathcal{T}}^{n,-1}$ . However, we have the following equivalence result taken from [Veeseer 2016]. The original proof, although more complicated and for a different notion of error relevant to discontinuous Galerkin approximations, can be traced back to [Bonito and Nochetto 2010, Proposition 5.2]; see Proposition 9.4.

**Proposition 6.2 (equivalence of classes)** *Assume that all stars of meshes  $\mathcal{T} \in \mathbb{T}$  are  $(d-1)$ -face-connected. Then, there exists a constant  $C_{dG}$  that depends on the shape regularity of  $\mathbb{T}$ , the dimension  $d$  and the polynomial degree  $n \geq 1$ , such that*

$$\sigma_N(v) \leq C_{dG} \sigma_N^{(-1)}(v) \quad \forall v \in H_0^1(\Omega), \quad N \geq \#\mathcal{T}_0.$$

Moreover, the approximation classes coincide  $\mathbb{A}_s^0 = \mathbb{A}_s^{-1}$ .

*Proof.* We simply resort to (3.19) of Proposition 3.9 (approximation of gradients), namely for  $v \in H_0^1(\Omega)$

$$1 \leq \frac{\min_{w \in \mathbb{S}_{\mathcal{T}}^{n,0}} |v - w|_{H_0^1(\Omega)}}{\min_{w \in \mathbb{S}_{\mathcal{T}}^{n,-1}} |v - w|_{H_0^1(\Omega; \mathcal{T})}} \leq C_{dG},$$

and use the definitions (6.3) and (6.6). This completes the proof.  $\square$

In the rest of the paper we will make the following approximability assumption.

**Assumption 6.3 (approximability of  $u$ )** The exact solution  $u \in H_0^1(\Omega)$  of problem (2.5) belongs to the approximation class  $\mathbb{A}_s(H_0^1(\Omega); \mathcal{T}_0)$  with  $s = s_u \in (0, \frac{n}{d}]$ .

The following condition (6.8) is simpler to handle in practice than (6.5).

**Lemma 6.4 (membership in  $\mathbb{A}_s$ )** Let  $v \in \mathbb{A}_s$  and  $\varepsilon_0 := \inf_{v_{\mathcal{T}_0} \in \mathbb{V}_{\mathcal{T}_0}} |v - v_{\mathcal{T}_0}|_{H_0^1(\Omega)}$ . Then for all  $0 < \varepsilon \leq \varepsilon_0$  there exist  $\mathcal{T}_\varepsilon \in \mathbb{T}$  and  $v_\varepsilon \in \mathbb{V}_{\mathcal{T}_\varepsilon}$  such that

$$|v - v_\varepsilon|_{H_0^1(\Omega)} \leq \varepsilon, \quad \#\mathcal{T}_\varepsilon \leq 1 + |v|_{\mathbb{A}_s}^{\frac{1}{s}} \varepsilon^{-\frac{1}{s}}. \quad (6.8)$$

*Proof.* Given  $0 < \varepsilon \leq \varepsilon_0$ , let  $\mathcal{T}_\varepsilon \in \mathbb{T}$  be a conforming refinement of  $\mathcal{T}_0$  with minimal cardinality and  $v_\varepsilon \in \mathbb{V}_{\mathcal{T}_\varepsilon}$  such that

$$|v - v_\varepsilon|_{H_0^1(\Omega)} \leq \varepsilon.$$

Therefore, if  $\varepsilon < \varepsilon_0$  we deduce from the minimal property of  $\mathcal{T}_\varepsilon$  that

$$\inf_{v_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}} |v - v_{\mathcal{T}}|_{H_0^1(\Omega)} > \varepsilon \quad \forall \mathcal{T} \in \mathbb{T} : \quad \#\mathcal{T} \leq \#\mathcal{T}_\varepsilon - 1.$$

If  $N := \#\mathcal{T}_\varepsilon - 1$ , definition (6.5) implies

$$\varepsilon < \inf_{\mathcal{T} \in \mathbb{T}_N} \inf_{v_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}} |v - v_{\mathcal{T}}|_{H_0^1(\Omega)} = \sigma_N(v) \leq |v|_{\mathbb{A}_s} N^{-s},$$

whence

$$\#\mathcal{T}_\varepsilon = 1 + N \leq 1 + |v|_{\mathbb{A}_s}^{\frac{1}{s}} \varepsilon^{-\frac{1}{s}},$$

as asserted in (6.8). On the other hand, if  $\varepsilon = \varepsilon_0$  we see that

$$\varepsilon_0 \leq |v|_{\mathbb{A}_s} (\#\mathcal{T}_0)^{-s} \quad \Rightarrow \quad \#\mathcal{T}_0 \leq |v|_{\mathbb{A}_s}^{\frac{1}{s}} \varepsilon_0^{-\frac{1}{s}} < 1 + |v|_{\mathbb{A}_s}^{\frac{1}{s}} \varepsilon_0^{-\frac{1}{s}}.$$

This completes the proof.  $\square$

**Remark 6.5** If  $d = 2$  and  $n = 1$ , then Corollary 3.20 (optimal convergence rate) shows that  $W_p^2(\Omega) \subset \mathbb{A}^{1/2}$  for  $p > 1$ . The space  $W_p^2(\Omega)$  is much larger than  $H^2(\Omega)$ , fits within the nonlinear Sobolev scale, and delivers the same decay rate as (6.4). We will investigate the connection between approximation classes  $\mathbb{A}_s$  and regularity classes in any dimension  $d$  and for any polynomial degree  $n \geq 1$  later in Section 6.8.

### 6.1.2 Nonlinear approximation classes for data in $\mathbb{D}$

Given data  $\mathcal{D} = (A, c, f) \in \mathbb{D}$  and a mesh  $\mathcal{T} \in \mathbb{T}$ , we consider the best approximation of  $\mathcal{D}$  by discrete (piecewise polynomial) data  $\widehat{\mathcal{D}} = (\widehat{A}, \widehat{c}, \widehat{f}) \in \mathbb{D}_{\mathcal{T}}$ , where  $\mathbb{D}$  and  $\mathbb{D}_{\mathcal{T}}$  are defined in (5.1) and (5.2). We measure the error in the space  $D(\Omega)$  defined in (5.60) with  $q = 2$  for  $d < 4$  or  $q > \frac{d}{2}$  for  $d \geq 4$ . We now discuss the best approximation errors for the components of data in  $D(\Omega)$ , which are used to define the corresponding approximation classes. For the coefficients  $(A, c)$ , they

are characterized by the quantities

$$\delta_{\mathcal{T}}(\mathbf{A})_r := \inf_{\widehat{\mathbf{A}} \in [\mathbb{S}_{\mathcal{T}}^{n-1, -1}]^{d \times d}} \|\mathbf{A} - \widehat{\mathbf{A}}\|_{L^r(\Omega)}, \quad \delta_{\mathcal{T}}(c)_q := \inf_{\widehat{c} \in \mathbb{S}_{\mathcal{T}}^{n-1, -1}} \|c - \widehat{c}\|_{L^q(\Omega)} \quad (6.9)$$

for  $r, q \in [2, \infty]$  as above. Note that  $\widehat{\mathbf{A}}$  and  $\widehat{c}$  in (6.9) are unconstrained in the sense that they do not necessarily satisfy the structural assumption (5.51) and are thus not suited coefficient for the perturbed problem (5.5). We define the best constrained approximation errors for  $\mathbf{A} \in M(\alpha_1, \alpha_2)$  and  $c \in R(c_1, c_2)$  by

$$\begin{aligned} \widetilde{\delta}_{\mathcal{T}}(\mathbf{A})_r &:= \inf_{\widehat{\mathbf{A}} \in [\mathbb{S}_{\mathcal{T}}^{n-1, -1}]^{d \times d} \cap M(\widehat{\alpha}_1, \widehat{\alpha}_2)} \|\mathbf{A} - \widehat{\mathbf{A}}\|_{L^r(\Omega)}, \\ \widetilde{\delta}_{\mathcal{T}}(c)_q &:= \inf_{\widehat{c} \in \mathbb{S}_{\mathcal{T}}^{n-1, -1} \cap R(\widehat{c}_1, \widehat{c}_2)} \|c - \widehat{c}\|_{L^q(\Omega)}, \end{aligned} \quad (6.10)$$

where in view of (5.52)

$$\widehat{\alpha}_1 = \frac{\alpha_1}{2}, \quad \widehat{\alpha}_2 = C_{\text{ctr}} \alpha_2, \quad \widehat{c}_1 = -\frac{\alpha_1}{4C_P^2}, \quad \widehat{c}_2 = C_{\text{ctr}}(\alpha_1 + c_1). \quad (6.11)$$

We anticipate that in Section 7.4, we prove the equivalences

$$\delta_{\mathcal{T}}(\mathbf{A})_r \leq \widetilde{\delta}_{\mathcal{T}}(\mathbf{A})_r \leq C_{\text{data}} \delta_{\mathcal{T}}(\mathbf{A})_r, \quad \delta_{\mathcal{T}}(c)_q \leq \widetilde{\delta}_{\mathcal{T}}(c)_q \leq C_{\text{data}} \delta_{\mathcal{T}}(c)_q, \quad (6.12)$$

for all  $\mathbf{A} \in M(\alpha_1, \alpha_2)$  and  $c \in R(c_1, c_2)$ ; see Remarks 7.13 and 7.17. For the load function  $f$ , the definition (4.56) of  $\text{osc}_{\mathcal{T}}(f)_{-1}$  suggests to consider

$$\delta_{\mathcal{T}}(f)_{-1} := \left( \sum_{T \in \mathcal{T}} \inf_{\widehat{f} \in \mathbb{F}_{\mathcal{T}} \omega_T} \|f - \widehat{f}\|_{H^{-1}(\omega_T)}^2 \right)^{\frac{1}{2}},$$

All these best approximation errors are hard to evaluate and are thus replaced by the computable oscillations defined in (5.79) and (5.80) in practice. We recall that they rely on the local  $L^2$ -projection operator  $\Pi_{\mathcal{T}}$  for  $(\mathbf{A}, c)$  and the local  $H^{-1}$ -projection operator  $P_{\mathcal{T}}$  for  $f$  to compute linear approximations  $\widehat{\mathcal{D}}$  of  $\mathcal{D}$  to a desired accuracy. These projections are later modified nonlinearly to give rise to  $\widehat{\mathcal{D}}$  satisfying the side constraints (5.51) without compromising accuracy. We recall that the DATA module is assumed to construct approximations so that

$$\text{osc}_{\mathcal{T}}(\mathbf{A})_r \leq \Lambda_{\text{data}} \delta_{\mathcal{T}}(\mathbf{A})_r, \quad \text{osc}_{\mathcal{T}}(c)_q \leq \Lambda_{\text{data}} \delta_{\mathcal{T}}(c)_q, \quad \text{osc}_{\mathcal{T}}(f)_{-1} \leq \Lambda_{\text{data}} \delta_{\mathcal{T}}(f)_{-1}$$

with a mesh independent constant  $\Lambda_{\text{data}}$ ; see Assumption 5.22. We discuss in Section 5.4.2 practical realizations of DATA.

For the purpose of assessing the cardinality of AFEM, we do not need the specific form of  $\widehat{\mathcal{D}}$  but rather the decay of the best approximation errors in terms of degrees of freedom. Therefore, we postpone until Section 7 the construction of  $\widehat{\mathcal{D}}$  for  $n \geq 1$  and to Section 6.8 the discussion of regularity properties of  $\mathcal{D}$  that guarantee membership in the following approximation classes.

**Definition 6.6 (approximation classes of  $\mathbf{A}$ )** For  $0 < \alpha_1 \leq \alpha_2$ ,  $2 \leq r \leq \infty$ , let

$\mathbb{M}_s := \mathbb{M}_s(L^r(\Omega; \mathbb{R}^{d \times d}); \mathcal{T}_0)$  be the set of matrix-valued functions  $\mathbf{A} \in M(\alpha_1, \alpha_2)$  satisfying

$$|\mathbf{A}|_{\mathbb{M}_s} := \sup_{N \geq \#\mathcal{T}_0} \left( N^s \inf_{\mathcal{T} \in \mathbb{T}_N} \widetilde{\delta}_{\mathcal{T}}(\mathbf{A})_r \right) < \infty \Rightarrow \inf_{\mathcal{T} \in \mathbb{T}_N} \widetilde{\delta}_{\mathcal{T}}(\mathbf{A})_r \leq |\mathbf{A}|_{\mathbb{M}_s} N^{-s}. \quad (6.13)$$

**Definition 6.7 (approximation classes of  $c$ )** The class  $\mathbb{C}_s := \mathbb{C}_s(L^q(\Omega); \mathcal{T}_0)$  is the set of functions  $c \in R(c_1, c_2)$  such that

$$|c|_{\mathbb{C}_s} := \sup_{N \geq \#\mathcal{T}_0} \left( N^s \inf_{\mathcal{T} \in \mathbb{T}_N} \widetilde{\delta}_{\mathcal{T}}(c)_q \right) < \infty \Rightarrow \inf_{\mathcal{T} \in \mathbb{T}_N} \widetilde{\delta}_{\mathcal{T}}(c)_q \leq |c|_{\mathbb{C}_s} N^{-s}. \quad (6.14)$$

**Definition 6.8 (approximation classes of  $f$ )** The class  $\mathbb{F}_s := \mathbb{F}_s(H^{-1}(\Omega); \mathcal{T}_0)$  is the set of functions  $f \in H^{-1}(\Omega)$  such that

$$|f|_{\mathbb{F}_s} := \sup_{N \geq \#\mathcal{T}_0} \left( N^s \inf_{\mathcal{T} \in \mathbb{T}_N} \delta_{\mathcal{T}}(f)_{-1} \right) < \infty \Rightarrow \inf_{\mathcal{T} \in \mathbb{T}_N} \delta_{\mathcal{T}}(f)_{-1} \leq |f|_{\mathbb{F}_s} N^{-s}. \quad (6.15)$$

Since the polynomial degree of discrete coefficients  $(\widehat{\mathbf{A}}, \widehat{c})$  in definition (5.2) is  $n - 1$ , we expect decay rates  $s_A, s_c \leq n/d$  according to nonlinear approximation theory. The specific values of  $(s_A, s_c)$  depend on the regularity of  $(\mathbf{A}, c)$ , a delicate topic that we further investigate in Sections 6.8 and 7. However, because  $u$  and  $\mathcal{D} = (\mathbf{A}, c, f)$  satisfy the elliptic problem (2.5), the above approximation classes are somewhat related. We now quantify this statement.

**Lemma 6.9 (relation between approximation classes)** *Let  $2 \leq r, q \leq \infty$  be so that  $\frac{d}{2} < q$ . If  $u \in \mathbb{A}_{s_u}(H_0^1(\Omega); \mathcal{T}_0)$ ,  $\mathbf{A} \in \mathbb{M}_{s_A}(L^r(\Omega; \mathbb{R}^{d \times d}); \mathcal{T}_0)$ , and  $c \in \mathbb{C}_{s_c}(L^q(\Omega); \mathcal{T}_0)$ , with  $0 < s_u, s_A, s_c \leq \frac{n}{d}$ , then  $f \in \mathbb{F}_{s_f}(H^{-1}(\Omega); \mathcal{T}_0)$  and*

$$|f|_{\mathbb{F}_{s_f}} \leq C \left( |u|_{\mathbb{A}_{s_u}} + |\mathbf{A}|_{\mathbb{M}_{s_A}} + |c|_{\mathbb{C}_{s_c}} \right), \quad s_f = \min\{s_u, s_A, s_c\}, \quad (6.16)$$

where the constant  $C > 0$  depends  $\|u\|_{W_p^1(\Omega)}$ ,  $p = \frac{2r}{r-2}$ , and  $\alpha_1, \alpha_2, c_1, c_2$ . In particular, if  $(\mathbf{A}, c)$  are discrete in  $\mathcal{T}_0$ , then

$$|f|_{\mathbb{F}_{s_f}} \leq C |u|_{\mathbb{A}_{s_u}}, \quad s_f = s_u. \quad (6.17)$$

*Proof.* Let  $L[u] := -\operatorname{div}(\mathbf{A} \nabla u) + cu$  be the operator in (2.5) and note that  $f = L[u] \in H^{-1}(\Omega)$  can be approximated by  $\widehat{f} = -\operatorname{div}(\widehat{\mathbf{A}} \nabla \widehat{v}) + \widehat{c} \widehat{v} \in \mathbb{F}_{\mathcal{T}}$ , where the discrete space  $\mathbb{F}_{\mathcal{T}}$  is given in Definition 4.17 and  $\widehat{v} \in \mathbb{S}_{\mathcal{T}}^{n,0}$ ,  $\widehat{\mathbf{A}} \in (\mathbb{S}_{\mathcal{T}}^{n-1,-1})^{d \times d}$ ,  $\widehat{c} \in \mathbb{S}_{\mathcal{T}}^{n-1,-1}$ . Let's now express  $f - \widehat{f}$  as follows

$$f - \widehat{f} = -\operatorname{div}((\mathbf{A} - \widehat{\mathbf{A}}) \nabla u) + (c - \widehat{c})u - \operatorname{div}(\widehat{\mathbf{A}} \nabla(u - \widehat{v})) + \widehat{c}(u - \widehat{v}),$$

and recall that we have to estimate  $\|f - \widehat{f}\|_{H^{-1}(\omega_T)}$  for every  $T \in \mathcal{T}$ , rather than a global norm in  $H^{-1}(\Omega)$ , to get an upper bound on  $\delta_{\mathcal{T}}(f)_{-1}$ . Therefore, we proceed as in the proof of Lemma 5.20 (continuous dependence on data) to obtain

$$\sum_{T \in \mathcal{T}} \|f - \widehat{f}\|_{H^{-1}(\omega_T)}^2 \lesssim \|\nabla u\|_{L^p(\Omega)}^2 \|\mathbf{A} - \widehat{\mathbf{A}}\|_{L^r(\Omega)}^2 + \|\nabla u\|_{L^2(\Omega)}^2 \|c - \widehat{c}\|_{L^q(\Omega)}^2$$

$$+ \|\widehat{A}\|_{L^\infty(\Omega)}^2 \|\nabla(u - \widehat{v})\|_{L^2(\Omega)}^2 + \|\widehat{c}\|_{L^\infty(\Omega)}^2 \|u - \widehat{v}\|_{L^2(\Omega)}^2,$$

where  $p = \frac{2r}{r-1}$  and  $\|\nabla u\|_{L^p(\Omega)} < \infty$  according to (2.41) and  $\frac{d}{2} < q \leq \infty$ . Note that thanks to (6.11),  $\|\widehat{A}\|_{L^\infty(\Omega)} \leq \widehat{\alpha}_2 = C_{\text{ctr}}\alpha_2$  and  $\|\widehat{c}\|_{L^\infty(\Omega)} \leq \widehat{c}_2 = C_{\text{ctr}}(\alpha_1 + c_2)$ . Moreover, since  $\widehat{v}$ ,  $\widehat{A}$ ,  $\widehat{c}$  can be chosen separately, invoking (6.5), (6.13), (6.14), and (6.15), we realize that

$$\begin{aligned} \inf_{\mathcal{T} \in \mathbb{T}_N} \delta_{\mathcal{T}}(f)_{-1} &\lesssim \inf_{\mathcal{T} \in \mathbb{T}_N} \inf_{\widehat{v} \in \mathbb{S}_{\mathcal{T}}^{n,0}} \|\nabla(u - \widehat{v})\|_{L^2(\Omega)} \\ &+ \inf_{\mathcal{T} \in \mathbb{T}_N} \inf_{\widehat{A} \in [\mathbb{S}_{\mathcal{T}}^{n-1,-1}]^{d \times d}} \|\mathbf{A} - \widehat{A}\|_{L^r(\Omega)} \\ &+ \inf_{\mathcal{T} \in \mathbb{T}_N} \inf_{\widehat{c} \in \mathbb{S}_{\mathcal{T}}^{n-1,-1}} \|c - \widehat{c}\|_{L^q(\Omega)} \\ &\leq |u|_{\mathbb{A}_{s_u}} N^{-s_u} + |\mathbf{A}|_{\mathbb{M}_{s_A}} N^{-s_A} + |c|_{\mathbb{C}_{s_c}} N^{-s_c} \end{aligned}$$

gives (6.16) with  $s_f = \min\{s_u, s_A, s_c\}$ ; (6.17) is a trivial consequence.  $\square$

Estimate (6.17) will be useful later in Theorem 6.20 (rate-optimality of one-step AFEMs). It is important to realize that the multiplicative structure between solution  $u$  and coefficients  $(\mathbf{A}, c)$  is hidden in the constants  $C$  in (6.16) and (6.17). Moreover, these estimates are possible due to the fact that the space  $H^{-1}(\Omega)$  is the range of the linear operator  $L : H_0^1(\Omega) \rightarrow H^{-1}(\Omega)$  and that the discrete functions in  $\mathbb{F}_{\mathcal{T}}$  are images by  $L$  of functions in  $\mathbb{V}_{\mathcal{T}}$ . This would not be true for  $L^2$ -weighted surrogates of  $\delta_{\mathcal{T}}(f)_{-1}$  that typically overestimate the error in  $H^{-1}(\Omega)$ .

**Assumption 6.10 (approximability of data)** There exist  $s_A, s_c, s_f \in (0, \frac{n}{d}]$  such that data  $\mathcal{D} = (\mathbf{A}, c, f) \in \mathbb{D}$  satisfies  $\mathbf{A} \in \mathbb{M}_{s_A}$ ,  $c \in \mathbb{C}_{s_c}$ ,  $f \in \mathbb{F}_{s_f}$ .

We recall that, if  $\text{osc}_{\mathcal{T}}(\mathcal{D}) = \|\mathcal{D} - \widehat{\mathcal{D}}\|_{D(\Omega)} > C_{\text{data}}\tau$  over a conforming refinement  $\mathcal{T} \in \mathbb{T}$  of  $\mathcal{T}_0$ , then the call

$$[\widehat{\mathcal{T}}, \widehat{\mathcal{D}}] = \text{DATA}(\mathcal{T}, \mathcal{D}, \tau)$$

produces a conforming refinement  $\widehat{\mathcal{T}}$  of  $\mathcal{T}$  and approximate data  $\widehat{\mathcal{D}} = (\widehat{\mathbf{A}}, \widehat{c}, \widehat{f}) \in \mathbb{D}_{\widehat{\mathcal{T}}}$  over  $\widehat{\mathcal{T}}$  that satisfies  $\text{osc}_{\widehat{\mathcal{T}}}(\mathcal{D}) \leq \Lambda_{\text{data}}\delta_{\widehat{\mathcal{T}}}(\mathcal{D})$  and for  $r, q \in [2, \infty]$

$$\text{osc}_{\widehat{\mathcal{T}}}(\mathcal{D}) = \text{osc}_{\widehat{\mathcal{T}}}(\mathbf{A})_r + \text{osc}_{\widehat{\mathcal{T}}}(c)_q + \text{osc}_{\widehat{\mathcal{T}}}(f)_{-1} \leq C_{\text{data}}\tau,$$

as well as the constraints  $\widehat{\mathbf{A}} \in M(\widehat{\alpha}_1, \widehat{\alpha}_2)$  and  $\widehat{c} \in R(\widehat{c}_1, \widehat{c}_2)$  defined in (5.51). We will show in Section 7 that the routine responsible for reducing oscillations, namely GREEDY, exhibits optimal performance in the sense that the cardinalities  $N_{\mathcal{T}}(\mathbf{A}), N_{\mathcal{T}}(c), N_{\mathcal{T}}(f)$  of the sets of elements necessary to reduce the individual oscillations of  $(\mathbf{A}, c, f)$  below the threshold  $C_{\text{data}}\frac{\tau}{3}$  starting from any  $\mathcal{T} \geq \mathcal{T}_0$  satisfy

$$N_{\mathcal{T}}(\mathbf{A}) \lesssim |\mathbf{A}|_{\mathbb{M}_{s_A}}^{\frac{1}{s_A}} \tau^{-\frac{1}{s_A}}, \quad N_{\mathcal{T}}(c) \lesssim |c|_{\mathbb{C}_{s_c}}^{\frac{1}{s_c}} \tau^{-\frac{1}{s_c}}, \quad N_{\mathcal{T}}(f) \lesssim |f|_{\mathbb{F}_{s_f}}^{\frac{1}{s_f}} \tau^{-\frac{1}{s_f}}. \quad (6.18)$$

Therefore, the cost of one call to DATA can be quantified by the total number  $N_{\mathcal{T}}(\mathcal{D})$  of elements marked, which obeys the relation

$$\begin{aligned} N_{\mathcal{T}}(\mathcal{D}) &= N_{\mathcal{T}}(\mathbf{A}) + N_{\mathcal{T}}(c) + N_{\mathcal{T}}(f) \\ &\lesssim |\mathbf{A}|_{\mathbb{M}_{s_A}}^{\frac{1}{s_A}} \tau^{-\frac{1}{s_A}} + |c|_{\mathbb{C}_{s_c}}^{\frac{1}{s_c}} \tau^{-\frac{1}{s_c}} + |f|_{\mathbb{F}_{s_f}}^{\frac{1}{s_f}} \tau^{-\frac{1}{s_f}} \leq |\mathcal{D}|_{\mathbb{A}_{\mathcal{D}}}^{\frac{1}{s_{\mathcal{D}}}} \tau^{-\frac{1}{s_{\mathcal{D}}}} \end{aligned}$$

with

$$s_{\mathcal{D}} := \min \{s_A, s_c, s_f\}, \quad |\mathcal{D}|_{\mathbb{A}_{\mathcal{D}}} := \left( |\mathbf{A}|_{\mathbb{M}_{s_A}}^{\frac{1}{s_A}} + |c|_{\mathbb{C}_{s_c}}^{\frac{1}{s_c}} + |f|_{\mathbb{F}_{s_f}}^{\frac{1}{s_f}} \right)^{s_{\mathcal{D}}}. \quad (6.19)$$

It is thus natural to make the following assumption on DATA.

**Assumption 6.11 (quasi-optimality of DATA)** The call  $[\widehat{\mathcal{T}}, \widehat{\mathcal{D}}] = \text{DATA}(\mathcal{T}, \mathcal{D}, \tau)$ , from an arbitrary conforming refinement  $\mathcal{T}$  of  $\mathcal{T}_0$  with tolerance  $\tau$ , marks the number of elements  $N_{\mathcal{T}}(\mathcal{D})$  to produce an approximation  $\widehat{\mathcal{D}}$  of  $\mathcal{D}$  over  $\widehat{\mathcal{T}}$  so that

$$\text{osc}_{\widehat{\mathcal{T}}}(\mathcal{D}) = \|\mathcal{D} - \widehat{\mathcal{D}}\|_{D(\Omega)} \leq C_{\text{data}} \tau, \quad N_{\mathcal{T}}(\mathcal{D}) \lesssim |\mathcal{D}|_{\mathbb{A}_{\mathcal{D}}}^{\frac{1}{s_{\mathcal{D}}}} \tau^{-\frac{1}{s_{\mathcal{D}}}}. \quad (6.20)$$

## 6.2 $\varepsilon$ -approximation of order $s$

Inspection of the structure of algorithm AFEM-TS (Algorithm 5.28) reveals that the approximate data  $\mathcal{D}_k$  is fixed inside GALERKIN. Therefore, the performance of GALERKIN is dictated by the regularity of the exact solution  $\widehat{u}_k = \widehat{u}_k(\mathcal{D}_k) \in H_0^1(\Omega)$  with data  $\mathcal{D}_k$ , rather than the exact solution  $u = u(\mathcal{D})$  with data  $\mathcal{D}$ . We know that  $u \in \mathbb{A}_s$  and wonder what regularity is inherited by  $\widehat{u}_k$ . This leads to the following concept introduced in [Bonito *et al.* 2013b, Def. 3.1 and Lemma 3.2].

**Definition 6.12 ( $\varepsilon$ -approximation of order  $s$ )** Given  $u \in \mathbb{A}_s(H_0^1(\Omega); \mathcal{T}_0)$  and  $\varepsilon > 0$ , a function  $v \in H_0^1(\Omega)$  is said to be an  $\varepsilon$ -approximation of order  $s$  to  $u$  if  $|u - v|_{H_0^1(\Omega)} \leq \varepsilon$  and there exists a constant  $C > 0$  independent of  $\varepsilon$ ,  $u$  and  $v$  such that for all  $\delta \geq \varepsilon$  there exists  $N \geq \#\mathcal{T}_0$  satisfying

$$\sigma_N(v) \leq \delta \quad N \leq 1 + C |u|_{\mathbb{A}_s}^{\frac{1}{s}} \delta^{-\frac{1}{s}}. \quad (6.21)$$

**Lemma 6.13 ( $\varepsilon$ -approximation of  $u$  of order  $s$ )** Let  $u \in \mathbb{A}_s(H_0^1(\Omega); \mathcal{T}_0)$  and  $v \in H_0^1(\Omega)$  satisfy  $|u - v|_{H_0^1(\Omega)} \leq \varepsilon$  for some  $0 < \varepsilon \leq \varepsilon_0$  with  $\varepsilon_0$  defined in Lemma 6.4. Then  $v$  is a  $2\varepsilon$ -approximation of order  $s$  to  $u$ .

*Proof.* Let  $\delta \geq 2\varepsilon$ . By definition (6.3) of  $\sigma_N(v)$ , it suffices to invoke the triangle inequality to realize that

$$\sigma_N(v) \leq |u - v|_{H_0^1(\Omega)} + \sigma_N(u) \leq \frac{\delta}{2} + \sigma_N(u).$$

Since  $u \in \mathbb{A}_s(H_0^1(\Omega))$ , in view of Lemma 6.4, there exist  $N \geq \#\mathcal{T}_0$  and  $\mathcal{T} \in \mathbb{T}_N$

$$\sigma_N(u) \leq \frac{\delta}{2}, \quad N \leq 1 + |u|_{\mathbb{A}_s}^{\frac{1}{s}} \left( \frac{\delta}{2} \right)^{-\frac{1}{s}}.$$



The estimate (6.21) thus follows with constant  $C = 2^{\frac{1}{s}}$ .  $\square$

This is a simple but crucial result to study AFEM-TS. It says that any function  $v$  that is  $\varepsilon$ -close to a function  $u \in \mathbb{A}_s(X; \mathcal{T}_0)$  in the norm of the space  $X$  defining the approximation class  $\mathbb{A}_s(X; \mathcal{T}_0)$  can be approximated with a similar decay rate as  $u$  in  $X$  for as long as the desired accuracy does not exceed  $\varepsilon$ . In other words, the approximability of  $u$  is inherited by  $v$  up to scale  $\varepsilon$ . However, beyond the scale  $\varepsilon$  the approximability of  $v$  may differ from that of  $u$ . Note that neither the definition (6.3) of  $\sigma_N(v)$  nor Lemma 6.13 require  $X = H_0^1(\Omega)$ .

### 6.3 Properties of Dörfler marking

We follow the ideas of [Stevenson 2007], [Cascón *et al.* 2008] and [Carstensen *et al.* 2014] to explore the insight (6.2) about Dörfler marking. Hereafter, we recall (5.6) and consider two admissible partitions  $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$  such that  $\mathcal{T} \leq \mathcal{T}_*$ , i.e. the latter is a refinement of the former obtained by applying (newest-vertex) bisection to some of the elements of  $\mathcal{T}$ .

In the sequel, we let  $u = \widehat{u} \in H_0^1(\Omega)$  be the exact solution with discrete coefficients  $(\widehat{A}, \widehat{c})$  over a fixed mesh  $\widehat{\mathcal{T}} \leq \mathcal{T}$  and forcing function  $f$  that may or may not be discrete. We rewrite the a posteriori error estimates (5.12)

$$C_L \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f) \leq |\widehat{u} - u_{\mathcal{T}}|_{H_0^1(\Omega)} \leq C_U \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f), \quad (6.22)$$

where the total estimator  $\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f)$  consists of the PDE estimator  $\eta_{\mathcal{T}}(u_{\mathcal{T}}, f)$  and the oscillation  $\text{osc}_{\mathcal{T}}(f)_{-1}$  and reads, according to (5.11),

$$\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f)^2 = \eta_{\mathcal{T}}(u_{\mathcal{T}}, f)^2 + \text{osc}_{\mathcal{T}}(f)_{-1}^2.$$

We also recall that when  $f = P_{\mathcal{T}}f \in \mathbb{F}_{\mathcal{T}}$  is discrete,  $\text{osc}_{\mathcal{T}}(f)_{-1} = 0$  and  $\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f)$  reduces to  $\eta_{\mathcal{T}}(u_{\mathcal{T}}, f)$ , and that  $P_{\mathcal{T}}f$  is used within  $\eta_{\mathcal{T}}(u_{\mathcal{T}}, f)$  rather than  $f$ . The global nature of the elliptic boundary value problem (2.5) prevents upper a posteriori energy error estimates such as (6.22) between the continuous and discrete solution to be local. Remarkably, the situation for two Galerkin solutions  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}$  and  $u_{\mathcal{T}_*} \in \mathbb{V}_{\mathcal{T}_*}$  is different, as stated in Theorem 4.48 (upper bound for corrections)

$$|u_{\mathcal{T}_*} - u_{\mathcal{T}}|_{H_0^1(\Omega)} \leq \widetilde{C}_U \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f, \mathcal{R}), \quad (6.23)$$

where  $\mathcal{R} = \mathcal{T} \setminus \mathcal{T}_*$  is the refined set defined in (5.14). It thus turns out that  $|u_{\mathcal{T}_*} - u_{\mathcal{T}}|_{H_0^1(\Omega)}$  is controlled by the estimator  $\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f, \mathcal{R})$  on the set of elements  $\mathcal{R}$  where the meshes differ. This crucial observation goes back to Stevenson [Stevenson 2007]; see also [Cascón *et al.* 2008, Nochetto *et al.* 2009].

In the sequel we impose restrictions on the ranges of the Dörfler marking parameter (5.23) and the threshold parameter  $\omega$  for GALERKIN and AFEM-SW. We will impose a different restriction later on  $\omega$  for AFEM-TS.

**Assumption 6.14 (marking parameter)** Let  $\theta$  satisfy  $\theta \in (0, \theta_0)$  with

$$\theta_0 := \min \left\{ (2C_{\text{Lip}}\tilde{C}_U)^{-1}, 1 \right\},$$

where  $C_{\text{Lip}}, \tilde{C}_U$  are the constants in (4.68) and (6.23) respectively.

**Assumption 6.15 (restriction on  $\omega$ )** We assume  $0 \leq \omega \leq \omega_0 < 1$  with

$$\omega_0 := \sqrt{\frac{\theta_0^2 - \theta^2}{2 + \theta_0^2 - \theta^2}}$$

We now ready to make Stevenson's insight (6.2) precise.

**Lemma 6.16 (Dörfler marking)** *Let Assumptions 6.14 and 6.15 hold and  $0 < \mu \leq \frac{1}{2}$ . Let  $\mathcal{T} \in \mathbb{T}$ , and let  $\mathcal{T}_* \in \mathbb{T}$  be a refinement of  $\mathcal{T}$  with respective Galerkin solutions  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}$  and  $u_{\mathcal{T}_*} \in \mathbb{V}_{\mathcal{T}_*}$ ; let  $\mathcal{R} = \mathcal{T} \setminus \mathcal{T}_*$  be the refined set. Assume that the oscillation on  $\mathcal{T}$  is dominated by the total estimator*

$$\text{osc}_{\mathcal{T}}(f)_{-1} \leq \omega \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f) \quad (6.24)$$

and that

$$\eta_{\mathcal{T}_*}(u_{\mathcal{T}_*}, f) \leq \mu \eta_{\mathcal{T}}(u_{\mathcal{T}}, f). \quad (6.25)$$

Then Dörfler marking is valid for any  $0 < \theta < \theta_0$

$$\theta \eta_{\mathcal{T}}(u_{\mathcal{T}}, f) \leq \eta_{\mathcal{T}}(u_{\mathcal{T}}, f, \mathcal{R}). \quad (6.26)$$

*Proof.* We invoke Proposition 4.56 (estimator reduction) with  $\delta = 1$  along with the localized upper bound (6.23) to write

$$\eta_{\mathcal{T}}(u_{\mathcal{T}}, f)^2 \leq 2\eta_{\mathcal{T}_*}(u_{\mathcal{T}_*}, f)^2 + 2C_{\text{Lip}}^2 \left( \tilde{C}_U^2 \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f, \mathcal{R})^2 + \text{osc}_{\mathcal{T}}(f)_{-1}^2 \right).$$

The last term accounts for the presence of  $P_{\mathcal{T}}f$  and  $P_{\mathcal{T}_*}f$  in the definitions of  $\eta_{\mathcal{T}}(u_{\mathcal{T}}, f)$  and  $\eta_{\mathcal{T}_*}(u_{\mathcal{T}_*}, f)$ . In view of (6.24) and the definition (5.11) of the total estimator, we have

$$\text{osc}_{\mathcal{T}}(f)_{-1} \leq \sigma \eta_{\mathcal{T}}(u_{\mathcal{T}}), \quad \sigma^2 := \frac{\omega^2}{1 - \omega^2}$$

so that

$$\eta_{\mathcal{T}}(u_{\mathcal{T}}, f)^2 \leq 2\eta_{\mathcal{T}_*}(u_{\mathcal{T}_*}, f)^2 + 2C_{\text{Lip}}^2 \tilde{C}_U^2 \left( \eta_{\mathcal{T}}(u_{\mathcal{T}}, f, \mathcal{R})^2 + 2\sigma^2 \eta_{\mathcal{T}}(u_{\mathcal{T}}, f)^2 \right).$$

Using (6.25) and rearranging the above expression we obtain

$$(\theta_0^2 - 2\sigma^2) \eta_{\mathcal{T}}(u_{\mathcal{T}}, f)^2 \leq \left( \frac{1 - 2\mu^2}{2C_{\text{Lip}}^2 \tilde{C}_U^2} - 2\sigma^2 \right) \eta_{\mathcal{T}}(u_{\mathcal{T}}, f)^2 \leq \eta_{\mathcal{T}}(u_{\mathcal{T}}, f, \mathcal{R})^2$$

provided  $0 < \mu \leq \frac{1}{2}$ , because of the definition of  $\theta_0$  in Assumption 6.14. Finally,

for any  $\theta < \theta_0$  we realize that  $\omega_0$  from Assumption 6.15 satisfies

$$\sigma_0^2 := \frac{\omega_0^2}{1 - \omega_0^2} = \frac{1}{2}(\theta_0^2 - \theta^2) \quad \Rightarrow \quad \theta^2 = \theta_0^2 - 2\sigma_0^2 \leq \theta_0^2 - 2\sigma^2,$$

and Dörfler marking is valid for  $\mathcal{R}$  with parameter  $\theta$ .  $\square$

We remark that Lemma 6.16 requires that the oscillation on  $\mathcal{T}$  is dominated by the total estimator to guarantee a Dörfler marking property. This is always the case when  $f$  is discrete as in Algorithm 5.4 (GALERKIN) because in that case  $\text{osc}_{\mathcal{T}}(f)_{-1} = 0$ , or within Algorithm 5.16 (one-step AFEM with switch), which marks elements for refinement only if this property holds.

We also see that  $\theta_0$  in Assumption 6.14 corresponds to the choices  $\mu = \frac{1}{2}$  and  $\delta = 1$ . However, the proof reveals that for  $\mu \rightarrow 0$  we could obtain the largest possible value  $\theta_0 = (C_{\text{Lip}} \widetilde{C}_U)^{-1}$ , thereby the less restrictive. Since this is just twice the value of  $\theta_0$  in Assumption 6.14, the practical choice  $\mu = \frac{1}{2}$  is justified.

Lemma 6.16 hinges on two ingredients: the Lipschitz property (4.68) and the localized upper bound (6.23) of the estimator. In particular, it does not rely on the lower a posteriori error estimate in (6.22), as the original proofs in [Stevenson 2007] and [Cascón *et al.* 2008], and easily extends to discontinuous Galerkin methods in Section 9 and inf-sup stable methods in Section 10. The original statement is, however, a bit more insightful: if  $\theta_0^2 = \frac{C_2^2}{2C_1^2}$ , then for all  $0 < \theta < \theta_0$ ,  $\omega^2 \leq \theta_0^2 - \theta^2$

$$\|u - u_{\mathcal{T}_*}\|_{\Omega} \leq \mu \|u - u_{\mathcal{T}}\|_{\Omega} \quad \Rightarrow \quad \eta_{\mathcal{T}}(u_{\mathcal{T}}, \mathcal{R}) \geq \theta \eta_{\mathcal{T}}(u_{\mathcal{T}})$$

provided  $0 < \mu \leq 2^{-\frac{1}{2}}$ . We see that the threshold  $\theta_0$  is related to the gap between *reliability* constant  $C_1$  and *efficiency* constant  $C_2$  in the a posteriori bounds (5.13) in the energy norm; hence the ratio  $\frac{C_2}{C_1} \leq 1$  is a quality measure of the estimator  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$ . It is thus reasonable to be cautious about marking decisions if the constants  $C_1$  and  $C_2$  are very disparate, and thus the ratio  $C_2/C_1$  is far from 1. This justifies the constraint  $\theta < \theta_0$ .

#### 6.4 Rate-optimality of one-step AFEMs

Recall that  $\mathcal{M}_j$  is the output of the module MARK and that  $\mathcal{T}_j, u_j$  are the meshes and associated Galerkin solutions generated within Algorithms 5.4 (GALERKIN) and 5.16 (one-step AFEM with switch). To express the cardinality  $N_j(u)$  of  $\mathcal{M}_j$  in terms of  $|u - u_j|_{H_0^1(\Omega)}$  we must relate the performance of these one-step AFEMs with the approximation classes  $\mathbb{A}_s = \mathbb{A}_s(H_0^1(\Omega); \mathcal{T}_0)$  for  $u$  and  $\mathbb{F}_s = \mathbb{F}_s(H^{-1}(\Omega); \mathcal{T}_0)$  for  $f$ , which are never used in the design of these algorithms. Even though this might appear like an undoable task, the key to unravel this connection is given by Lemma 6.16 (Dörfler marking) and the following assumption.

**Assumption 6.17 (cardinality of  $\mathcal{M}$ )** The module MARK selects a set  $\mathcal{M}$  in (5.23) with *minimal* cardinality.

According to the equidistribution principle (3.24) and the local lower bound (4.54) in the proof of Theorem 4.45 (modified residual estimator) for discrete coefficients, i.e.,

$$C_L \eta_{\mathcal{T}_j}(u_j, T) \leq C_L \mathcal{E}_{\mathcal{T}_j}(u_j, f, T) \leq |u - u_j|_{H^1(\omega_T)},$$

it is natural to mark elements with largest error indicators. This explains the choice of a minimal set  $\mathcal{M}$  in Assumption 6.17.

We are now ready to bound the cardinality of  $\mathcal{M}_j$  in terms of  $|u - u_j|_{H_0^1(\Omega)}$ .

**Proposition 6.18 (cardinality of  $\mathcal{M}_j$ )** *Let Assumptions 6.14, 6.15 and 6.17 be valid. If  $u \in \mathbb{A}_s$  and*

$$\text{osc}_{\mathcal{T}_j}(f)_{-1} \leq \omega \mathcal{E}_{\mathcal{T}_j}(u_j, f),$$

*then the cardinality  $N_j(u)$  of  $\mathcal{M}_j$  satisfies*

$$N_j(u) \lesssim |u|_{\mathbb{A}_s}^{\frac{1}{s}} |u - u_j|_{H_0^1(\Omega)}^{-\frac{1}{s}} \quad \forall j \geq 0.$$

*Proof.* Let  $C_{\text{cea}} := \sqrt{\frac{C_{\mathcal{B}}}{c_{\mathcal{B}}}}$  and let  $\delta = \mu \frac{C_L}{C_{\text{cea}}} \eta_j(u_j)$  with  $\mu \leq \frac{1}{2}$  the quasi-monotonicity constant in (3.8). We invoke (6.8) for  $u \in \mathbb{A}_s$  to find a mesh  $\mathcal{T}_{\delta} \in \mathbb{T}$  and a Galerkin solution  $u_{\delta} \in \mathbb{V}_{\mathcal{T}_{\delta}}$  so that

$$|u - u_{\mathcal{T}_{\delta}}|_{H_0^1(\Omega)} \leq \delta, \quad \#\mathcal{T}_{\delta} \lesssim |u|_{\mathbb{A}_s}^{\frac{1}{s}} \delta^{-\frac{1}{s}}.$$

Since  $\mathcal{T}_{\delta}$  may be totally unrelated to  $\mathcal{T}_j$ , we introduce the overlay  $\mathcal{T}_* = \mathcal{T}_j \oplus \mathcal{T}_{\delta}$ . We exploit that  $\mathcal{T}_* \geq \mathcal{T}_{\delta}$ , hence the space nestedness  $\mathbb{V}_{\mathcal{T}_{\delta}} \subset \mathbb{V}_{\mathcal{T}_*}$ , along with the property that the Galerkin solution  $u_{\mathcal{T}_*} \in \mathbb{V}_{\mathcal{T}_*}$  minimizes the energy error in  $\mathbb{V}_{\mathcal{T}_*}$

$$\eta_{\mathcal{T}_*}(u_{\mathcal{T}_*}) \leq \frac{1}{C_L} |u - u_{\mathcal{T}_*}|_{H_0^1(\Omega)} \leq \frac{C_{\text{cea}}}{C_L} |u - u_{\mathcal{T}_{\delta}}|_{H_0^1(\Omega)} \leq \frac{C_{\text{cea}}}{C_L} \delta = \mu \eta_j(u_j),$$

because of the lower bound in (5.12) and (3.8). Therefore, Lemma 6.16 (Dörfler marking) implies that the refined set  $\mathcal{R} = \mathcal{T} \setminus \mathcal{T}_*$  satisfies Dörfler marking with parameter  $\theta < \theta_0$ . Since MARK delivers a minimal set  $\mathcal{M}_j$  with this property, according to Assumption 6.17, we deduce

$$N_j(u) = \#\mathcal{M}_j \leq \#\mathcal{R} \leq \#\mathcal{T}_* - \#\mathcal{T} \leq \#\mathcal{T}_{\varepsilon} - \#\mathcal{T}_0 \lesssim |u|_{\mathbb{A}_s}^{\frac{1}{s}} \delta^{-\frac{1}{s}},$$

where we have used Lemma 3.17 (mesh overlay). The assertion follows from  $\text{osc}_j(f)_{-1} \leq \omega \mathcal{E}_j(u_j, f)$  in Assumption 6.15 and the upper bound in (5.12)

$$|u - u_j|_{H_0^1(\Omega)} \leq C_U \mathcal{E}_j(u_j, f) \leq \frac{C_U}{\sqrt{1 - \omega^2}} \eta_j(u_j),$$

and completes the proof.  $\square$

We next prove rate optimality of the one-step AFEMs of Algorithm 5.4 and Algorithm 5.16. To this end, we need an additional assumption.

**Assumption 6.19 (initial labeling)** If the initial mesh  $\mathcal{T}_0$  is made of simplices, then let the initial labeling (3.35) for  $d = 2$ , or that of [Stevenson 2008, Section 4] for  $d > 2$ , be valid.

This assumption ensures the validity of Theorem 3.16 (complexity of REFINES): if  $\mathcal{M}_j \subset \mathcal{T}_j$  is a set of marked elements for a sequence  $\{\mathcal{T}_j\}_{j=0}^{k-1}$  of consecutive refinements of  $\mathcal{T}_0$ , then the cardinality of the  $k$ -th mesh satisfies

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \leq D \sum_{j=0}^{k-1} \#\mathcal{M}_j, \quad (6.27)$$

with a universal constant  $D$  depending only on  $\mathcal{T}_0$  and  $d$ . We always assume that  $\#\mathcal{T}_k \geq \frac{3}{2}\#\mathcal{T}_0$ , whence  $\#\mathcal{T}_k - \#\mathcal{T}_0 \geq \frac{1}{3}\#\mathcal{T}_k$  and, if  $\tilde{D} = 3D$ , (6.27) reads instead

$$\#\mathcal{T}_k \leq \tilde{D} \sum_{j=0}^{k-1} \#\mathcal{M}_j, \quad (6.28)$$

**Theorem 6.20 (rate optimality of one-step AFEMs)** For Algorithms 5.4 (with  $\mathcal{T} = \mathcal{T}_0$ ) and 5.16, let Assumptions 6.14, 6.17 and 6.19 be valid, and in addition let the parameter  $\omega > 0$  satisfy Assumption 6.15 for Algorithm 5.16. If  $u \in \mathbb{A}_s$ , then both one-step AFEMs give rise to sequences  $\{\mathcal{T}_k, \mathbb{V}_k, u_k\}_{k=0}^\infty$  such that

$$|u - u_k|_{H_0^1(\Omega)} \lesssim |u|_{\mathbb{A}_s} (\#\mathcal{T}_k)^{-s}. \quad (6.29)$$

*Proof.* We consider first Algorithm 5.4, for which the forcing  $f \in \mathbb{F}_{\mathcal{T}_0}$  is discrete, whence  $\text{osc}_j(f)_{-1} = 0$  for all  $j \geq 0$  and  $\omega = \sigma = 0$  in Assumption 6.15. In view of (6.28), we apply Proposition 6.18 (cardinality of  $\mathcal{M}_j$ ) to infer that

$$\#\mathcal{T}_k \leq \tilde{D} \sum_{j=0}^{k-1} \#\mathcal{M}_j \lesssim |u|_{\mathbb{A}_s}^{\frac{1}{s}} \sum_{j=0}^{k-1} |u - u_j|_{H_0^1(\Omega)}^{-\frac{1}{s}}.$$

We now recall the inequality  $|u - u_k|_{H_0^1(\Omega)} \leq C_* \alpha^{k-j} |u - u_j|_{H_0^1(\Omega)}$  from Corollary 5.6 (linear convergence), and replace the sum above by

$$\sum_{j=0}^{k-1} |u - u_j|_{H_0^1(\Omega)}^{-\frac{1}{s}} \leq |u - u_k|_{H_0^1(\Omega)}^{-\frac{1}{s}} \sum_{j=0}^k \alpha^{\frac{k-j}{s}} \leq \frac{\alpha^{\frac{1}{s}}}{1 - \alpha^{\frac{1}{s}}} |u - u_k|_{H_0^1(\Omega)}^{-\frac{1}{s}}$$

because  $0 < \alpha < 1$  and the geometric series is summable.

We next deal with Algorithm 5.16. If the algorithm calls MARK, then  $\text{osc}_j(f)_{-1} \leq \omega \mathcal{E}_j(u_j, f)$  and the number of marked elements  $N_j(u)$  obeys Proposition 6.18

$$N_j(u) \lesssim |u|_{\mathbb{A}_s}^{\frac{1}{s}} |u - u_j|_{H_0^1(\Omega)}^{-\frac{1}{s}}.$$

Instead, if the algorithm calls DATA, then  $\text{osc}_j(f)_{-1} > \sigma_j = \omega \mathcal{E}_j(u_j, f)$  and DATA returns a mesh  $\mathcal{T}_{j+1}$  and reduces the oscillation  $\text{osc}_{j+1}(f)_{-1} \leq \sigma_j$  with optimal complexity. To quantify the cost, we recall that  $u \in \mathbb{A}_s$  yields  $f \in \mathbb{F}_s$  according to

Lemma 6.9 (relation between approximation classes) and  $|f|_{\mathbb{F}_s} \lesssim |u|_{\mathbb{A}_s}$ . Therefore, the number of marked elements  $N_j(f)$  to reduce  $\text{osc}_j(f)_{-1}$  to tolerance  $\sigma_j$  satisfies

$$N_j(f) \lesssim |f|_{\mathbb{F}_s}^{\frac{1}{s}} \sigma_j^{-\frac{1}{s}} \lesssim |u|_{\mathbb{A}_s}^{\frac{1}{s}} \mathcal{E}_j(u_j, f)^{-\frac{1}{s}} \lesssim |u|_{\mathbb{A}_s}^{\frac{1}{s}} |u - u_j|_{H_0^1(\Omega)}^{-\frac{1}{s}},$$

because of (5.12). It thus remains to sum over  $j$ , apply again (6.28)

$$\#\mathcal{T}_k \leq \tilde{D} \sum_{j=0}^{k-1} (N_j(u) + N_j(f)) \lesssim |u|_{\mathbb{A}_s}^{1/s} \sum_{j=0}^{k-1} |u - u_j|_{H_0^1(\Omega)}^{-\frac{1}{s}},$$

and finally argue as before with the help of Corollary 5.18 (linear convergence of error).  $\square$

### 6.5 Rate-optimality of two-Step AFEM

The output of  $[\widehat{\mathcal{T}}_k, \widehat{\mathcal{D}}_k] = \text{DATA}(\mathcal{T}_k, \mathcal{D}, \omega \varepsilon_k)$ , in the  $k$ -step of AFEM-TS (Algorithm 5.28), is fed to  $[\mathcal{T}_{k+1}, u_{k+1}] = \text{GALERKIN}(\widehat{\mathcal{T}}_k, \widehat{\mathcal{D}}_k, \varepsilon_k)$ , which in turn iterates  $J_k$  times. We denote by  $(\mathcal{T}_{k,j}, \mathcal{M}_{k,j}, u_{\mathcal{T}_{k,j}})$  the triplets of grids, marked sets and discrete solutions computed within  $\text{GALERKIN}(\widehat{\mathcal{T}}_k, \widehat{\mathcal{D}}_k, \varepsilon_k)$  for  $0 \leq j < J_k$ . We further assume that

$$\widehat{\varepsilon}_k := \eta_{\mathcal{T}_{k,0}}(u_{\mathcal{T}_{k,0}}, \widehat{\mathcal{D}}_k) > \varepsilon_k$$

for otherwise the module GALERKIN is skipped. In view of the lower a posteriori error estimate in (6.22) for discrete data  $\widehat{\mathcal{D}}_k$ , we infer that

$$|\widehat{u}_k - u_{\mathcal{T}_{k,0}}|_{H_0^1(\Omega)} \geq C_L \widehat{\varepsilon}_k > C_L \varepsilon_k,$$

where  $\widehat{u}_k \in H_0^1(\Omega)$  is the exact solution of (5.5) with approximate data  $\widehat{\mathcal{D}}_k$ . The module DATA guarantees (5.63) and (5.65), namely

$$\|\mathcal{D} - \widehat{\mathcal{D}}_k\|_{D(\Omega)} \leq C_{\text{data}} \omega \varepsilon_k \quad \Rightarrow \quad |u - \widehat{u}_k|_{H_0^1(\Omega)} \leq C_D \omega \varepsilon_k, \quad (6.30)$$

where  $u = u(\mathcal{D}) \in H_0^1(\Omega)$  is the exact solution of (2.7). We see that the parameter  $\omega$  controls the discrepancy between  $u$  and  $\widehat{u}_k = \widehat{u}_k(\mathcal{D}_k)$  relative to  $\varepsilon_k$ . We now make an assumption on the appropriate size of  $\omega$ , which replaces Assumption 6.15 for AFEM-SW.

**Assumption 6.21 (size of  $\omega$ )** The parameter  $\omega$  in AFEM-TS satisfies  $\omega \in (0, \omega_0]$  where  $\omega_0 := \frac{\mu C_L}{2C_D C_{\text{Cea}}}$  with  $C_{\text{cea}}$  as in (3.8) and the parameter  $0 < \mu \leq \frac{1}{2}$  appears in Lemma 6.16 (Dörfler marking).

Consequently, if Assumption 6.21 is valid then (6.30) yields for  $\omega \leq \omega_0$

$$|u - \widehat{u}_k|_{H_0^1(\Omega)} \leq \frac{\mu C_L}{2C_{\text{Cea}}} \varepsilon_k. \quad (6.31)$$

**Corollary 6.22 (cardinality of marked sets)** Let Assumptions 6.14, 6.17, and 6.21

hold. If  $u \in \mathbb{A}_s(H_0^1(\Omega); \mathcal{T}_0)$  and  $\widehat{\varepsilon}_k > \varepsilon_k$ , then GALERKIN is called and there exists a constant  $C_0$  such that for all  $0 \leq j < J_k$

$$\#\mathcal{M}_{k,j} \leq C_0 |u|_{\mathbb{A}_s}^{1/s} |u - u_{\mathcal{T}_{k,j}}|_{H_0^1(\Omega)}^{-1/s}, \quad (6.32)$$

and

$$\#\mathcal{M}_{k,j} \leq C_0 |u|_{\mathbb{A}_s}^{1/s} \varepsilon_k^{-1/s}. \quad (6.33)$$

*Proof.* We argue as in Proposition 6.18. Fix  $0 \leq j < J_k$  and set

$$\delta := \mu \frac{C_L}{C_{\text{Cea}}} \eta_{\mathcal{T}_{k,j}}(u_{\mathcal{T}_{k,j}}) \quad \Rightarrow \quad \delta \geq \mu \frac{C_L}{C_{\text{Cea}}} \varepsilon_k.$$

Since  $|u - \widehat{u}_k|_{H_0^1(\Omega)} \leq \frac{\delta}{2}$ , by virtue of (6.31), we deduce that  $\widehat{u}_k$  is an  $\delta$ -approximation of order  $s$  to  $u$  according to Lemma 6.13 ( $\varepsilon$ -approximation of order  $s$ ). Therefore, there exists an admissible mesh  $\mathcal{T}_\delta \in \mathbb{T}$  such that

$$|\widehat{u}_k - u_{\mathcal{T}_\delta}|_{H_0^1(\Omega)} \leq \delta, \quad \#\mathcal{T}_\delta \lesssim |u|_{\mathbb{A}_s}^{\frac{1}{s}} \delta^{-\frac{1}{s}},$$

and we proceed exactly as in Proposition 6.18 to show that

$$\#\mathcal{M}_{k,j} \lesssim |u|_{\mathbb{A}_s}^{\frac{1}{s}} \delta^{-\frac{1}{s}} \approx |u|_{\mathbb{A}_s}^{\frac{1}{s}} |u - u_{\mathcal{T}_{k,j}}|_{H_0^1(\Omega)}^{-1/s} \lesssim |u|_{\mathbb{A}_s}^{\frac{1}{s}} \varepsilon_k^{-\frac{1}{s}}.$$

This concludes the proof.  $\square$

**Corollary 6.23 (quasi-optimality of GALERKIN)** *Let Assumptions 6.3, 6.14, 6.17, and 6.21 be valid. Then, the number of marked elements  $N_k(u)$  within the  $k$ -th call to GALERKIN satisfies*

$$N_k(u) \leq J C_0 |u|_{\mathbb{A}_s}^{\frac{1}{s}} \varepsilon_k^{-\frac{1}{s}},$$

where  $J \geq J_k$  is a uniform upper bound for the number of iterations of GALERKIN.

*Proof.* Use  $N_k(u) = \sum_{j=0}^{J_k-1} \#\mathcal{M}_{k,j}$  and combine Corollary 6.22 (cardinality of marked sets) and Proposition 5.27 (computational cost of GALERKIN).  $\square$

We finally address the rate-optimality of the two-step algorithm AFEM-TS, by proving the announced bound (6.1).

**Theorem 6.24 (rate-optimality of AFEM-TS)** *Let Assumptions 6.3 (approximability of  $u$ ), 6.10 (approximability of data), 6.11 (quasi-optimality of the module DATA), 6.14 (marking parameter), 6.21 (size of  $\omega$ ), and 6.19 (initial labeling) be valid. Then, AFEM-TS gives rise to a sequence  $(\mathcal{T}_k, \mathbb{V}_{\mathcal{T}_k}, u_{\mathcal{T}_k})_{k=0}^K$  such that*

$$|u - u_k|_{H_0^1(\Omega)} \leq C(u, \mathcal{D}) (\#\mathcal{T}_k)^{-s} \quad 1 \leq k \leq K,$$

where  $0 < s = \min\{s_u, s_{\mathcal{D}}\} = \min\{s_u, s_A, s_c, s_f\} \leq \frac{n}{d}$  and

$$C(u, \mathcal{D}) = C_* \left( |u|_{\mathbb{A}_{su}}^{\frac{1}{s_u}} + |A|_{\mathbb{M}_{s_A}}^{\frac{1}{s_A}} + |c|_{\mathbb{C}_{s_c}}^{\frac{1}{s_c}} + |f|_{\mathbb{F}_{s_f}}^{\frac{1}{s_f}} \right)^s$$

with constant  $C_* > 0$  independent of  $u$  and  $\mathcal{D}$ .

*Proof.* In view of Assumption 6.3, Corollary 6.23 implies that the number of marked elements  $N_k(u)$  within the  $(k + 1)$ -th call to GALERKIN satisfies

$$N_k(u) \leq C_3 |u|_{\mathbb{A}_{s_u}}^{\frac{1}{s_u}} \varepsilon_k^{-\frac{1}{s_u}}$$

with  $s_u \leq \frac{n}{d}$  and  $C_3 > 0$  a suitable constant. Moreover, by Assumption 6.11 the number of marked elements  $N_k(\mathcal{D})$  within the  $(k + 1)$ -th call to DATA satisfies

$$N_k(\mathcal{D}) \leq C_3 |\mathcal{D}|_{\mathbb{A}_{s_{\mathcal{D}}}}^{\frac{1}{s_{\mathcal{D}}}} \varepsilon_k^{-\frac{1}{s_{\mathcal{D}}}}$$

with  $s_{\mathcal{D}} \leq \frac{n}{d}$ . The total number of marked elements in the  $(k + 1)$ -th loop of AFEM-TS is thus

$$N_k(\mathcal{D}) + N_k(u) \leq C_3 \left( |u|_{\mathbb{A}_{s_u}}^{\frac{1}{s_u}} + |\mathcal{D}|_{\mathbb{A}_{s_{\mathcal{D}}}}^{\frac{1}{s_{\mathcal{D}}}} \right) \varepsilon_k^{-\frac{1}{s}}.$$

Upon termination, DATA and GALERKIN give

$$|u - \widehat{u}_k|_{H_0^1(\Omega)} \leq \frac{\mu C_L}{2C_{\text{Cea}}} \varepsilon_k \leq \frac{C_L}{4C_{\text{Cea}}} \varepsilon_k,$$

$$|\widehat{u}_k - u_{k+1}|_{H_0^1(\Omega)} \leq C_U \eta_{\mathcal{T}_{k+1}}(u_{k+1}, \widehat{\mathcal{D}}_k) \leq C_U \varepsilon_k,$$

because of (6.31), (6.22) and the fact that  $\mu < \frac{1}{2}$ . This implies by triangle inequality

$$|u - u_{k+1}|_{H_0^1(\Omega)} \leq \left( \frac{C_L}{4C_{\text{Cea}}} + C_U \right) \varepsilon_k = C_4 \varepsilon_k.$$

Therefore, applying Theorem 3.16 (complexity of REFINE), the total amount of elements created by  $k + 1$  iterations within AFEM-TS, besides those in  $\mathcal{T}_0$ , obeys the expression

$$\#\mathcal{T}_{k+1} \leq \widetilde{D} \sum_{j=0}^k (N_j(\mathcal{D}) + N_j(u)) \leq \widetilde{D} C_3 \left( |u|_{\mathbb{A}_{s_u}}^{\frac{1}{s_u}} + |\mathcal{D}|_{\mathbb{A}_{s_{\mathcal{D}}}}^{\frac{1}{s_{\mathcal{D}}}} \right) \sum_{j=0}^k \varepsilon_j^{-\frac{1}{s}}.$$

according to (6.28). Since  $\varepsilon_j = 2^{-j} \varepsilon_0$  and  $\sum_{j=0}^{k-1} (2^{-\frac{1}{s}})^j \leq \frac{1}{1-2^{-1/s}}$  we obtain

$$\#\mathcal{T}_{k+1} \leq C \left( |u|_{\mathbb{A}_{s_u}}^{\frac{1}{s_u}} + |\mathcal{D}|_{\mathbb{A}_{s_{\mathcal{D}}}}^{\frac{1}{s_{\mathcal{D}}}} \right) \varepsilon_k^{-\frac{1}{s}}$$

with  $C = \frac{\widetilde{D} C_3 \varepsilon_0}{1-2^{-1/s}}$  provided  $\#\mathcal{T}_{k+1} \geq \frac{3}{2} \#\mathcal{T}_0$ . This together with  $|u - u_{k+1}|_{H_0^1(\Omega)} \leq C_4 \varepsilon_k$  gives the asserted estimate after  $1 \leq k + 1 \leq K$  loops.  $\square$

**Remark 6.25** The thresholds  $\theta_0, \omega_0$  play no role in Proposition 5.29 (convergence of AFEM-TS) but are critical in Theorem 6.24 (rate-optimality of AFEM-TS). The former takes care of the discrepancy between error and estimator [Stevenson 2007, Cascón *et al.* 2008, Nochetto *et al.* 2009, Bonito and Nochetto 2010, Nochetto



and Veeseer 2012]. The latter guarantees that the perturbation error (6.30) is much smaller than  $\varepsilon_k$  and enables GALERKIN to learn the regularity of  $u$  from  $\widehat{u}_{\widehat{\mathcal{T}}_k}$  [Stevenson 2007, Bonito *et al.* 2013b].

**Remark 6.26** We claim that the convergence rate  $s = \min\{s_u, s_{\mathcal{D}}\}$  cannot be improved to  $s_u$  (the optimal rate for approximations of  $u \in \mathbb{A}_{s_u}(H_0^1(\Omega); \mathcal{T}_0)$ ) when  $s_{\mathcal{D}} < s_u$  by any algorithm that uses approximations  $\widehat{\mathcal{D}} = (\widehat{A}, \widehat{c}, \widehat{f})$  of data  $\mathcal{D} = (A, c, f)$ . In fact, given any  $\delta > 0$  consider the ball

$$B(\mathcal{D}, \delta) := \{\widehat{\mathcal{D}} \in \mathbb{D} : \|\mathcal{D} - \widehat{\mathcal{D}}\|_{D(\Omega)} \leq \delta\}, \quad (6.34)$$

where  $D(\Omega)$  is defined in (5.60). If  $u, \widehat{u} \in H_0^1(\Omega)$  are the exact solutions for data  $\mathcal{D}, \widehat{\mathcal{D}}$ , then there are constants  $0 < c_* \leq C_*$  such that

$$c_*\delta \leq \sup_{\widehat{\mathcal{D}} \in B(\mathcal{D}, \delta)} |u - \widehat{u}|_{H_0^1(\Omega)} \leq C_*\delta.$$

The rightmost inequality is a consequence of Lemma 5.20 (continuous dependence on data). For the leftmost inequality, consider first a perturbation  $\widehat{f} = (1 + \delta)f$  of the source term with coefficients  $(\widehat{A}, \widehat{c}) = (A, c)$ , whence  $\|\mathcal{D} - \widehat{\mathcal{D}}\|_{D(\Omega)} = \delta$ . Proceeding as in (2.30), the coercivity and continuity of the bilinear form  $\mathcal{B}$  imply

$$c_{\mathcal{B}} |u - \widehat{u}|_{H_0^1(\Omega)} \leq \|f - \widehat{f}\|_{H^{-1}(\Omega)} = \delta \leq C_{\mathcal{B}} |u - \widehat{u}|_{H_0^1(\Omega)}.$$

On the other hand, if  $\widehat{f} = f$  and  $(\widehat{A}, \widehat{c}) = \frac{1}{\alpha}(A, c)$  with  $\alpha = 1 + \frac{\delta}{\|\mathcal{D}\|_{D(\Omega)}}$ , then

$$\|\mathcal{D} - \widehat{\mathcal{D}}\|_{D(\Omega)} < \delta, \quad |u - \widehat{u}|_{H_0^1(\Omega)} = \frac{|u|_{H_0^1(\Omega)}}{\|\mathcal{D}\|_{D(\Omega)}} \delta \geq \frac{\|f\|_{H^{-1}(\Omega)}}{C_{\mathcal{B}} \|\mathcal{D}\|_{D(\Omega)}} \delta.$$

This argument takes care of the multiplicative nature of  $(A, c)$  in (2.5), which makes  $\widehat{u} = \alpha u$ , and proves our claim.

## 6.6 Rate-optimality of AFEM with other boundary conditions

The key ingredient for rate-optimality of AFEM, regardless of boundary conditions, is the validity of Lemma 6.16 (Dörfler marking). This lemma provides a bridge between FEM meshes and optimal meshes and, in turn, hinges on three properties of the PDE residual estimator  $\eta_{\mathcal{T}}(u_{\mathcal{T}}, f)$ : Theorem 4.48 (upper bound for corrections), Lemma 4.54 (Lipschitz property of the estimator), and Lemma 4.55 (estimator dependence on discrete forcing) to account for the possible change in the discrete forcing  $P_{\mathcal{T}}f$ . Since their proofs are insensitive to boundary conditions, because they do not alter the structure of  $\eta_{\mathcal{T}}(u_{\mathcal{T}}, f)$ , we conclude their validity as well as for Robin, Neumann and non-homogeneous Dirichlet conditions.

Therefore, our three AFEMs based on Dörfler marking deliver the same asymptotic convergence rates associated with the approximations classes  $\mathbb{A}_{s_u}$  for the solution  $u \in H^1(\Omega)$  and  $\mathbb{A}_{s_{\mathcal{D}}}$  for data  $\mathcal{D} = (A, c, p, \ell)$  for Robin and Neumann boundary conditions and  $\mathcal{D} = (A, c, f, g)$  for Dirichlet boundary conditions. We need three

new approximation classes for  $p \in \mathbb{P}_{s_p}(L^\infty(\partial\Omega); \mathcal{T}_0)$ ,  $\ell \in \mathbb{L}_{s_\ell}(H^1(\Omega)^*; \mathcal{T}_0)$  for Robin or Neumann conditions and  $g \in \mathbb{G}_{s_g}(H^{1/2}(\partial\Omega); \mathcal{T}_0)$ .

If coefficients  $(A, c, p)$  are discrete for the Robin condition, then Lemma 6.9 (relation between approximation classes) extends and yields

$$\langle \ell - \widehat{\ell}, v \rangle = \mathcal{B}[u - \widehat{u}, v] \quad \Rightarrow \quad \|\ell - \widehat{\ell}\|_{\mathbb{V}^*} \leq C \|u - \widehat{u}\|_{\mathbb{V}},$$

with  $\mathbb{V} = H^1(\Omega)$  and  $\mathcal{B} = \widehat{\mathcal{B}}, \widehat{\ell}$  given in (5.85). This in turn implies  $|\ell|_{\mathbb{L}_{s_\ell}} \leq C |u|_{\mathbb{A}_{s_u}}$ ,  $s_\ell = s_u$  and the validity of Theorem 6.20 (rate optimality of one-step AFEMs). In AFEM-TS, DATA approximates  $\ell$  along with the other data and Theorem 6.24 (rate-optimality of AFEM-TS) is also valid for Robin and Neumann boundary conditions. We do not explore this matter any longer.

For non-homogeneous Dirichlet boundary conditions the analysis is simpler. If  $g$  is discrete, then there is no difference with  $g = 0$ . If not, we note that the solution map  $g \mapsto u$  (all other data being fixed) is affine and that the error and augmented total estimator  $\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f, g) := \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f) + \text{osc}_{\mathcal{T}}(g)_{1/2}$  are equivalent (Theorem 4.74). This indicates that the role of  $g$  is similar to the role of  $f$ . Therefore, it suffices to replace  $\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f)$  by  $\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f, g)$  and  $\text{osc}_{\mathcal{T}}(f)_{-1}$  by  $\text{osc}_{\mathcal{T}}(f)_{-1} + \text{osc}_{\mathcal{T}}(g)_{1/2}$  in AFEM-SW. For AFEM-TS, the approximation of  $g$  is handled by DATA along with the other data. Hence, we again conclude that Theorems 6.20 and 6.24 extend to non-vanishing Dirichlet conditions.

### 6.7 Rate-optimality of AFEM driven by alternative estimators

We recall the notation  $\zeta_{\mathcal{T}}(u_{\mathcal{T}})$  of Section 5.6 for any of the three alternative estimators in Section 4.9 and the crucial local properties (5.86) and (5.87).

As already alluded to in Section 6.6, the key instrument for rate-optimality is Lemma 6.16 (Dörfler marking). We now check the validity of its three main pillars: Theorem 4.48 (upper bound for corrections) and Lemma 4.54 (Lipschitz property of the estimator) and Lemma 4.55 (estimator dependence on discrete forcing) to account for possible change in the discrete forcing  $P_{\mathcal{T}}f$ . It turns out that if they were valid for  $\zeta_{\mathcal{T}}(u_{\mathcal{T}}) = \zeta_{\mathcal{T}}(u_{\mathcal{T}}, f)$ , then statements about rates of convergence similar to those for  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$  would follow for  $\zeta_{\mathcal{T}}(u_{\mathcal{T}})$ .

**Lemma 6.27 (localized discrete upper bound)** *Let  $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$  and  $\mathcal{T}_*$  be a refinement of  $\mathcal{T}$ . Let the coefficients  $(\widehat{A}, \widehat{c})$  be discrete over  $\mathcal{T}$  and  $f \in H^{-1}(\Omega)$ . Then the error between the corresponding Galerkin solutions  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}$  and  $u_{\mathcal{T}_*} \in \mathbb{V}_{\mathcal{T}_*}$  is bounded by the indicator in the refined set  $\overline{\mathcal{R}}$  plus data oscillation*

$$|u_{\mathcal{T}} - u_{\mathcal{T}_*}|_{H_0^1(\Omega)} \leq \overline{C}_U \left( (C_L^{eq})^{-1} \zeta_{\mathcal{T}}(u_{\mathcal{T}}, \overline{\mathcal{R}})^2 + \text{osc}_{\mathcal{T}}(f)_{-1}^2 \right)^{\frac{1}{2}},$$

where  $\overline{\mathcal{R}} := \{z \in \mathcal{V} : T \in \mathcal{T} \setminus \mathcal{T}_*, T \subset \omega_z\}$  collects all vertices whose associated stars change from  $\mathcal{T}$  to  $\mathcal{T}_*$ .

*Proof.* It suffices to realize that  $\mathcal{T} \setminus \mathcal{T}_* \subset \bigcup \{\omega_z : z \in \overline{\mathcal{R}}\}$ , and appeal to Theorem

4.48 and (5.87) to arrive at

$$\begin{aligned} |u_{\mathcal{T}} - u_{\mathcal{T}^*}|_{H_0^1(\Omega)}^2 &\leq \bar{C}_U^2 \left( \eta_{\mathcal{T}}(u_{\mathcal{T}}, \mathcal{R})^2 + \text{osc}_{\mathcal{T}}(f, \mathcal{R})_{-1}^2 \right) \\ &\leq \bar{C}_U^2 \left( \eta_{\mathcal{T}}(u_{\mathcal{T}}, \bar{\mathcal{R}})^2 + \text{osc}_{\mathcal{T}}(f)_{-1}^2 \right) \\ &\leq \bar{C}_U^2 \left( (C_L^{\text{eq}})^{-2} \zeta_{\mathcal{T}}(u_{\mathcal{T}}, \bar{\mathcal{R}})^2 + \text{osc}_{\mathcal{T}}(f)_{-1}^2 \right). \end{aligned}$$

This is the desired estimate.  $\square$

**Lemma 6.28 (Lipschitz property of the estimator)** *Let the coefficients  $(\hat{A}, \hat{c})$  be discrete over  $\mathcal{T}$ . There exists  $C_{\text{Lip}}$  such that*

$$|\zeta_{\mathcal{T}}(v_1) - \zeta_{\mathcal{T}}(v_2)| \leq C_{\text{Lip}} |v_1 - v_2|_{H_0^1(\Omega)} \quad \forall v_1, v_2 \in \mathbb{V}_{\mathcal{T}}.$$

*Proof.* We resort to the star equivalence (5.86) between discrete residual and estimator. It thus suffices to derive the Lipschitz property for  $\|P_{\mathcal{T}} R_{\mathcal{T}}(v)\|_{H^{-1}(\omega_z)}$  with respect to  $v \in \mathbb{V}_{\mathcal{T}}$  for all  $z \in \mathcal{V}$ . Since  $P_{\mathcal{T}} R_{\mathcal{T}}(v) = P_{\mathcal{T}} f - \hat{\mathcal{B}}[v, \cdot]$ , we get

$$\langle P_{\mathcal{T}} R_{\mathcal{T}}(v_1) - P_{\mathcal{T}} R_{\mathcal{T}}(v_2), w \rangle = \int_{\omega_z} \nabla w \cdot \hat{A} \nabla(v_1 - v_2) + \hat{c}(v_1 - v_2)w$$

for all  $w \in H_0^1(\omega_z)$ . Therefore, Lemma 2.2 (first Poincaré inequality) yields

$$\|P_{\mathcal{T}} R_{\mathcal{T}}(v) - P_{\mathcal{T}} R_{\mathcal{T}}(v_2)\|_{H^{-1}(\omega_z)} \leq C(\hat{A}, \hat{c}) \|v_1 - v_2\|_{H^1(\omega_z)}$$

where  $C(\hat{A}, \hat{c})$  depends on the  $L^\infty$ -norms of  $(\hat{A}, \hat{c})$ . Finally, using the triangle inequality to accumulate over  $z \in \mathcal{V}$  together with (5.86) gives the assertion.  $\square$

Lemmas 6.27 and 6.28 lead to Lemma 6.16 (Dörfler marking) for  $\zeta_{\mathcal{T}}(u_{\mathcal{T}})$ . If we further choose a *minimal* set  $\mathcal{M}$  of vertices that satisfies Dörfler property (5.88), then the previous rates of convergence for the three algorithms GALERKIN, AFEM-SW, and AFEM-TS but now driven by  $\zeta_{\mathcal{T}}(u_{\mathcal{T}})$  are valid provided  $u \in \mathbb{A}_s$ , the approximation class in Definition 6.1. We do not restate these results.

## 6.8 Approximation vs regularity classes

The purpose of this section is to reconcile the notion of approximation classes, discussed above, with that of regularity classes. We recall the DeVore Diagram of Fig. 2.1 that depicts the Sobolev line for the energy space  $H_0^1(\Omega)$ , namely

$$\text{sob}(H_0^1) = \text{sob}(W_p^s) \quad \Rightarrow \quad s - \frac{d}{p} = 1 - \frac{d}{2}.$$

The differentiability  $s \geq 1$  is only limited by the polynomial degree  $n$ , so  $s \in [1, n+1]$ . On the other hand, the integrability  $p$  is not restricted to be  $p \geq 1$  as is customary with Sobolev spaces. For example, for  $d = 2$  and  $s = n+1$ , we get  $p = \frac{2}{n+1} < 1$  provided  $n \geq 2$ . Therefore, to take full advantage of nonlinear approximation theory, we need to abandon the framework of Sobolev spaces  $W_p^s(\Omega)$

and deal with *Besov spaces*  $B_{p,q}^s(\Omega)$  (frequently denoted  $B_q^s(L^p(\Omega))$  or  $B_q^s(L_p(\Omega))$  in the literature) with integrability index  $p \in (0, \infty]$ . The second index  $q \in (0, \infty]$  is useful to characterize special limiting cases; we will provide below a few interesting examples but take  $p = q$  most of the time. At this point, we only mention that when  $s$  is non-integer and  $1 \leq p \leq \infty$ ,  $B_{p,p}^s(\Omega) = W_p^s(\Omega)$  while when  $r$  is integer  $W_p^r(\Omega)$  for  $p \neq 2$  is not a Besov space but it is slightly smaller than  $B_{p,\infty}^r(\Omega)$ . The case  $p = 2$  is special since  $B_{2,2}^s(\Omega) = H^s(\Omega)$  even when  $s$  is an integer.

This section is devoted to the definition and properties of Besov and Lipschitz spaces, including their close relation to approximation classes. Our presentation follows closely [Binev et al. 2002] for  $n = 1$  and [Gaspaz and Morin 2014, 2017] for  $n \geq 1$ , but it adds a few new ingredients. Since our results involve three different type of spaces to account for the particular cases when the differentiability is integer, it is pertinent to introduce the following abstract space  $X_p^s(\Omega)$  with differentiability index  $s \in (0, n + 1]$  and integrability index  $p \in (0, \infty]$

$$X_p^s(\Omega) := \begin{cases} B_{p,p}^s(\Omega) & s \in (0, n + 1), p \in (0, \infty], \\ W_p^{n+1}(\Omega) & s = n + 1, p \in [1, \infty], \\ \text{Lip}_p^{n+1}(\Omega) & s = n + 1, p \in (0, 1). \end{cases} \quad (6.35)$$

Here  $\text{Lip}_p^s(\Omega) = \text{Lip}(s, L^p(\Omega))$ ,  $s \in \mathbb{N}$ , are the Lipschitz spaces; see (6.58) below. For  $s \in \mathbb{N}$  and  $1 < p < \infty$  the Sobolev spaces coincide with the Lipschitz spaces [Leoni 2009, Theorem 10.55], i.e.,

$$\text{Lip}_p^s(\Omega) = W_p^s(\Omega), \quad s \in \mathbb{N}, \quad 1 < p < \infty, \quad (6.36)$$

while for  $p = 1$  we only have

$$W_1^s(\Omega) \hookrightarrow \text{Lip}_1^s(\Omega), \quad s \in \mathbb{N}. \quad (6.37)$$

We use the following conventions:  $X_p^s(\Omega) := L^p(\Omega)$  for  $s = 0$ ;  $X_p^s(\Omega; \mathcal{T})$  is the space of functions with piecewise regularity  $X_p^s$  over  $\mathcal{T} \in \mathbb{T}$ ;  $X_p^s(\Omega; \mathbb{R}^m)$  is the space  $X_p^s(\Omega)$  of vector or matrix-valued functions.

We will prove in Section 6.8.4 the following crucial approximation results for functions in  $L^q(\Omega)$  by discontinuous piecewise polynomials  $\mathbb{S}_{\mathcal{T}}^{n,-1}$  of degree  $n \geq 1$  over conforming refinements  $\mathcal{T}$  of  $\mathcal{T}_0$ . It turns out that this will also allow us to deal with approximations in  $H_0^1(\Omega)$  by continuous piecewise polynomials  $\mathbb{V}_{\mathcal{T}}$  of degree  $n \geq 1$ .

**Theorem 6.29 (regularity yields approximation)** *Let  $q \in [1, \infty]$ ,  $p \in (0, \infty]$ ,  $s \in (0, n + 1]$  and a function  $g \in L^q(\Omega)$  satisfy  $g \in X_p^s(\Omega)$  with  $s - \frac{d}{p} + \frac{d}{q} > 0$ . Then, there exists a constant  $C = C(p, q, s, t, d, \Omega, \mathcal{T}_0)$  such that*

$$E_n(g, \Omega)_q := \inf_{\mathcal{T} \in \mathbb{T}_N} \inf_{v \in \mathbb{S}_{\mathcal{T}}^{n,-1}} |g - v|_{L^q(\Omega)} \leq C |g|_{X_p^s(\Omega)} N^{-\frac{s}{d}}. \quad (6.38)$$

Therefore,  $g \in \mathbb{A}_{\frac{s}{d}} = \mathbb{A}_{\frac{s}{d}}(L^q(\Omega); \mathcal{T}_0)$  and

$$|g|_{\mathbb{A}_{\frac{s}{d}}} \leq C |g|_{X_p^s(\Omega)}. \quad (6.39)$$

We see that the decay rate  $\frac{s}{d}$  in (6.38) is proportional to the difference of the differentiability indices between the space  $X_p^s(\Omega)$  and  $L^q(\Omega)$  provided the Sobolev numbers satisfy the relation

$$\text{sob}(X_p^s(\Omega)) > \text{sob}(L^q(\Omega)),$$

which implies that the embedding of  $X_p^s(\Omega)$  into  $L^q(\Omega)$  is compact. The factor  $d$  in the denominator is a manifestation of the so-called *curse of dimensionality*. The limiting case  $s = n + 1$  entails dealing with Sobolev spaces  $W_p^s(\Omega)$  and Lipschitz spaces  $\text{Lip}_p^s(\Omega)$  depending on whether  $p \geq 1$  or  $p < 1$ .

### 6.8.1 Modulus of smoothness

*Difference Operators.* Since we intend to allow  $p \in (0, 1)$ , the underlying functions in  $B_{p,p}^s(\Omega)$  might not be locally integrable, whence they might not be distributions in  $\Omega$ . Therefore, the notion of weak derivative does not apply, which in turn has the drawback of being defined for integers and not for fractional numbers. This leads to the most standard definition of Besov spaces  $B_{p,q}^s(\Omega)$  using difference operators, which only requires integrability in  $L^p(\Omega)$  and is valid for any  $s > 0$ ,  $p, q \in (0, \infty]$ . Other definitions, which provides equivalent results in the range  $1 \leq p, q \leq \infty$  can be found in [Adams and Fournier 2003, Bergh and Löfström 1976].

Given a bounded Lipschitz domain  $\Omega \subset \mathbb{R}^d$ , and a vector  $h \in \mathbb{R}^d$ , we set

$$\Omega_h := \{x \in \Omega : [x, x+h] \subset \Omega\}$$

where  $[x, x+h]$  denotes the closed segment connecting  $x$  and  $x+h$ , and define the first-order difference operator to be

$$\Delta_h^1 g(x) = \Delta_h^1(g, x, \Omega) := \begin{cases} g(x+h) - g(x) & x \in \Omega_h, \\ 0 & \text{otherwise.} \end{cases} \quad (6.40)$$

For  $k \in \mathbb{N}$ ,  $k \geq 1$ , we define the  $k$ -th difference operator by iteration

$$\Delta_h^k g(x) := \Delta_h^1(\Delta_h^{k-1})g(x) \quad x \in \Omega_{kh} \quad (6.41)$$

and observe that it has the explicit form

$$\Delta_h^k g(x) = \begin{cases} \sum_{j=0}^k (-1)^{k+j} \binom{k}{j} g(x+jh) & [x, x+kh] \subset \Omega, \\ 0 & \text{otherwise.} \end{cases}$$

Note the property

$$p \in \mathbb{P}_k \quad \Rightarrow \quad \Delta_h^{k+1} p = 0 \quad \forall h. \quad (6.42)$$

*Smoothness.* Given  $p \in (0, \infty]$  and  $t > 0$ , we define the *modulus of smoothness* of order  $k$  in  $L^p(\Omega)$  to be

$$\omega_k(g, t)_p = \omega_k(g, t, \Omega)_p := \sup_{|h| \leq t} \|\Delta_h^k g\|_{L^p(\Omega)}. \quad (6.43)$$

We note that if  $\omega_k(g, t)_p = o(t^{n+1})$  as  $t \rightarrow 0$ , then  $g$  is a.e a polynomial in  $\mathbb{P}_n$  and

$$g \notin \mathbb{P}_n \Rightarrow \omega_k(g, t)_p \geq Ct^{n+1} \quad 0 < t \leq 1 \quad (6.44)$$

for some  $C > 0$  [DeVore and Lorentz 1993, Proposition 7.4]. We also observe that the definition (6.43) only requires  $L^p$ -integrability of  $g$  and leads to the following celebrated Whitney estimate of the *best approximation error*

$$E_n(g, G)_p := \inf_{v \in \mathbb{P}_n} \|g - v\|_{L^p(G)}$$

of  $g$  by polynomials of degree  $\leq n$  in  $G \subset \Omega$  [Binev et al. 2002], [Dekel and Leviatan 2004, Theorem 1.4] [Gaspoz and Morin 2014, 2017, Lemma 4.4].

**Lemma 6.30 (Whitney's lemma)** *Let  $\mathcal{T} \in \mathbb{T}$  be an admissible grid, and let  $T \in \mathcal{T}$  be a generic element. If  $0 < p \leq \infty$  and  $n \geq 0$ , then*

$$E_n(g, T)_p \leq C\omega_{n+1}(g, h_T, T)_p \quad \forall g \in L^p(T),$$

where  $C = C(p, n, d, \mathcal{T}_0)$  but is independent of  $g$  and the size of  $T$ .

### 6.8.2 Besov spaces

Given  $s > 0$  and  $0 < p, q \leq \infty$ , the Besov space  $B_{p,q}^s(\Omega)$  is the set of all functions  $v \in L^p(\Omega)$  such that the following quantity is finite

$$|v|_{B_{p,q}^s(\Omega)} := \begin{cases} \left( \int_0^\infty [t^{-s}\omega_k(v, t)_p]^q \frac{dt}{t} \right)^{\frac{1}{q}} & 0 < q < \infty, \\ \sup_{t>0} [t^{-s}\omega_k(v, t)_p] & q = \infty, \end{cases} \quad (6.45)$$

with  $k = [s] + 1 \in \mathbb{N}$  and  $[s]$  stands for the integer part of  $s$ . If we split the integral in (6.45) for  $0 < q < \infty$  in dyadic intervals, we obtain the following equivalent expression for  $|v|_{B_{p,q}^s(\Omega)}$

$$|v|_{B_{p,q}^s(\Omega)}^q = \sum_{m \in \mathbb{Z}} \int_{2^{-m-1}}^{2^{-m}} t^{-sq} \omega_k(v, t)_p^q \frac{dt}{t} \approx \sum_{m \in \mathbb{Z}} 2^{msq} \omega_k(v, 2^{-m})_p^q; \quad (6.46)$$

here we have used that both  $\omega_k(v, t)_p$  and  $t^{-s}$  are monotone functions of  $t$ . The hidden constants depend on  $s$  and  $q$  but are otherwise independent of  $v, k$  and  $p$ . Note that, with obvious changes, (6.46) is also valid for  $q = \infty$

$$|v|_{B_{p,\infty}^s(\Omega)} \approx \sup_{m \in \mathbb{Z}} 2^{ms} \omega_k(v, 2^{-m})_p. \quad (6.47)$$

We point out that  $|v|_{B_{p,q}^s(\Omega)}$  is a semi-norm for  $p, q \geq 1$  and is otherwise a semi-(quasi)norm in that the triangle inequality is valid up to a constant larger than 1; note that  $|1|_{B_{p,q}^s(\Omega)} = 0$ . The quasi-norm of  $B_{p,q}^s(\Omega)$  is defined to be

$$\|v\|_{B_{p,q}^s(\Omega)} := \|v\|_{L^p(\Omega)} + |v|_{B_{p,q}^s(\Omega)}.$$

If an integer  $k' > k$  is chosen in (6.45), then the ensuing quasi-norms  $\|v\|_{B_{p,q}^s(\Omega)}$

are equivalent. This hinges on the *Marchaud inequality* [DeVore and Popov 1988, eq (2.6)], [Ditzian 1988, Theorems 1 and 3]

$$\omega_k(v, t)_p \leq Ct^k \left( \|v\|_{L^p(\Omega)} + \left( \int_t^\infty (z^{-k} \omega_{k'}(v, z)_p)^p \frac{dz}{z} \right)^{\frac{1}{p}} \right). \quad (6.48)$$

The following lemma characterizes the precise blow-up of  $|v|_{B_{p,p}^s(\Omega)}$  as  $s \rightarrow n+1$ .

**Lemma 6.31 (blow-up of  $|v|_{B_{p,p}^s(\Omega)}$ )** *Let  $s \in (0, n+1)$ ,  $p \in (0, \infty]$ . Then*

$$|v|_{B_{p,p}^s(\Omega)} \leq (p(n+1-s))^{-\frac{1}{p}} \|v\|_{B_{p,p}^{n+1}(\Omega)} \quad \forall v \in B_{p,p}^{n+1}(\Omega).$$

*Proof.* We take  $p < \infty$  and combine the definition (6.45) with (6.48), after replacing the upper limit of integration by  $\text{diam}(\Omega) \approx 1$ , to write

$$|v|_{B_{p,p}^s(\Omega)}^p \approx \int_0^1 \left( t^{-s} \omega_{n+1}(v, t)_p \right)^p \frac{dt}{t} \lesssim I + II,$$

with

$$I = \int_0^1 t^{(n+1-s)p} \int_t^1 \left( z^{-(n+1)} \omega_{n+2}(v, z)_p \right)^p \frac{dz}{z} \frac{dt}{t}, \quad II = \int_0^1 t^{(n+1-s)p} \|v\|_{L^p(\Omega)}^p \frac{dt}{t}.$$

Exchanging the order of integration yields

$$\begin{aligned} I &= \int_0^1 \left( z^{-(n+1)} \omega_{n+2}(v, z)_p \right)^p \left( \int_0^z t^{(n+1-s)p-1} dt \right) \frac{dz}{z} \\ &= \frac{1}{p(n+1-s)} \int_0^1 \left( z^{-s} \omega_{n+2}(v, z)_p \right)^p \frac{dz}{z} \leq \frac{1}{p(n+1-s)} |v|_{B_{p,p}^{n+1}(\Omega)}^p. \end{aligned}$$

Since  $II = (p(n+1-s))^{-1} \|v\|_{L^p(\Omega)}^p$ , the proof is thus complete.  $\square$

The following equivalence between Sobolev and Besov spaces is valid for fractional differentiability  $s$  [Leoni 2009, Proposition 14.40] (see also [Bergh and Löfström 1976, §6.4.4], [Adams and Fournier 2003, §7.33, §7.67]): for all  $s \geq 0$ ,  $s \notin \mathbb{N}$  and  $p \in [1, \infty]$

$$B_{p,p}^s(\Omega) = W_p^s(\Omega). \quad (6.49)$$

However, if  $s \in \mathbb{N}$  is an integer, then  $B_{p,q}^s(\Omega)$  is defined using  $k = s+1$  differences whereas  $W_p^s(\Omega)$  involves  $s$  weak derivatives in  $L^p(\Omega)$  provided  $p \in [1, \infty]$ . It turns out that for integer values  $s \in \mathbb{N}$  the spaces differ

$$B_{p,p}^s(\Omega) \neq W_p^s(\Omega), \quad p \neq 2, \quad (6.50)$$

except for the exceptional case  $p = 2$  for which  $B_{2,2}^s(\Omega) = H^s(\Omega)$  [DeVore 1998].

The Besov semi-norm is *sub-additive* in the following sense: if  $\{T_i\}_{i=1}^N$  is a disjoint collection of elements  $T_i \in \mathcal{T}$  and  $\mathcal{T} \in \mathbb{T}$ ,  $p \in (0, \infty]$  and  $s > 0$ , then there

exists a constant  $C$  depending on  $p, s, d$  and  $\mathcal{T}_0$  but independent of  $N$  such that

$$\sum_{i=1}^N |v|_{B_{p,p}^s(T_i)}^p \leq C |v|_{B_{p,p}^s(\Omega)}^p \quad \forall v \in B_{p,p}^s(\Omega). \quad (6.51)$$

The localization of Besov norms is more general than (6.51). In fact, if  $\omega_{\mathcal{T}}(T)$  denotes the patch of elements in  $\mathcal{T}$  around  $T \in \mathcal{T}$  (first ring), then the following is valid with equivalence constants depending  $p, s, d$  and  $\mathcal{T}_0$  but independent of  $N$  [Binev *et al.* 2002, Lemmas 4.3 and 4.4]

$$\sum_{T \in \mathcal{T}} |v|_{B_{p,p}^s(\omega_{\mathcal{T}}(T))}^p \approx C |v|_{B_{p,p}^s(\Omega)}^p \quad \forall v \in B_{p,p}^s(\Omega). \quad (6.52)$$

The following statements about embeddings between Besov spaces on bounded Lipschitz domains  $\Omega$  will turn to be useful in the sequel [Triebel 2010, §3.2.4, §3.3.1]: if  $0 < p \leq \infty$ ,  $0 < q_1, q_2 \leq \infty$ , and  $s_1, s_2, s > 0$ , then

$$\begin{aligned} s_1 > s_2 &\Rightarrow B_{p,q_1}^{s_1}(\Omega) \hookrightarrow B_{p,q_1}^{s_2}(\Omega), \\ q_1 < q_2 &\Rightarrow B_{p,q_1}^s(\Omega) \hookrightarrow B_{p,q_2}^s(\Omega). \end{aligned} \quad (6.53)$$

Because of the second relation in (6.53), statements valid for all second index  $q$  are written for the largest space corresponding to  $q = \infty$ . In addition, for all  $0 < p, q, r \leq \infty$  and  $s > 0$ , the *discrepancy* between the spaces  $B_{p,r}^s(\Omega)$  and  $L^q(\Omega)$  is the quantity

$$\delta := s - \frac{d}{p} + \frac{d}{q}. \quad (6.54)$$

The discrepancy  $\delta$  governs the embedding between these two spaces [Leoni 2009, Theorems 14.29 and 14.32], [ DeVore 1998], namely

$$\delta > 0 \Rightarrow B_{p,\infty}^s(\Omega) \hookrightarrow L^q(\Omega); \quad \delta = 0 \Rightarrow B_{p,p}^s(\Omega) \hookrightarrow L^q(\Omega), \quad q \neq \infty, \quad (6.55)$$

and the embedding is compact when  $\delta > 0$ . Notice that  $\delta = 0$  determines the Sobolev embedding line of the DeVore diagram in Fig. 2.1.

We stress that when  $\delta > 0$  the third parameter  $r$  in  $B_{p,r}^s(\Omega)$  plays no role in (6.55) involving the largest space  $B_{p,\infty}^s(\Omega)$ , see (6.53). However, it turns out to be useful to quantify regularity in extreme cases. For instance, the characteristic function  $\chi_G$  of a smooth set  $G \subsetneq \Omega$  satisfies

$$\chi_G \in B_{p,\infty}^{\frac{1}{p}}(\Omega) \setminus B_{p,r}^{\frac{1}{p}}(\Omega) \quad 0 < p, r < \infty.$$

Moreover, the Lagrange basis functions  $\{\phi_z\}_{z \in \mathcal{N}}$  of  $\mathbb{V}_{\mathcal{T}}$  satisfy for any  $0 < p \leq \infty$  [Gaspoz and Morin 2014, Proposition 4.7]

$$\omega_{n+1}(\phi_z, t)_p = \begin{cases} |\text{supp } \phi_z|^{\frac{d-1-p}{dp}} t^{1+\frac{1}{p}} & 0 < t \leq |\text{supp } \phi_z|^{\frac{1}{d}}, \\ |\text{supp } \phi_z| & t > |\text{supp } \phi_z|^{\frac{1}{d}}. \end{cases}$$



This readily implies that for all  $0 < s < 1 + \frac{1}{p}$  and  $0 < q < \infty$

$$\mathbb{V}_{\mathcal{T}} \subset B_{p,q}^s(\Omega), \quad \mathbb{V}_{\mathcal{T}} \subset B_{p,\infty}^{1+\frac{1}{p}}(\Omega). \quad (6.56)$$

### 6.8.3 Local approximation

We are now in the position to prove a key approximation estimate. In finite element theory it comes by the name of Bramble-Hilbert Lemma whereas in nonlinear approximation theory it is called Jackson Theorem. We distinguish between the case  $0 < s < n + 1$  and the limit integral case  $s = n + 1$ .

**Proposition 6.32 (Bramble-Hilbert for Besov spaces)** *Let  $\mathcal{T} \in \mathbb{T}$  and  $T \in \mathcal{T}$ . Let  $0 < p, q \leq \infty$ ,  $0 < s < n + 1$ , and either  $s - \frac{d}{p} + \frac{d}{q} \geq 0$ ,  $q < \infty$  or  $s > \frac{d}{p}$ ,  $q = \infty$ . Set  $r = \infty$  when  $s - \frac{d}{p} + \frac{d}{q} > 0$  and  $r = p$  otherwise. Then we have*

$$\inf_{P \in \mathbb{P}_n} \|v - P\|_{L^q(T)} \leq Ch_T^{s - \frac{d}{p} + \frac{d}{q}} |v|_{B_{p,r}^s(T)} \quad \forall v \in B_{p,r}^s(T), \quad (6.57)$$

where the constant  $C = C(p, q, s, d, n, \mathcal{T}_0)$  is independent of  $v$  and  $T$ .

*Proof.* We first point out that we could use  $k = n + 1 \geq [s] + 1$  in the definition of  $|v|_{B_{p,r}^s(T)}$  according to (6.48). We next proceed in three steps.

[1] Suppose first that  $T$  is the master element, namely  $|T| \approx 1$ . If  $P \in \mathbb{P}_n$  is an arbitrary polynomial, using that the discrepancy  $\delta = s - \frac{d}{p} + \frac{d}{q} \geq 0$  yields

$$E_n(v, T)_q \leq \|v - P\|_{L^q(T)} \lesssim \|v - P\|_{B_{p,r}^s(T)} = \|v - P\|_{L^p(T)} + |v - P|_{B_{p,r}^s(T)}$$

due to the embedding (6.55). Since the definition of  $\omega_{n+1}(v, t)_p$  involves  $n + 1$  differences, we deduce  $\Delta_h^{n+1} P = 0$  in view of (6.42), whence  $|v - P|_{B_{p,r}^s(T)} = |v|_{B_{p,r}^s(T)}$ . We now take  $P$  to be the best approximation of  $v$  in  $L^p(T)$  to derive

$$E_n(v, T)_q \lesssim E_n(v, T)_p + |v|_{B_{p,r}^s(T)}.$$

[2] We perform a scaling argument from the element  $T \in \mathcal{T}$  to the master element  $\widehat{T}$ . Let  $\widehat{x} = |T|^{-1/d} x$  be the change of variables and note that

$$\begin{aligned} \omega_{n+1}(v, t, T)_p &= \sup_{|h| \leq t} \|\Delta_h^{n+1} v\|_{L^p(T)} \\ &= |T|^{\frac{1}{p}} \sup_{|h| \leq t} \|\Delta_{h|T|^{-1/d}}^{n+1} \widehat{v}\|_{L^p(\widehat{T})} = |T|^{\frac{1}{p}} \omega_{n+1}(\widehat{v}, \widehat{t}, \widehat{T})_p \end{aligned}$$

with  $\widehat{t} = t|T|^{-1/d}$ , whence

$$\begin{aligned} |v|_{B_{p,r}^s(T)}^r &= \int_0^\infty \left( t^{-s} \omega_{n+1}(v, t, T)_p \right)^r \frac{dt}{t} \\ &= |T|^{\frac{r}{p} - \frac{sr}{d}} \int_0^\infty \left( \widehat{t}^{-s} \omega_{n+1}(\widehat{v}, \widehat{t}, \widehat{T})_p \right)^r \frac{d\widehat{t}}{\widehat{t}} = |T|^{\frac{r}{p} - \frac{sr}{d}} |\widehat{v}|_{B_{p,r}^s(\widehat{T})}^r. \end{aligned}$$

Therefore, since  $E_n(v, T)_q = |T|^{\frac{1}{q}} E_n(\widehat{v}, \widehat{T})_q$ , we obtain

$$E_n(v, T)_q \lesssim |T|^{\frac{1}{q} - \frac{1}{p}} E_n(v, T)_p + |T|^{\frac{s}{d} - \frac{1}{p} + \frac{1}{q}} |v|_{B_{p,r}^s(T)}.$$

3 It remains to estimate  $E_n(v, T)_p$  which, in view of Lemma 6.30 (Whitney's lemma), satisfies  $E_n(v, T)_p \leq C \omega_{n+1}(v, h_T, T)_p$  with  $h_T \approx |T|^{1/d} \approx 2^{-m}$  for some  $m \in \mathbb{Z}$ . Since  $k = n + 1 \geq [s] + 1$ , invoking the equivalent definition (6.48) of  $|v|_{B_{p,r}^s(T)}$  yields

$$E_n(v, T)_p^r \lesssim \omega_{n+1}(v, 2^{-m}, T)_p^r \lesssim h_T^{sr} \sum_{m \in \mathbb{Z}} 2^{msr} \omega_{n+1}(v, 2^{-m}, T)_p^r = h_T^{sr} |v|_{B_{p,r}^s(T)}^r.$$

Inserting this estimate into that of Step 2 gives (6.57), as asserted.  $\square$

We now consider the integer case  $s = n + 1$ . The first thing to notice is that (6.57) cannot possibly be valid: the definition (6.45) requires  $k = [s] + 1 = n + 2$ , whence any polynomial  $g \in \mathbb{P}_{n+1} \setminus \mathbb{P}_n$  satisfies  $E_n(g, T)_p > 0$  as well as  $\omega_{n+2}(g, T)_p = 0$  according to (6.42). Lemma 6.31 (blow-up of  $|v|_{B_{p,p}^s(\Omega)}$ ) reveals that replacing the semi-norm  $|v|_{B_{p,p}^s(\Omega)}$  by the full norm  $\|v\|_{B_{p,p}^{n+1}(\Omega)}$  is not a good idea either. To overcome this problem, we now introduce the space  $\text{Lip}_p^s(\Omega) := \text{Lip}(s, L^p(\Omega))$  of  $s$ -Lipschitz functions with values in  $L^p(\Omega)$ ,  $0 < p < \infty$  [DeVore 1998, p.92]

$$|g|_{\text{Lip}_p^s(\Omega)} := \sup_{t>0} \left( t^{-s} \omega_{n+1}(g, t, \Omega)_p \right). \quad (6.58)$$

Comparing with (6.45) we realize that  $\text{Lip}_p^s(\Omega) = B_{p,\infty}^s(\Omega)$  provided  $s \notin \mathbb{N}$  but  $\text{Lip}_p^s(\Omega) \neq B_{p,\infty}^s(\Omega)$  when  $s \in \mathbb{N}$ . Moreover,

$$\delta = s - \frac{d}{p} + \frac{d}{q} > 0, \quad s \in \mathbb{N} \quad \Rightarrow \quad \text{Lip}_p^s(\Omega) \hookrightarrow L^q(\Omega), \quad (6.59)$$

with compact embedding. If  $\delta = 0$  and  $p \geq 1$ ,  $q \neq \infty$ , the above embedding is continuous in view of (6.36) and (6.55).

**Proposition 6.33 (Bramble-Hilbert for Lipschitz spaces)** *Let  $\mathcal{T} \in \mathbb{T}$  and  $T \in \mathcal{T}$ . If  $p \in (0, \infty)$ ,  $q \in (0, \infty]$ ,  $k \geq 0$  integer, and  $k + 1 - \frac{d}{p} + \frac{d}{q} \geq 0$ , with strict inequality when  $p < 1$  or  $q = \infty$ , then we have*

$$\inf_{P \in \mathbb{P}_k} \|v - P\|_{L^q(T)} \leq C h_T^{k+1 - \frac{d}{p} + \frac{d}{q}} |v|_{\text{Lip}_p^{k+1}(T)} \quad \forall v \in \text{Lip}_p^{k+1}(T), \quad (6.60)$$

where the constant  $C = C(p, q, d, k, \mathcal{T}_0)$  is independent of  $v$  and  $T$ .

*Proof.* In view of (6.59), we proceed as in the proof of Proposition 6.32, except for the following change in Step 3. For  $h_T \approx 2^{-m}$ , we have instead

$$\begin{aligned} E_k(v, T)_p &\lesssim \omega_{k+1}(v, 2^{-m}, T)_p \\ &\lesssim h_T^{k+1} \sup_{m \in \mathbb{Z}} \left( 2^{m(k+1)} \omega_{k+1}(v, 2^{-m}, T)_p \right) = h_T^{k+1} |v|_{\text{Lip}_p^{k+1}(T)}. \end{aligned}$$

This concludes the proof.  $\square$

It is instructive to realize that Propositions 6.32 and 6.33 extend to Besov and Lipschitz spaces the usual Bramble-Hilbert lemma for Sobolev spaces [Brenner and Scott 2008, Lemma 4.3.8].

**Proposition 6.34 (Bramble-Hilbert for Sobolev spaces)** *Let  $\mathcal{T} \in \mathbb{T}$  and  $T \in \mathcal{T}$ . For all  $1 \leq p, q, \leq \infty$  and  $0 < s \leq n+1$  such that  $s - \frac{d}{p} + \frac{d}{q} \geq 0$  with strict inequality when  $q = \infty$ , then*

$$\inf_{P \in \mathbb{P}_n} \|v - P\|_{L^q(T)} \leq Ch_T^{s - \frac{d}{p} + \frac{d}{q}} |v|_{W_p^s(T)} \quad \forall v \in W_p^s(T), \quad (6.61)$$

where the constant  $C = C(p, q, s, d, n, \mathcal{T}_0)$  but is independent of  $v$  and  $T$ .

*Proof.* When  $s$  is fractional,  $W_p^s(T) = B_{p,p}^s(T)$  in view of (6.49) and the result follows from Proposition 6.32. Instead, when  $s$  is integral, we invoke (6.36) and (6.37) to deduce the result from Proposition 6.33.  $\square$

#### 6.8.4 Global approximation: direct estimates

We now collect local contributions from Propositions 6.32, 6.33 and 6.34, depending on the range of parameters  $p, q, s$ , to find global error estimates for the solution  $u$  as well as the coefficients  $(A, c)$  of (2.7). They are trivial consequences of Theorem 6.29, which we prove first. The analysis of the forcing function  $f$  is somewhat different, due to the non-locality of the corresponding norm  $H^{-1}(\Omega)$ , and is postponed to Section 7.3.

**Proof of Theorem 6.29.** Since the discrepancy  $\delta = s - \frac{d}{p} + \frac{d}{q} > 0$ , the embedding  $X_p^s(\Omega) \hookrightarrow L^q(\Omega)$  is compact according to (6.55) and (6.59). Given  $g \in X_p^s(\Omega)$ , we consider the surrogate quantity  $e_{\mathcal{T}}(g, T) := Ch_T^\delta |g|_{X_p^s(T)}$ , which satisfies

$$E_n(g, T)_q = \inf_{v \in \mathbb{P}_n} \|g - v\|_{L^q(T)} \leq e_{\mathcal{T}}(g, T) \quad \forall T \in \mathcal{T}$$

by virtue of Bramble-Hilbert Propositions 6.32, 6.33 and 6.34. We finally combine Proposition 3.19 (abstract greedy) with the subadditivity property (6.51) to deduce the desired estimate (6.3). The remaining estimate (6.39) follows from the definition of  $|g|_{\mathbb{A}_{\frac{s}{d}}}$ . This concludes the proof.

Inspection of this proof reveals that our estimate is stronger than (6.38). In fact, we need the weaker regularity

$$|g|_{X_p^s(\Omega; \mathcal{T})}^p = \sum_{T \in \mathcal{T}} |g|_{X_p^s(T)}^p < \infty,$$

which allows for piecewise Besov smoothness of  $g$  very much in the spirit of (3.20). This may accommodate singular behavior of  $g$  aligned with the initial mesh  $\mathcal{T}_0$ .

**Corollary 6.35 (approximation class of  $u$ )** *Let the solution  $u \in H_0^1(\Omega)$  of (2.7) satisfy  $u \in X_p^s(\Omega)$  with  $s \in (0, n+1]$ ,  $p \in (0, \infty]$  and  $s - 1 - \frac{d}{p} + \frac{d}{2} > 0$ , where*

$X_p^s(\Omega)$  is defined in (6.35). Then  $u \in \mathbb{A}_{\frac{s-1}{d}}(H_0^1(\Omega); \mathcal{T}_0)$  and

$$|u|_{\mathbb{A}_{\frac{s-1}{d}}} \lesssim |u|_{X_p^s(\Omega)}. \quad (6.62)$$

Equivalently,  $\sigma_N(u)$  defined in (6.3) satisfies

$$\sigma_N(u) \lesssim |u|_{X_p^s(\Omega)} N^{-\frac{s-1}{d}} \quad N \geq \#\mathcal{T}_0. \quad (6.63)$$

*Proof.* In view of (3.19) of Proposition 3.9 (approximation of gradients), namely

$$\inf_{v \in \mathbb{S}_{\mathcal{T}}^{n,0}} \|\nabla(u - v)\|_{L^2(\Omega)} \lesssim \inf_{v \in \mathbb{S}_{\mathcal{T}}^{n,-1}} \|\nabla(u - v)\|_{L^2(\Omega)},$$

we realize that it suffices to bound the element errors for  $g = \nabla u \in L^2(\Omega; \mathbb{R}^d)$  by vector-valued discontinuous piecewise polynomials of degree  $\leq n - 1$ . Therefore, applying Theorem 6.29 (regularity yields approximation) with  $n$  replaced by  $n - 1$  gives the desired estimates (6.62) and (6.63).  $\square$

We now turn our attention to the coefficients  $(A, c)$ . Regarding  $A$ , Lemma 5.20 (continuous dependence on data) shows that the natural function space for  $A$  is  $L^\infty(\Omega, \mathbb{R}^{d \times d})$  provided  $u \in H_0^1(\Omega)$ . However, Lemma 5.20 also allows for  $A \in L^r(\Omega, \mathbb{R}^{d \times d})$ ,  $2 \leq r < \infty$ , provided  $u \in W_p^1(\Omega)$  with  $2 \leq p = \frac{2r}{r-2} < p_1$  which in turn is guaranteed by Lemma 2.13 ( $W_p^1$ -regularity). The latter permits discontinuities of  $A$  within elements, which is of practical importance. Therefore, we consider the most general situation  $2 \leq r \leq \infty$  in the sequel.

**Corollary 6.36 (approximation class of  $A$ )** For  $0 < \alpha_1 \leq \alpha_2$  and  $2 \leq r \leq \infty$  let the diffusion coefficient  $A \in M(\alpha_1, \alpha_2)$  of (2.7) satisfy  $A \in X_p^s(\Omega; \mathbb{R}^{d \times d})$  with  $s \in (0, n]$ ,  $p \in (0, \infty]$  and  $s - \frac{d}{p} + \frac{d}{r} > 0$ . Then  $A \in \mathbb{M}_{\frac{s}{d}} = \mathbb{M}_{\frac{s}{d}}((L^r(\Omega))^{d \times d}; \mathcal{T}_0)$  and

$$|A|_{\mathbb{M}_{\frac{s}{d}}} \lesssim |A|_{X_p^s(\Omega)}. \quad (6.64)$$

Equivalently,  $\tilde{\delta}_{\mathcal{T}}(A)_r$  defined in Section 6.1.2 satisfies

$$\inf_{\mathcal{T} \in \mathbb{T}_N} \tilde{\delta}_{\mathcal{T}}(A)_r \lesssim |A|_{X_p^s(\Omega)} N^{-\frac{s}{d}} \quad \forall \mathcal{T} \in \mathbb{T}_N, N \geq \#\mathcal{T}_0, \quad (6.65)$$

and this error decay is achieved by Algorithm 3.18 (greedy algorithm).

*Proof.* Simply recall the relation (6.12) between the best constrained and unconstrained approximation errors and apply Theorem 6.29 (regularity yields approximation).  $\square$

Consider the special case  $s = n$  and  $r = \infty$  in Corollary 6.36. We readily see that  $p > \frac{d}{n}$  which might be less than 1 for  $n > d$ , hence the need for Besov spaces.

We finally deal with the reaction coefficient  $c \in L^\infty(\Omega)$ . According to Lemma 5.20 (continuous dependence on data), and the discussion in Section 5.4.2, a natural space for  $c$  is  $L^q(\Omega)$  with  $\frac{d}{2} < q \leq \infty$ ,  $s = 0$ ; we could take  $q = 2$  for  $d < 4$ . Section 5.4.2 also reveals that the case  $n = 1$  is somewhat special in that we can

exploit superconvergence in  $W_q^{-1}(\Omega)$  with  $q > d$ . In fact, combining the argument following (5.75) with (5.68) yields

$$\inf_{\widehat{c} \in \mathbb{S}_{\mathcal{T}}^{n-1, -1}} \|c - \widehat{c}\|_{W_q^{-1}(\Omega)}^2 \lesssim \sum_{T \in \mathcal{T}} h_T^{2t} \|c - \Pi_T c\|_{L^2(T)}^2 \lesssim \sum_{T \in \mathcal{T}} h_T^{2t} \delta_{\mathcal{T}}(c, T)_2^2 = \text{osc}_{\mathcal{T}}(c)_2^2$$

with  $0 < t = 1 - \frac{d}{2} + \frac{d}{q} < 2 - \frac{d}{2}$  provided  $d < 4$ . This gives the following statement. We note that (5.76) could also be combined with (5.68) for  $n = 1$  to obtain a similar result for  $\text{osc}_{\mathcal{T}}(c)_\infty$  with  $t = 1$  and any  $d \geq 2$ ; however, we do not elaborate further.

**Corollary 6.37 (approximation of  $c$ )** *Let  $0 \leq c_1 \leq c_2$  and the reaction coefficient  $c \in R(c_1, c_2)$  satisfy  $c \in X_p^s(\Omega)$  with  $s \in (0, n]$ ,  $p \in (0, \infty]$ . If  $n \geq 1$ ,  $q > \frac{d}{2}$ , and  $s - \frac{d}{p} + \frac{d}{q} > 0$ , then  $c \in \mathbb{C}_{\frac{s}{d}} = \mathbb{C}_{\frac{s}{d}}(L^q(\Omega); \mathcal{T}_0)$  and*

$$|c|_{\mathbb{C}_{\frac{s}{d}}} \lesssim |c|_{X_p^s(\Omega)}. \quad (6.66)$$

*If instead  $n = 1$ ,  $q > d$ ,  $s - \frac{d}{p} + \frac{d}{2} \geq 0$ ,  $0 < t = 1 - \frac{d}{2} + \frac{d}{q} < 2 - \frac{d}{2}$  and  $d < 4$ , then  $c \in \mathbb{C}_{\frac{s+t}{d}} = \mathbb{C}_{\frac{s+t}{d}}(L^2(\Omega); \mathcal{T}_0)$  and*

$$|c|_{\mathbb{C}_{\frac{s+t}{d}}} \lesssim |c|_{X_p^s(\Omega)}. \quad (6.67)$$

*Equivalently for all  $n \geq 1$ ,  $\widetilde{\delta}_{\mathcal{T}}(c)_q$  defined in Section 6.1.2 satisfies*

$$\inf_{\mathcal{T} \in \mathbb{T}_N} \widetilde{\delta}_{\mathcal{T}}(c)_q \lesssim |c|_{X_p^s(\Omega)} N^{-\frac{s+t}{d}} \quad \forall \mathcal{T} \in \mathbb{T}_N, N \geq \#\mathcal{T}_0,$$

*with  $t = 0$  when  $n > 1$ . This error decay is achieved by Algorithm 3.18 (greedy algorithm).*

*Proof.* In view of (6.12), inequality (6.66) is a direct application of Theorem 6.29 (regularity yields approximation). The superconvergence rate in (6.67) is a consequence of (6.12) and the proof of Proposition 3.19 (abstract greedy) with  $s$  replaced by  $s + t$ .  $\square$

We finally go back to the abstract space  $X_p^s(\Omega)$ , defined in (6.35), and introduce the corresponding abstract approximation class  $\mathbb{X}_{\frac{s}{d}} = \mathbb{X}_{\frac{s}{d}}(L^q(\Omega); \mathcal{T}_0)$  of functions  $v \in L^q(\Omega)$  such that

$$|v|_{\mathbb{X}_{\frac{s}{d}}} = \sup_{N \geq \#\mathcal{T}_0} \left( N^{\frac{s}{d}} \inf_{\mathcal{T} \in \mathbb{T}_N} \text{osc}_{\mathcal{T}}(v)_q \right) < \infty \quad \Rightarrow \quad \inf_{\mathcal{T} \in \mathbb{T}_N} \text{osc}_{\mathcal{T}}(v)_q \leq |v|_{\mathbb{X}_{\frac{s}{d}}} N^{-\frac{s}{d}}.$$

Consequently, Theorem 6.29 (regularity yields approximation) implies

$$X_p^s(\Omega) \subset \mathbb{X}_{\frac{s}{d}}, \quad |v|_{\mathbb{X}_{\frac{s}{d}}} \lesssim |v|_{X_p^s(\Omega)}. \quad (6.68)$$

We will utilize this abstract notation and estimates in Section 7 while discussing the approximation of data  $\mathcal{D} = (\mathbf{A}, c, f)$  by a greedy algorithm.

### 6.8.5 Global approximation: inverse estimates

Theorem 6.29 gives sufficient regularity properties for a function  $g \in L^q(\Omega)$  to belong to an approximation class  $\mathbb{A}_{\frac{s}{d}}(L^q(\Omega); \mathcal{T}_0)$ ; this is called *direct estimate*.

Such regularity is written in terms of a Besov space  $B_{p,p}^s(\Omega)$ , except in the limiting case  $s = n + 1$ . The converse statement is also true and is called an *inverse estimate*: if  $g$  belongs to an approximation class  $\mathbb{A}_{\frac{t}{d}}(L^q(\Omega); \mathcal{T}_0)$ , then it is a member of a Besov space  $\widehat{B}_{p,p}^s(\Omega)$  provided  $t > s$  and  $0 < s < n + 1$ ,  $s - \frac{d}{p} + \frac{d}{q} = 0$  [Binev *et al.* 2002, Gaspoz and Morin 2014].

Several comments are in order. The Besov space  $\widehat{B}_{p,p}^s(\Omega)$  is defined via a multilevel decomposition of  $L^p(\Omega)$  and coincides with  $B_{p,p}^s(\Omega)$  only when  $s < 1 + \frac{1}{p}$ . This restriction of  $s$  is natural because  $\mathbb{V}_{\mathcal{T}} \subset \widehat{B}_{p,p}^s(\Omega)$  for all  $s$  but  $\mathbb{V}_{\mathcal{T}} \subset B_{p,p}^s(\Omega)$  requires  $s < 1 + \frac{1}{p}$  according to (6.56). The discrepancy between the spaces  $\widehat{B}_{p,p}^s(\Omega)$  and  $L^q(\Omega)$  is  $\delta = s - \frac{d}{p} + \frac{d}{q} = 0$ , but the decay rate  $t/d$  of  $\mathbb{A}_{\frac{t}{d}}(L^q(\Omega); \mathcal{T}_0)$  is larger than  $s/d$ . This accounts for the embedding of  $\mathbb{A}_{\frac{t}{d}}(L^q(\Omega); \mathcal{T}_0)$

$$\sum_{n \in \mathbb{N}} \left( \sigma_{2^n}(g) 2^{\frac{s}{d}n} \right)^p \leq \sup_{n \in \mathbb{N}} \left( \sigma_{2^n}(g) 2^{\frac{t}{d}n} \right)^p \sum_{n \in \mathbb{N}} 2^{\frac{s-t}{d}pn} \lesssim |g|_{\mathbb{A}_{\frac{t}{d}}}^p$$

into a space with decay  $s/d$  and summability  $\ell^q$  that in turn embeds into  $\widehat{B}_{p,p}^s(\Omega)$  [Binev *et al.* 2002, Gaspoz and Morin 2014]. This reveals that there is no complete characterization of the approximation class  $\mathbb{A}_{\frac{s}{d}}$  in terms of Besov regularity.

## 7 Data Approximation

This section focuses on the module DATA of Algorithms 5.1 (AFEM-TS) and 5.16 (one-step AFEM with switch). According to Assumption 6.11 (quasi-optimality of DATA), the call

$$[\widehat{\mathcal{T}}, \widehat{\mathcal{D}}] = \text{DATA}(\mathcal{T}, \mathcal{D}, \tau) \quad (7.1)$$

is meant to construct a quasi-optimal conforming refinement  $\widehat{\mathcal{T}}$  of  $\mathcal{T} \in \mathbb{T}$  and approximate piecewise polynomial data  $\widehat{\mathcal{D}} = (\widehat{A}, \widehat{c}, \widehat{f}) \in \mathbb{D}_{\widehat{\mathcal{T}}}$  over  $\widehat{\mathcal{T}}$  that satisfies

$$\|\mathcal{D} - \widehat{\mathcal{D}}\|_{D(\Omega)} \leq C_{\text{data}} \tau \quad (7.2)$$

as well as the constraints  $\widehat{A} \in M(\widehat{a}_1, \widehat{a}_2)$  and  $\widehat{c} \in R(\widehat{c}_1, \widehat{c}_2)$  defined in (5.50).

Sections 7.2.2 and 7.2.3 are devoted to the construction of  $(\widehat{A}, \widehat{c})$ . To approximate the coefficients  $(A, c)$  we proceed in two steps. First, we solve an *unconstrained* approximation problem upon computing the  $L^2$ -projection  $(\widetilde{A}, \widetilde{c})$  of  $(A, c)$  onto the space of discontinuous piecewise polynomials of degree  $\leq n - 1$ ; this step is linear, easily achieves the desired accuracy, but does not guarantee the monotonicity of oscillations with respect to refinement (5.72) and violates the constraints in (5.50) unless  $n = 1$ . Second, we resort to the nonlinear selection (5.70) of the local  $L^2$  approximation to force the resulting oscillations to be monotone. Third, we solve a *constrained* problem, which modifies  $(\widetilde{A}, \widetilde{c})$  locally into  $(\widehat{A}, \widehat{c})$  and restores (5.50) without accuracy degradation; this is a delicate nonlinear procedure executed element by element, introduced and discussed in Section 7.2.

The approximation of the right-hand side  $f \in H^{-1}(\Omega)$  is a conceptually different linear process. Without further structural assumptions on  $f$  it is not possible to evaluate  $\text{osc}_{\mathcal{T}}(f)_{-1}$  and reduce it. Hence we introduce surrogate estimators  $\widetilde{\text{osc}}_{\mathcal{T}}(f)_{-1}$ , which are larger than  $\text{osc}_{\mathcal{T}}(f)_{-1}$ , but computable, for several classes of forcing functions  $f$  relevant in practice. We discuss this in Section 7.3.

We start in Section 7.1 with a presentation and assessment of quasi-optimal GREEDY algorithms to reduce the data error. An important consideration is that the local error estimators  $\{\text{osc}_{\mathcal{T}}(v, T)\}_{T \in \mathcal{T}}$  may accumulate in  $\ell^\infty$  as well as in  $\ell^q$  for  $q < \infty$ . Both are handled via a GREEDY algorithm similar to Algorithm 3.18 but with different stopping criteria when the local errors accumulate in  $\ell^q$  with  $q < \infty$ . The module DATA combines both: its structure is displayed in Algorithm 7.23 and its performance is elucidated in Corollary 7.24 below.

### 7.1 Quasi-optimal GREEDY algorithms for data reduction

Algorithm 3.18 (greedy algorithm) is well suited for dealing with local error estimators  $\text{osc}_{\mathcal{T}}(v, T)_q$  that accumulate with respect to  $T \in \mathcal{T}$  in the space  $\ell^\infty$ . This is the framework for approximating coefficients  $v = A, c$  in  $L^\infty(\Omega)$ , in which case the local error estimators  $\text{osc}_{\mathcal{T}}(v, T)_\infty$  are defined in (5.73) for  $r = q = \infty$ . This requires that  $v = A, c$  be piecewise uniformly continuous on  $\mathcal{T}$  for  $\text{osc}_{\mathcal{T}}(v; T)_\infty \rightarrow 0$  as  $h_T \rightarrow 0$ . However, for discontinuous  $(A, c)$  and the forcing function  $f$ , the accumulation of  $\text{osc}_{\mathcal{T}}(v, T)_q$  for  $v = A, c, f$  is in  $\ell^q$  for  $q < \infty$ . In this case, Algorithm 3.18 does not provide a direct relation between a desired output tolerance  $\tau$  for the total error

$$E_{\mathcal{T}}(v)_q := \|\{\text{osc}_{\mathcal{T}}(v, T)_q\}_{T \in \mathcal{T}}\|_{\ell^q} = \left( \sum_{T \in \mathcal{T}} \text{osc}_{\mathcal{T}}(v, T)_q^q \right)^{1/q}$$

and the threshold  $\delta$ ; recall that  $\text{osc}_{\mathcal{T}}(v, T)_q := \|v - \widehat{v}\|_{L^q(T)}$  for  $T \in \mathcal{T}$ .

Another subtle difference from Algorithm 3.18 is that the algorithm GREEDY below does not start from  $\mathcal{T}_0$  but from any  $\mathcal{T} \in \mathbb{T}$ . Since DATA and thus GREEDY is called repeatedly within AFEM, it seems advantageous to exploit the mesh refinement already performed in the adaptive process rather than restarting from scratch; this thus improve the computational efficiency.

**Algorithm 7.1 (GREEDY)** Given a tolerance  $\tau > 0$ ,  $0 < q \leq \infty$ , a number of bisection  $b \geq 1$  performed per element to be refined, and an arbitrary conforming grid  $\mathcal{T} \in \mathbb{T}$ , not necessarily  $\mathcal{T}_0$ , GREEDY finds a conforming refinement  $\widehat{\mathcal{T}} \geq \mathcal{T}$  of  $\mathcal{T}$  by bisection and  $\widehat{v} \in \mathbb{S}_{\widehat{\mathcal{T}}}^{n-1, -1}$  such that  $E_{\widehat{\mathcal{T}}}(v)_q \leq \tau$ :

```

 $[\widehat{\mathcal{T}}, \widehat{v}] = \text{GREEDY}(\mathcal{T}, \tau, q, b, v)$ 
 $[\widehat{v}] = \text{PROJECT}(\widehat{\mathcal{T}}, v)$ 
while  $E_{\widehat{\mathcal{T}}}(v)_q > \tau$ 
   $[\mathcal{M}] = \text{argmax}\{\text{osc}_{\mathcal{T}}(v, T)_q : T \in \mathcal{T}\}$ 
   $[\mathcal{T}] = \text{REFINE}(\mathcal{T}, \mathcal{M}, b)$ 

```

$[\widehat{v}] = \text{PROJECT}(\mathcal{T}, v)$   
 return  $\mathcal{T}, \widehat{v}$

In GREEDY above, the element  $T$  with largest error is refined as long as the total error  $E_{\mathcal{T}}(v)_q$  exceeds the target tolerance  $\tau$ . When the largest error is achieved by several elements, an ad-hoc criteria such as lexical order is used to break ties. We also recall that the routine REFINE bisects all the elements in  $\mathcal{M}$  (in this case only one)  $b$  times and performs additional refinements necessary to produce a conforming subdivision. PROJECT computes the local approximations  $\widehat{v}$  of  $v$  needed to evaluate  $\text{osc}_{\mathcal{T}}(v, T)_q$ ; refer to Section 5.4.2 and (5.70) for the definition of  $\widehat{v}$ . The dependency on  $\widehat{v}$  in  $\text{osc}_{\mathcal{T}}(v, T)_q$  and  $E_{\mathcal{T}}(v)$  is not indicated.

To discuss the performances of the GREEDY algorithm, we recall that  $X_p^s(\Omega; \mathcal{T}_0)$  is the abstract space defined in (6.35) which satisfies

$$\sum_{T \in \mathcal{T}} |v|_{X_p^s(T)}^p \lesssim |v|_{X_p^s(\Omega; \mathcal{T}_0)}^p \quad (7.3)$$

for all  $\mathcal{T} \in \mathbb{T}$  and  $v \in X_p^s(\Omega; \mathcal{T}_0)$ .

The GREEDY algorithm analyzed in Proposition 3.19 (abstract greedy) relies on the abstract assumptions (3.40), (3.41), and (3.42). With the aim of reducing the data oscillations, we make these assumptions more concrete.

**Assumption 7.2 (admissible set of parameters for GREEDY)** We say that the set of parameters  $(v, s, t, p, q)$  is admissible for GREEDY with local oscillations  $\{\text{osc}_{\mathcal{T}}(v, T)_q\}_{T \in \mathcal{T}}$  if  $0 < p, q \leq \infty, s, t \geq 0$  satisfy

- (i)  $v \in X_p^s(\Omega; \mathcal{T}_0)$ ;
- (ii)  $t + s > 0, \quad s - \frac{d}{p} + \frac{d}{q} \geq 0 \quad \text{with strict inequality when } q = \infty \text{ or } s = n + 1,$   
 $p < 1$ ;
- (iii) for  $r := t + s - \frac{d}{p} + \frac{d}{q} > 0$

$$\text{osc}_{\mathcal{T}}(v, T)_q \lesssim h_T^r |v|_{X_p^s(T)} \quad \forall T \in \mathcal{T}, \mathcal{T} \in \mathbb{T}. \quad (7.4)$$

When the local oscillations considered are clear from the context, we say that  $(v, s, t, p, q)$  is admissible for GREEDY.

Relation (7.4) replaces (3.41) and is a regularity assumption guaranteeing a convergence rate when approximating  $v$  by  $\widehat{v} = \text{PROJECT}(\mathcal{T}, v)$  (appearing in the definition of  $\text{osc}_{\mathcal{T}}(v, T)_q$ ). We refer to Propositions 6.32, 6.33, and 6.34 for examples where Assumption 7.2 holds. Note that in view of (6.55) and (6.59), the conditions (ii) in Assumption 7.2 guarantees  $v \in X_p^s(\Omega) \subset L^q(\Omega)$ . The parameter  $t \geq 0$  reflects a possible additional power of  $h$  in the oscillation term, see for e.g. (5.75), (5.76) and (5.78). Furthermore, in view of (7.3) assumption (6.51) is always satisfied by the  $X_p^s(\Omega; \mathcal{T}_0)$  semi-norms, and (3.28) is not needed any longer.

As alluded to above, the case  $q < \infty$  is more complex to analyze and cannot solely rely on the decay property (7.4) as in the proof of Proposition 3.19. It requires the local oscillations to be monotone with respect to refinements.



**Assumption 7.3 (monotonicity of local oscillations)** We say that for  $0 < q \leq \infty$ , the local errors satisfy the monotonicity property in  $\ell^q$  if for any  $v \in L^q(\Omega)$ , any  $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$  with  $\mathcal{T}_* \geq \mathcal{T}$  and any  $T_* \in \mathcal{T}_*, T \in \mathcal{T}$  with  $T_* \subset T$ , we have

$$\text{osc}_{\mathcal{T}_*}(v, T_*)_q \leq \text{osc}_{\mathcal{T}}(v, T)_q. \quad (7.5)$$

In view of Lemma 5.26 (monotonicity of oscillation), Assumption 7.3 holds for the oscillations on  $\mathbf{A}$  and  $c$  given in (5.73) and (5.75), but not for the oscillations (5.77) of  $f$ . However, in Section 7.3 below we derive computable surrogates for the local error  $\text{osc}_{\mathcal{T}}(f, T)_{-1}$ . These surrogates satisfy the monotonicity property and are used in turn to drive the GREEDY algorithm. In passing, we note that we refrain from using the right-hand side of inequality (7.4) as surrogate for the local oscillation. In fact, it is monotone with respect to refinements but at the expense of being difficult to evaluate because it involves the semi-norm  $|v|_{X_p^s(\mathcal{T})}$ .

The following result is the counterpart of Proposition 3.19 (abstract greedy) for GREEDY with errors accumulating in  $\ell^q$ ,  $0 < q \leq \infty$ , and still starting from  $\mathcal{T}_0$ . We address the case where  $\mathcal{T} \neq \mathcal{T}_0$  in Lemma 7.5 below.

**Proposition 7.4 (performance of GREEDY)** *Let the initial subdivision  $\mathcal{T}_0$  of  $\Omega \subset \mathbb{R}^d$  satisfy Assumption 6.19 (initial labeling). Let  $\tau > 0$  be the target tolerance and  $b \geq 1$  be the number of bisections performed on each marked element. Let  $(v, s, t, p, q)$  satisfy Assumption 7.2 (admissible set of parameters for GREEDY) with local errors  $\{\text{osc}_{\mathcal{T}}(v, T)_q\}_{T \in \mathcal{T}}$  which in turn verify Assumption 7.3 (monotonicity of local oscillations) in  $\ell^q$ . Then GREEDY( $\mathcal{T}_0, \tau, q, b, v$ ) terminates in a finite number of iterations and*

$$E_{\mathcal{T}}(v)_q \leq \tau \leq C |v|_{X_p^s(\Omega; \mathcal{T}_0)} (\#\mathcal{T})^{-\frac{s+t}{d}}. \quad (7.6)$$

with a constant  $C = C(p, q, s, b, d, \Omega, \mathcal{T}_0)$ . Furthermore,  $v \in \mathbb{X}_{\frac{s+t}{d}}$  and  $|v|_{\mathbb{X}_{\frac{s+t}{d}}} \lesssim |v|_{X_p^s(\Omega; \mathcal{T}_0)}$ . Moreover, the estimate (7.6) is valid for tensor-valued functions  $v$ .

*Proof.* Since the proof is similar to that of Proposition 3.19 (abstract greedy) with  $e_{\mathcal{T}}(v, T)_q = \text{osc}_{\mathcal{T}}(v, T)_q$ , we only report the new ingredients. We recall that we use the convention  $1/\infty = 0$ . Let  $\mathcal{T}_1, \dots, \mathcal{T}_k$  be the sequence of refinements produced by GREEDY, and  $T_1, \dots, T_k$  be the sequence of marked elements. We need to estimate  $\#\mathcal{M} = k$  with  $\mathcal{M} = \{T_1, \dots, T_k\}$ . Set

$$\delta_i := \text{osc}_{\mathcal{T}_i}(v, T_i)_q, \quad (1 \leq i \leq k) \quad \text{and} \quad \delta := \delta_{k-1}.$$

Then, there holds

$$E_{\mathcal{T}_k}(v)_q \leq \tau < E_{\mathcal{T}_{k-1}}(v)_q \leq \delta (\#\mathcal{T}_{k-1})^{\frac{1}{q}} \leq \delta (\#\mathcal{T}_k)^{\frac{1}{q}}. \quad (7.7)$$

On the other hand, since REFINE does not increase the element estimators  $\text{osc}_{\mathcal{T}_i}(v, T_i)$  thanks to (7.5), one has  $\delta_i \geq \delta$  for any  $i$ , whence

$$\text{osc}_{\mathcal{T}_i}(v, T_i)_q = \delta_i \geq \delta, \quad \forall 1 \leq i < k.$$

Let us now partition  $\mathcal{M}$  into disjoint subsets  $\mathcal{P}_j$  as in the proof of Proposition 3.19.

If  $T_i \in \mathcal{P}_j$ , (7.4) implies

$$\delta \leq \text{osc}_{\mathcal{T}_i}(v, T_i)_q \lesssim h_T^r |v|_{X_P^s(T_i)} \leq 2^{-\frac{jr}{2}} |v|_{X_P^s(T_i)},$$

whence, exploiting the  $\ell^p$  summability (7.3) gives

$$\#\mathcal{P}_j \lesssim \delta^{-p} 2^{-\frac{jr p}{2}} |v|_{X_P^s(\Omega; \mathcal{T}_0)}^p.$$

which is similar to (3.45). Recalling (3.44), and proceeding as in the proof of Proposition 3.19, yields

$$\delta \lesssim |v|_{X_P^s(\Omega; \mathcal{T}_0)} (\#\mathcal{T}_k - \#\mathcal{T}_0)^{-\frac{s+t}{d} - \frac{1}{q}}.$$

We conclude the proof using (7.7) and the bound  $\#\mathcal{T}_k \geq c_0 \#\mathcal{T}_0$  for  $c_0 > 1$ .  $\square$

In contrast to Section 3, and most of the existing literature, Algorithm 7.1 starts from a refinement  $\mathcal{T}$  of  $\mathcal{T}_0$  rather than  $\mathcal{T}_0$  and thus exploits the mesh refinement already performed in the adaptive process. We now give a simple argument, updated from [Bonito *et al.* 2013b], that shows that the number of elements  $N(\mathcal{T}, \tau, b, v)$  marked by GREEDY starting from  $\mathcal{T}$  with target tolerance  $\tau$  and refined  $b \geq 1$  times is dominated by  $N(\mathcal{T}_0, \tau, 1, v)$ , namely

$$N(\mathcal{T}, \tau, b, v) \leq N(\mathcal{T}_0, \tau, 1, v). \quad (7.8)$$

Estimate (7.8) is crucial because it avoids studying the cardinality of GREEDY starting from  $\mathcal{T} \neq \mathcal{T}_0$  directly, and simplifies the analysis. Even though (7.8) is plausible, the fact that the output of GREEDY  $(\mathcal{T}_0, \tau, q, 1, v)$  is unrelated to  $\mathcal{T}$  makes it non-obvious. In fact, note that we do not claim that  $N(\mathcal{T}, \tau, b, v) \leq N(\mathcal{T}_0, \tau, b, v)$ , which is unclear. The proof presented below hinges on the fact that all the elements refined within GREEDY  $(\mathcal{T}_0, \tau, q, 1, v)$  are either refined because they are marked by GREEDY (and thus of largest oscillation) or because their refinement is necessary to guarantee conformity of the resulting subdivision. For our purposes, (7.8) suffices.

**Lemma 7.5 (GREEDY starting from  $\mathcal{T}$ )** *Let  $\tau > 0$  be a target tolerance and  $b \geq 1$  be the number of bisections per marked element. Assume that the local errors employed by GREEDY satisfy Assumption 7.3 (monotonicity of local oscillations) in  $\ell^q$ . Then, the number of elements  $N(\mathcal{T}, \tau, b, v)$  marked by GREEDY  $(\mathcal{T}, \tau, q, b, v)$  satisfies (7.8) for any admissible refinement  $\mathcal{T} \in \mathbb{T}$  of  $\mathcal{T}_0$ .*

*Proof.* We simply write GREEDY  $(\mathcal{T}_0, \tau, 1)$  and GREEDY  $(\mathcal{T}, \tau, b)$  because  $v$  and  $q$  are fixed. Let  $N := N(\mathcal{T}_0, \tau, 1, v)$ , and recall that the bisection rules define a unique forest  $\mathbb{T}$  emanating from  $\mathcal{T}_0$  and a unique sequence of elements  $\{T_i\}_{i=1}^N$  marked by GREEDY  $(\mathcal{T}_0, \tau, 1)$ . We denote by  $\{\mathcal{T}^i\}_{i=1}^N$  the sequence of intermediate subdivisions built within GREEDY  $(\mathcal{T}_0, \tau, 1)$  starting with  $\mathcal{T}^0 = \mathcal{T}_0$ :  $T_i \in \mathcal{T}^{i-1}$  is bisected once by REFINE which also produces the smallest conforming refinement  $\mathcal{T}^i$  of  $\mathcal{T}^{i-1}$  containing the two children of  $T_i$ . We thus say that GREEDY  $(\mathcal{T}_0, \tau, 1)$  satisfies the *minimality property* that all the elements refined are either marked elements

because their error is largest or necessary to guarantee conforming subdivisions. Notice that this is not true for GREEDY  $(\mathcal{T}_0, \tau, b)$  when  $b > 1$ .

For any  $\mathcal{T} \in \mathbb{T}$ , we let  $\Lambda_{\mathcal{T}}$  be the set of indices  $j \in \{1, \dots, N\}$  such that  $T_j$  is never refined in the process to create  $\mathcal{T}$ , i.e.  $T_j$  is either an element of  $\mathcal{T}$  or a successor of an element of  $\mathcal{T}$ . We show that

$$N(\mathcal{T}, \tau, b, v) \leq \#\Lambda_{\mathcal{T}} \quad (7.9)$$

by induction on  $\#\Lambda_{\mathcal{T}}$ . If  $\#\Lambda_{\mathcal{T}} = 0$  then  $\mathcal{T}$  is a refinement of  $\mathcal{T}^N$ , whence the monotonicity of the total error

$$E_{\mathcal{T}}(v)_q \leq E_{\mathcal{T}^N}(v)_q \leq \tau,$$

guaranteed by (7.5), implies that  $N(\mathcal{T}, \tau, b, v) = 0$ ; this satisfies (7.9) as desired.

We now assume that (7.9) is valid for any  $\mathcal{T} \in \mathbb{T}$  such that  $\#\Lambda_{\mathcal{T}} \leq k$ , a non-negative integer, and deduce it must also hold for any  $\mathcal{T} \in \mathbb{T}$  such that  $\#\Lambda_{\mathcal{T}} \leq k+1$ . Let  $\mathcal{T} \in \mathbb{T}$  be one such mesh, namely  $\#\Lambda_{\mathcal{T}} = k+1$ . If  $E_{\mathcal{T}}(v)_q \leq \tau$  then  $N(\mathcal{T}, \tau, b, v) = 0$  and  $N(\mathcal{T}, \tau, b, v) \leq \#\Lambda_{\mathcal{T}}$  holds trivially.

When instead  $E_{\mathcal{T}}(v)_q > \tau$ , we let  $j$  be the smallest index in  $\Lambda_{\mathcal{T}}$  and show that  $T_j \in \mathcal{T}$  using the minimality property of GREEDY  $(\mathcal{T}_0, \tau, 1)$ . Assume by contradiction that  $T_j \notin \mathcal{T}$  but  $T_j$  belongs to a refinement  $\tilde{\mathcal{T}}$  of  $\mathcal{T}$  and is thus a successor of an element  $T \in \mathcal{T}$ . Note that  $T$  is refined by GREEDY  $(\mathcal{T}_0, \tau, 1)$  to produce  $T_j$  but was not marked because otherwise  $T = T_i$  for some  $i < j$  and  $i \in \Lambda_{\mathcal{T}}$ , which would contradict the minimality of  $j$ . Hence,  $T$  must have been refined by the REFINE routine to guarantee conformity when bisecting a marked element  $T_{\ell}$ ,  $\ell < j$ . Invoking the minimality of  $j$  again yields that  $\ell \notin \Lambda_{\mathcal{T}}$  and  $T_{\ell}$  cannot be in  $\mathcal{T}$  because  $T_{\ell}$  has been refined to get to  $\mathcal{T}$  by definition of  $\Lambda_{\mathcal{T}}$ . Since REFINE refines the minimal number of non-marked elements to guarantee conformity, and  $\mathcal{T}$  is conforming,  $T$  must have been refined as well when refining  $T_{\ell}$  in the process of constructing  $\mathcal{T}$  and therefore cannot be in  $\mathcal{T}$ . This is a contradiction and  $T_j \in \mathcal{T}$ .

Therefore,  $\mathcal{T}$  is a refinement of  $\mathcal{T}^{j-1}$  because all the elements marked or refined to ensure conformity by GREEDY  $(\mathcal{T}_0, \tau, 1)$  have been refined in the process of creating  $\mathcal{T}$ . Moreover,  $T_j \in \mathcal{T}$  is the element with largest error  $\text{osc}_{\mathcal{T}}(v, T_j)$  within  $\mathcal{T}$  (with ad-hoc criteria to break ties), because  $\text{osc}_{\mathcal{T}^{j-1}}(v, T_j)$  is largest in  $\mathcal{T}^{j-1}$  by definition of  $T_j$  and monotonicity of the local error (7.5); hence  $T_j$  must be the first element marked by GREEDY  $(\mathcal{T}, \tau, b)$ . Let  $\mathcal{T}^*$  be the subdivision obtained from  $\mathcal{T}$  upon bisecting  $b$  times  $T_j$ . Notice that  $\Lambda_{\mathcal{T}^*}$  is a strict subset of  $\Lambda_{\mathcal{T}}$ , because  $j \notin \Lambda_{\mathcal{T}^*}$ , so that the induction assumption yields

$$N(\mathcal{T}, \tau, b, v) = 1 + N(\mathcal{T}^*, \tau, b, v) \leq 1 + \#\Lambda_{\mathcal{T}^*} \leq \#\Lambda_{\mathcal{T}}.$$

This proves (7.9) and (7.8) follows immediately since  $\#\Lambda_{\mathcal{T}} \leq N(\mathcal{T}_0, \tau, 1, v)$ .  $\square$

Estimate (7.8) is critical to analyze the performances of GREEDY starting from any admissible subdivision  $\mathcal{T} \in \mathbb{T}$ . We emphasize that the complexity estimate provided by Corollary 7.6 is expressed in terms of number of marked elements

$N(\mathcal{T}, \tau, q, b, v)$  and tolerance  $\tau$  instead of error and cardinality of  $\mathcal{T}$ . This is why GREEDY can start from any mesh  $\mathcal{T} \in \mathbb{T}$ .

**Corollary 7.6 (performance of greedy)** *Let the initial subdivision  $\mathcal{T}_0$  of  $\Omega \subset \mathbb{R}^d$  satisfy Assumption 6.19 (initial labeling) and  $\mathcal{T} \in \mathbb{T}$  be any admissible refinement of  $\mathcal{T}_0$ . Let  $\tau > 0$  be the target tolerance and  $b \geq 1$  be the number of bisections performed on each marked element. Let  $(v, s, t, p, q)$  satisfy Assumption 7.2 (admissible set of parameters for GREEDY) with local errors  $\{\text{osc}_{\mathcal{T}}(v, T)_q\}_{T \in \mathcal{T}}$  which in turn verify Assumption 7.3 (monotonicity of local oscillations) in  $\ell^q$ . The number of marked elements  $N(\mathcal{T}, \tau, q, b, v)$  by GREEDY( $\mathcal{T}, \tau, q, b, v$ ) satisfies*

$$N(\mathcal{T}, \tau, q, b, v) \leq C |v|_{X_P^s(\Omega)}^{\frac{d}{s+t}} \tau^{-\frac{d}{s+t}}. \quad (7.10)$$

with a constant  $C = C(p, q, s, b, d, \Omega, \mathcal{T}_0)$ . Moreover, the estimate (7.10) is valid for tensor-valued functions  $v$ .

*Proof.* Invoking Proposition 7.4 (performance of GREEDY), which gives rise to a mesh  $\widehat{\mathcal{T}}$ , and Lemma 7.5 (GREEDY starting from  $\mathcal{T}$ ), we readily deduce

$$N(\mathcal{T}, \tau, q, b, v) \leq N(\mathcal{T}_0, \tau, q, 1, v) \leq \#\widehat{\mathcal{T}} \leq C |v|_{X_P^s(\Omega)}^{\frac{d}{s+t}} \tau^{-\frac{d}{s+t}},$$

which is the desired inequality (7.10).  $\square$

## 7.2 Constrained approximations

We discuss how the approximations produced by GREEDY (see Corollary 7.6) can be modified to satisfy the structural assumption (5.51) without sacrificing their accuracy.

### 7.2.1 Constrained approximations of scalar functions

The approximate data  $\widetilde{\mathcal{D}} = (\widetilde{\mathbf{A}}, \widetilde{c}, \widetilde{f})$  constructed in the previous sections using the GREEDY algorithm are not guaranteed to satisfy the necessary conditions for perturbed problem (5.5) with  $\widehat{\mathcal{D}} = \widetilde{\mathcal{D}}$  to have a solution  $\widehat{u} = \widehat{u}(\widehat{\mathcal{D}}) \in H_0^1(\Omega)$ . Recall that the data  $\mathcal{D} = (\mathbf{A}, c, f) \in D(\Omega)$  is assumed to satisfy the structural assumption (5.50), i.e.,  $\mathbf{A} \in M(\alpha_1, \alpha_2)$  and  $c \in R(c_1, c_2)$  with  $0 < \alpha_1 \leq \alpha_2$  and  $0 \leq c_1 \leq c_2$ . It turns out that constructing approximate data  $\widehat{\mathcal{D}}$  with the same constraints is a difficult task. We follow [Bonito et al. 2013b] and modify the data  $\widetilde{\mathcal{D}}$  to obtain  $\widehat{\mathcal{D}} = (\widehat{\mathbf{A}}, \widehat{c}, \widehat{f})$  in such a way that the approximation property of  $\widetilde{\mathcal{D}}$  is preserved,

$$\|\mathcal{D} - \widehat{\mathcal{D}}\|_{D(\Omega)} \leq C_{\text{data}} \|\mathcal{D} - \widetilde{\mathcal{D}}\|_{D(\Omega)},$$

while ensuring that

$$\widehat{\mathbf{A}} \in M\left(\frac{\alpha_1}{2}, C\widehat{\alpha}_2\right), \quad \widehat{c} \in R\left(-\frac{\alpha_1}{4C_P^2}, C\widehat{c}_2\right). \quad (7.11)$$

Here  $C$  is a constant independent of relevant quantities (we make this more precise below). In particular, the data  $\widehat{\mathcal{D}}$  satisfies the structural assumption (7.11) which guarantees that the perturbed problem (5.5) has a unique solution. Note that the general case is more subtle than when the data are approximated by piecewise constant approximations (5.74), which are directly satisfying the structural assumption and used as motivation in Section 5.4.2.

We start by discussing a process modifying the approximation of a strictly positive scalar function  $v \in L^\infty(\Omega)$ , i.e.  $v \in R(c_1, c_2)$  for some  $0 < c_1 \leq c_2$ ; see (5.49). Because the polynomial degree used to approximate the data might differ depending on the application, we use  $m \in \mathbb{N}$  to denote a generic polynomial degree. We think of  $\widetilde{v} \in \mathbb{S}_{\mathcal{T}}^{m,-1}$  to be an approximation to  $v$  not necessarily strictly positive. The following process modifies  $\widetilde{v}$  locally to construct  $\widehat{v} \in \mathbb{S}_{\mathcal{T}}^{m,-1}$ . It involves a parameter  $L > 2$  responsible for the truncation of  $\widetilde{v}$  whenever it is too large, i.e.,  $\widetilde{v} \geq Lc_2$ . For  $T \in \mathcal{T}$ , we set  $\widehat{v}|_T := \widehat{v}_T$  where

$$\widehat{v}_T := \begin{cases} c_2 & \text{when } \|\widetilde{v}\|_{L^\infty(T)} \geq Lc_2, \\ \widetilde{v}|_T - \min_{x \in T} \widetilde{v}(x) + \frac{c_1}{2} & \text{when otherwise } \min_{x \in T} \widetilde{v}(x) < \frac{c_1}{2}, \\ \widetilde{v}|_T & \text{otherwise.} \end{cases} \quad (7.12)$$

Corollary 7.9 below is in essence Proposition 3 of [Bonito *et al.* 2013b] and states that the constructed  $\widehat{v}$  satisfies

$$0 < \frac{c_1}{2} \leq \widehat{v} \leq \left( \frac{1}{2} + 2L \right) c_2, \quad \text{a.e. in } \Omega.$$

This is at the expense of inflating the approximation error in  $L^q$ ,  $1 \leq q \leq \infty$ , by a multiplicative constant  $C$  only depending on  $d$ ,  $m$ ,  $c_2/c_1$ ,  $q$ ,  $L$  and the shape regularity of  $\mathbb{T}$

$$\|v - \widehat{v}\|_{L^q(T)} \leq C \|v - \widetilde{v}\|_{L^q(T)}.$$

In preparation for this result, we introduce the following notations. We denote by  $C_I$  the smallest constant such that for any  $T \in \mathcal{T}$  and any polynomial  $P \in \mathbb{P}_m(T)$ , the inverse inequality

$$\|\nabla P\|_{L^\infty(T)} \leq C_I \|P\|_{L^\infty(T)} |T|^{-1/d}, \quad (7.13)$$

holds. The inverse inequality constant  $C_I$  only depends on the shape regularity of  $\mathbb{T}$ ,  $m$  and  $d$ . Note that for such polynomial  $P \in \mathbb{P}_m(T)$ , we have

$$|P(x) - P(y)| \leq C_I \|P\|_{L^\infty(T)} |T|^{-1/d} |x - y|, \quad \forall x, y \in T.$$

Consequently, for any  $\rho > 0$  and  $x \in T$ , we define

$$T(x, \rho) := T \cap \overline{B(x, \rho |T|^{1/d} / C_I)},$$

which is motivated by the fact that for  $x \in T$  and  $y \in T(x, \rho)$  we have

$$|P(x) - P(y)| \leq C_I \|P\|_{L^\infty(T)} |T|^{-1/d} |x - y| \leq \rho \|P\|_{L^\infty(T)}. \quad (7.14)$$

Critical for the analysis below is the existence of a constant  $0 < C_S(\rho) \leq 1$  depending on  $\rho$  but also on  $d, m$ , and the shape regularity of  $\mathbb{T}$ , such that

$$|T(x, \rho)| \geq C_S(\rho)|T| \quad \forall x \in T, \quad T \in \mathcal{T}. \quad (7.15)$$

This constant  $C_S(\rho)$  assesses the area of a subset of  $T$  where the polynomial  $P$  varies no more than  $\rho\|P\|_{L^\infty(T)}$  away from  $P(x)$ .

We are now in position to analyze the effect of the nonlinear correction (7.12). We proceed locally over each  $T \in \mathcal{T}$  and start with the case where  $\|\tilde{v}\|_{L^\infty(T)}$  is large (Lemma 7.7). We then discuss the case where  $\tilde{v}(x)$  is small on  $T$  (Lemma 7.8) while for the remaining case, the function  $\tilde{v}$  does not need to be modified on  $T$ . These three cases are collected in Corollary 7.9 for scalar valued functions and in Corollary 7.11 for matrix valued functions. In all the arguments below we used the convention  $a^{1/\infty} = 1$  for any  $a > 0$ .

**Lemma 7.7 (locally enforcing constraints for large approximations)** *Let  $\mathcal{T} \in \mathbb{T}$  be any conforming refinement of  $\mathcal{T}_0$  satisfying Assumption 6.19 (initial labelling). Let  $c_2 > 0$ ,  $T \in \mathcal{T}$  and  $v_T \in L^\infty(T)$  satisfying  $0 < v_T \leq c_2$  a.e. in  $T$ . Furthermore, for  $m \geq 0$  and  $L > 2$ , assume that  $\tilde{v}_T \in \mathbb{P}_m(T)$  satisfies*

$$\|\tilde{v}_T\|_{L^\infty(T)} \geq Lc_2. \quad (7.16)$$

*Then for the constant function  $\widehat{v}_T := c_2 \in \mathbb{P}_m(T)$  there holds*

$$\frac{c_1}{2} < \widehat{v}_T < Lc_2 < \left(\frac{1}{2} + 2L\right) c_2.$$

*Moreover, for  $1 \leq q \leq \infty$ , we have*

$$\|v_T - \widehat{v}_T\|_{L^q(T)} \leq C_2^+ \|v_T - \tilde{v}_T\|_{L^q(T)},$$

*where  $C_2^+ := \frac{4C_S^{-1/q}}{L-2}$  and  $C_S = C_S(1/2)$  is the constant appearing in (7.15) with  $\rho = 1/2$ .*

*Proof.* Let  $x_0 \in T$  and  $\widetilde{c}_{2,T}$  defined by the relation

$$\widetilde{c}_{2,T} := |\tilde{v}_T(x_0)| := \|\tilde{v}_T\|_{L^\infty(T)}.$$

In view of the Lipschitz property (7.14) applied to  $P = \tilde{v}_T$  and with  $\rho = \frac{1}{2}$ , we have

$$|\tilde{v}_T(x) - \tilde{v}_T(x_0)| \leq \frac{\widetilde{c}_{2,T}}{2},$$

for  $x \in T_0 := T(x_0, \frac{1}{2}) \subset T$ . Recall (7.15), which implies that  $|T_0| \geq \widetilde{C}_S|T|$  for some constant  $\widetilde{C}_S := C_S(1/2)$  only depending  $d, n$ , and the shape regularity of  $\mathbb{T}$ . On the one hand, this implies that  $\tilde{v}_T|_{T_0}$  is bounded below with  $|\tilde{v}_T(x)| \geq \frac{\widetilde{c}_{2,T}}{2}$  for  $x \in T_0$  and, on the other hand,  $v_T$  is bounded from above by

$$0 \leq v_T(x) \leq c_2 \leq L^{-1}\widetilde{c}_{2,T}, \quad x \in T.$$

Consequently, for  $x \in T_0$  and since  $L > 2$ , there holds

$$0 \leq v_T(x) \leq L^{-1} \tilde{c}_{2,T} \leq \frac{\tilde{c}_{2,T}}{2} \leq |\tilde{v}_T(x)|$$

and thus

$$|v_T(x) - \tilde{v}_T(x)| \geq |\tilde{v}_T(x)| - v_T(x) \geq \left(\frac{1}{2} - \frac{1}{L}\right) \tilde{c}_{2,T} = \frac{L-2}{2L} \tilde{c}_{2,T},$$

which indicates that  $v_T$  and  $\tilde{v}_T$  are sufficiently far apart on a substantial portion  $T_0$  of  $T$ . Thus is responsible for the  $L^q$ -bound below. In fact, we have

$$\|v_T - \tilde{v}_T\|_{L^q(T)} \geq \|v_T - \tilde{v}_T\|_{L^q(T_0)} \geq \frac{L-2}{2L} \tilde{c}_{2,T} |T_0|^{1/q}, \quad (7.17)$$

whence, from the definition  $\hat{v}_T := c_2$  and using (7.15), we deduce

$$\|v_T - \hat{v}_T\|_{L^q(T)} \leq 2c_2 |T|^{1/q} \leq 2L^{-1} \tilde{c}_{2,T} |T|^{1/q} \leq \frac{4C_S^{-1/q}}{L-2} \|v_T - \tilde{v}_T\|_{L^q(T)}$$

as desired.  $\square$

**Lemma 7.8 (locally enforcing constraints for small approximations)** *Let  $\mathcal{T} \in \mathbb{T}$  be any conforming refinement of  $\mathcal{T}_0$  satisfying Assumption 6.19 (initial labelling). Let  $0 < c_1 \leq c_2$ ,  $T \in \mathcal{T}$  and  $v_T \in L^\infty(T)$  satisfying  $c_1 < v_T \leq c_2$  a.e. in  $T$ . Furthermore, for  $m \geq 0$  and  $L > 2$  assume that  $\tilde{v}_T \in \mathbb{P}_m(T)$  satisfies*

$$\|\tilde{v}_T\|_{L^\infty(T)} \leq Lc_2 \quad (7.18)$$

and

$$\min_{x \in T} \tilde{v}_T(x) < \frac{c_1}{2}. \quad (7.19)$$

Then the function  $\hat{v}_T := \frac{c_1}{2} + \tilde{v}_T - \min_{x \in T} \tilde{v}_T(x) \in \mathbb{P}_n(T)$  is such that

$$\frac{c_1}{2} \leq \hat{v}_T \leq 2Lc_2 + \frac{c_1}{2} \leq \left(2L + \frac{1}{2}\right) c_2$$

and

$$\|v_T - \hat{v}_T\|_{L^q(T)} \leq C_1^+ \|v_T - \tilde{v}_T\|_{L^q(T)},$$

where  $C_1^+ := (1 + C_S^{-1/q}(\rho))$  and  $C_S(\rho)$  is the constant appearing in (7.15) with  $\rho = c_1/(2Lc_2)$ .

*Proof.* We define  $x_0 \in T$ ,  $\tilde{c}_{1,T} \in \mathbb{R}$  by the relations

$$\tilde{c}_{1,T} := \tilde{v}_T(x_0) := \min_{x \in T} \tilde{v}_T(x).$$

From the Lipschitz property (7.14) and the assumption (7.18), we find that

$$|\tilde{v}_T(x) - \tilde{v}_T(x_0)| \leq \frac{c_1}{2}$$

for  $x \in T_0 := T(x_0, \rho)$  with  $\rho := \frac{c_1}{2Lc_2}$ . Recall (7.15), which implies that  $|T_0| \geq \tilde{C}_S |T|$  for some constant  $\tilde{C}_S := C_S(\rho)$  only depending  $d, m, c_2/c_1, L$ , and the shape regularity of  $\mathbb{T}$ .

For  $x \in T_0$ , we proceed by estimating the difference

$$v_T(x) - \tilde{v}_T(x) = v_T(x) - (\tilde{v}_T(x) - \tilde{v}_T(x_0)) - \tilde{v}_T(x_0) \geq c_1 - \frac{c_1}{2} - \tilde{c}_{1,T} = \frac{c_1}{2} - \tilde{c}_{1,T} > 0$$

because  $\tilde{c}_{1,T} < c_1/2$  by assumption (7.19). This implies that

$$|T_0|^{1/q} \left( \frac{c_1}{2} - \tilde{c}_{1,T} \right) \leq \|v_T - \tilde{v}_T\|_{L^q(T)},$$

and  $v_T$  and  $\tilde{v}_T$  are uniformly far apart in the substantial part  $T_0$  of  $T$ . Therefore,  $\widehat{v}_T := \tilde{v}_T + (\frac{c_1}{2} - \tilde{c}_{1,T})$  satisfies

$$\frac{c_1}{2} \leq \widehat{v}_T \leq 2Lc_2 + \frac{c_1}{2}$$

because  $\tilde{c}_{1,T} \geq -\|\tilde{v}\|_{L^\infty(T)} \geq -Lc_2$  by assumption (7.18), and

$$\begin{aligned} \|v_T - \widehat{v}_T\|_{L^q(T)} &\leq \|v_T - \tilde{v}_T\|_{L^q(T)} + \left( \frac{c_1}{2} - \tilde{c}_{1,T} \right) |T|^{1/q} \\ &\leq \|v_T - \tilde{v}_T\|_{L^q(T)} + \left( \frac{c_1}{2} - \tilde{c}_{1,T} \right) \tilde{C}_S^{-1/q} |T_0|^{1/q} \\ &\leq (1 + \tilde{C}_S^{-1/q}) \|v_T - \tilde{v}_T\|_{L^q(T)}. \end{aligned}$$

This proves the assertions.  $\square$

**Corollary 7.9 (locally enforcing constraints)** *Let  $\mathcal{T} \in \mathbb{T}$  be any conforming refinement of  $\mathcal{T}_0$  satisfying Assumption 6.19 (initial labelling). Let  $0 < c_1 \leq c_2$ ,  $T \in \mathcal{T}$  and  $v_T \in L^\infty(T)$  satisfying  $c_1 \leq v_T \leq c_2$  a.e. in  $T$ . Then, for  $m \geq 0$ ,  $L > 2$ , and  $\tilde{v}_T \in \mathbb{P}_m(T)$ , the function  $\widehat{v}_T \in \mathbb{P}_m(T)$  defined in (7.12) satisfies*

$$\frac{c_1}{2} \leq \widehat{v}_T \leq \left( \frac{1}{2} + 2L \right) c_2 \quad \text{a.e. in } T.$$

Moreover for  $1 \leq q \leq \infty$ , we have

$$\|v_T - \widehat{v}_T\|_{L^q(T)} \leq \max(C_1^+, C_2^+) \|v_T - \tilde{v}_T\|_{L^q(T)} \quad \forall T \in \mathcal{T},$$

where  $C_1^+$  and  $C_2^+$  are the constants appearing in Lemmas 7.8 and 7.7, which only depend on  $d, m, c_2/c_1, L$ , and the shape regularity of  $\mathbb{T}$ .

*Proof.* The desired results follows from Lemma 7.7 when

$$\|\tilde{v}_T\|_{L^\infty(T)} \geq Lc_2$$

and from Lemma 7.8 when

$$\|\tilde{v}_T\|_{L^\infty(T)} < Lc_2 \quad \text{and} \quad \min_{x \in T} \tilde{v}_T < \frac{c_1}{2}.$$



In the remaining case

$$\|\tilde{v}_T\|_{L^\infty(T)} < Lc_2 \quad \text{and} \quad \min_{x \in T} \tilde{v}_T(x) \geq \frac{c_1}{2},$$

since  $\widehat{v}_T = \tilde{v}_T$  satisfies the desired constraints there is nothing to prove.  $\square$

### 7.2.2 Constrained approximation of the diffusion coefficients

For matrix-valued functions, the constraints are on the eigenvalues of the matrix rather than on the coefficients themselves. Although this requires a few adjustments, the process is similar to the scalar case. We recall that for  $0 < \alpha_1 \leq \alpha_2$ ,  $M(\alpha_1, \alpha_2) \subset L^\infty(\Omega; \mathbb{R}_{\text{sym}}^{d \times d})$  denotes the class of symmetric matrix-valued functions whose eigenvalues lie between  $\alpha_1$  and  $\alpha_2$ ; see (5.48).

Algorithm CONSTRAINT-A is based on (7.12) and modifies approximations  $\tilde{\mathbf{A}} \in (\mathbb{S}_{\mathcal{T}}^{n_{\mathbf{A}}, -1})^{d \times d}$  of  $\mathbf{A} \in M(\alpha_1, \alpha_2)$  to produce uniformly positive definite approximations  $\widehat{\mathbf{A}} \in (\mathbb{S}_{\mathcal{T}}^{n_{\mathbf{A}}, -1})^{d \times d}$  of  $\mathbf{A}$ .

**Algorithm 7.10 (CONSTRAINT-A)** Given a threshold parameter  $L > 2$ ,  $0 < \alpha_1 \leq \alpha_2$ , a conforming refinement  $\mathcal{T} \in \mathbb{T}$  of  $\mathcal{T}_0$ , and  $\tilde{\mathbf{A}} \in (\mathbb{S}_{\mathcal{T}}^{n_{\mathbf{A}}, -1})^{d \times d}$ , this routine constructs a positive definite  $\widehat{\mathbf{A}} \in (\mathbb{S}_{\mathcal{T}}^{n_{\mathbf{A}}, -1})^{d \times d}$ .

$$[\widehat{\mathbf{A}}] = \text{CONSTRAINT-A}(\mathcal{T}, \alpha_1, \alpha_2, L, \tilde{\mathbf{A}})$$

For  $T \in \mathcal{T}$

$$\tilde{\alpha}_{1,T} = \inf\{y^t \tilde{\mathbf{A}}(x)y, x \in T, |y| = 1\}$$

$$\tilde{\alpha}_{2,T} = \sup\{|y^t \tilde{\mathbf{A}}(x)y|, x \in T, |y| = 1\}$$

$$\text{if } \tilde{\alpha}_{2,T} \geq L\alpha_2$$

$$\quad \widehat{\mathbf{A}}|_T = \alpha_2 \mathbf{I}_d$$

$$\text{else if } \tilde{\alpha}_{1,T} < \frac{\alpha_1}{2}$$

$$\quad \widehat{\mathbf{A}}|_T = \tilde{\mathbf{A}}|_T - \left(\frac{\alpha_1}{2} - \tilde{\alpha}_{1,T}\right) \mathbf{I}_d$$

else

$$\quad \widehat{\mathbf{A}}|_T = \tilde{\mathbf{A}}|_T$$

return  $\widehat{\mathbf{A}}$

Notice that CONSTRAINT-A preserves symmetry, i.e, if  $\tilde{\mathbf{A}}$  is symmetric so is the output  $\widehat{\mathbf{A}}$ . In addition, when  $n = 1$  and  $\tilde{\mathbf{A}} \in (\mathbb{S}_{\mathcal{T}}^{0, -1})^{d \times d}$  is the piecewise constant local average of  $\mathbf{A}$ , the output  $\widehat{\mathbf{A}}$  of CONSTRAINT-A is  $\widehat{\mathbf{A}} = \tilde{\mathbf{A}}$  since in that case the parameters  $\tilde{\alpha}_{1,T}$  and  $\tilde{\alpha}_{2,T}$  satisfy

$$\tilde{\alpha}_{1,T} \geq \alpha_1 > \frac{\alpha_1}{2} \quad \text{and} \quad \tilde{\alpha}_{2,T} \leq \alpha_2 < L\alpha_2, \quad \forall T \in \mathcal{T}.$$

This is consistent with the observation made in Section 5.4.2.

The next corollary hinges on Corollary 7.9 (locally enforcing constraints) to derive properties of CONSTRAINT-A. In passing, we recall that for  $\mathbf{A} \in L^p(\Omega; \mathbb{R}^{d \times d})$  we write

$$\|\mathbf{A}\|_{L^p(\Omega)} := \| \|\mathbf{A}\| \|_{L^p(\Omega)},$$

where for  $x \in \Omega$ ,  $|\mathbf{A}(x)|$  is the spectral norm of  $\mathbf{A}(x)$ .

**Corollary 7.11 (locally enforcing constraints for matrices)** *Let the threshold be  $L > 2$ ,  $0 < \alpha_1 \leq \alpha_2$ , and  $\mathbf{A} \in M(\alpha_1, \alpha_2)$ . Let  $\mathcal{T} \in \mathbb{T}$  be any conforming refinement of  $\mathcal{T}_0$  and  $\tilde{\mathbf{A}} \in (\mathbb{S}_{\mathcal{T}}^{n_{\mathbf{A}}, -1})^{d \times d}$  be a symmetric approximation of  $\mathbf{A}$ . Then the output  $[\hat{\mathbf{A}}] = \text{CONSTRAINT-A}(\mathcal{T}, \alpha_1, \alpha_2, L, \tilde{\mathbf{A}})$  is symmetric and satisfies*

$$\frac{\alpha_1}{2} \leq \lambda_j(\hat{\mathbf{A}}) \leq \left(\frac{1}{2} + 2L\right) \alpha_2 \quad \text{a.e. in } \Omega, \quad 1 \leq j \leq d.$$

Moreover for  $1 \leq q \leq \infty$ , we have

$$\|\mathbf{A} - \hat{\mathbf{A}}\|_{L^q(T)} \leq C_{\text{data}} \|\mathbf{A} - \tilde{\mathbf{A}}\|_{L^q(T)} \quad \forall T \in \mathcal{T},$$

where  $C_{\text{data}} := \max(C_1^+, C_2^+)$  and  $C_1^+$  and  $C_2^+$  are the constants appearing in Lemmas 7.8 and 7.7, which only depends on  $d$ ,  $n_{\mathbf{A}}$ ,  $\alpha_2/\alpha_1$ ,  $L$ , and the shape regularity of  $\mathbb{T}$ .

*Proof.* We observe that  $\tilde{\mathbf{A}}$  is not assumed positive semi-definite. We argue locally and fix  $T \in \mathcal{T}$ . Let  $\tilde{\alpha}_{2,T} > 0$  and  $y_0 \in \mathbb{R}^d$  be such that  $|y_0| = 1$  and

$$\tilde{\alpha}_{2,T} := \sup_{x \in T} |y_0^t \tilde{\mathbf{A}}(x) y_0| := \sup_{x \in T} \sup_{y \in \mathbb{R}^d, |y|=1} |y^t \tilde{\mathbf{A}}(x) y|.$$

We first consider the case  $\tilde{\alpha}_{2,T} \geq L\alpha_2$  for which  $\hat{\mathbf{A}}|_T := \alpha_2 \mathbf{I}_d$ . For  $x \in T$ , we set

$$a(x) := y_0^t \mathbf{A}(x) y_0 \quad \text{and} \quad \tilde{a}_T(x) = y_0^t \tilde{\mathbf{A}}(x) y_0 \in \mathbb{P}_{n_{\mathbf{A}}}(T).$$

These notations allows us to reduce to the scalar case upon noting that

$$\|a - \tilde{a}_T\|_{L^q(T)} \leq \|\mathbf{A} - \tilde{\mathbf{A}}\|_{L^q(T)}$$

and  $\alpha_1 \leq a \leq \alpha_2$  a.e. in  $\Omega$ . Because  $\tilde{\alpha}_{2,T} \geq L\alpha_2$ , Lemma 7.7 with  $m = n_{\mathbf{A}}$  guarantees that  $\hat{a}_T := \alpha_2$  satisfies

$$\|a - \hat{a}_T\|_{L^q(T)} \leq C_2^+ \|a - \tilde{a}_T\|_{L^q(T)} \leq C_2^+ \|\mathbf{A} - \tilde{\mathbf{A}}\|_{L^q(T)}.$$

Consequently, the matrix-valued approximation  $\hat{\mathbf{A}}|_T := \alpha_2 \mathbf{I}_d$  satisfies

$$\|\mathbf{A} - \hat{\mathbf{A}}\|_{L^q(T)} = \|a - \hat{a}_T\|_{L^q(T)} \leq C_2^+ \|\mathbf{A} - \tilde{\mathbf{A}}\|_{L^q(T)}.$$

This proves the desired result when  $\tilde{\alpha}_{2,T} \geq L\alpha_2$ .

We now consider the case where  $\tilde{\alpha}_{2,T} < L\alpha_2$  and define  $\tilde{\alpha}_{1,T} \in \mathbb{R}$ ,  $y_1 \in \mathbb{R}^d$  with  $|y_1| = 1$  by the relations

$$\tilde{\alpha}_{1,T} = \inf_{x \in T} y_1^t \tilde{\mathbf{A}}(x) y_1 = \inf_{x \in T} \inf_{|y|=1} y^t \tilde{\mathbf{A}}(x) y.$$

We also redefine the associated scalar functions for  $x \in T$  using  $y_1$  instead of  $y_0$

$$a(x) := y_1^t \mathbf{A}(x) y_1 \quad \text{and} \quad \tilde{a}_T(x) = y_1^t \tilde{\mathbf{A}}(x) y_1 \in \mathbb{P}_{n_{\mathbf{A}}}(T).$$

If  $\tilde{\alpha}_{1,T} < \frac{\alpha_1}{2}$  then  $\hat{\mathbf{A}}|_T = \tilde{\mathbf{A}}|_T + (\frac{\alpha_1}{2} - \tilde{\alpha}_{1,T}) \mathbf{I}_d$ . Lemma 7.8 with  $m = n_{\mathbf{A}}$  ensures

that  $\widehat{a}_T = \widetilde{a}_T + \frac{\alpha_1}{2} - \widetilde{\alpha}_{1,T}$  satisfies

$$\frac{\alpha_1}{2} \leq \widehat{a}_T \leq \left( \frac{1}{2} + 2M \right) \alpha_2$$

and

$$\|a - \widehat{a}_T\|_{L^q(T)} \leq C_1^+ \|a - \widetilde{a}_T\|_{L^q(T)} \leq C_1^+ \|A - \widetilde{A}\|_{L^q(T)}.$$

Thus,  $\widehat{A}|_T$  satisfies the desired properties provided  $\widetilde{\alpha}_{2,T} \geq L\alpha_2$  as well.

It the remaining case  $\widetilde{\alpha}_{2,T} < L\alpha_2$  and  $\widetilde{\alpha}_{1,T} \geq \frac{\alpha_1}{2}$ , the function  $\widehat{A}|_T = \widetilde{A}|_T$  satisfies the desired properties and there is nothing to prove.  $\square$

As a corollary, we report the complexity of an algorithm that concatenates the linear approximation of GREEDY with the nonlinear correction into the constraint of CONSTRAINT-A. We recall from Corollary 6.36 (approximation class of  $A$ ) that the admissible set of parameters of  $A$  for GREEDY are  $n_A \leq n - 1$

$$s_A \in (0, n_A], \quad p_A \in (0, \infty], \quad q_A \in [2, \infty], \quad s_A - \frac{d}{p_A} + \frac{d}{q_A} > 0, \quad t_A = 0.$$

**Corollary 7.12 (complexity of constrained GREEDY for  $A$ )** *Let the initial mesh  $\mathcal{T}_0$  of  $\Omega \subset \mathbb{R}^d$  satisfy Assumption 6.19 (initial labeling) and  $\mathcal{T} \in \mathbb{T}$  be any admissible refinement of  $\mathcal{T}_0$ . Let  $\tau > 0$  be the target tolerance,  $b \geq 1$  be the number of bisections performed on each marked element, and  $L > 2$  be a threshold parameter. Furthermore, assume that  $(A, s_A, t_A, p_A, q_A)$  satisfies Assumption 7.2 (admissible set of parameters for GREEDY) with local oscillations  $\{\|A - \widehat{A}\|_{L^q(T)}\}_{T \in \mathcal{T}}$  and, in addition,  $A \in M(\alpha_1, \alpha_2)$  for some  $0 < \alpha_1 \leq \alpha_2$ . The algorithm*

$$\begin{aligned} [\widehat{\mathcal{T}}, \widetilde{A}] &= \text{GREEDY}(\mathcal{T}, \tau, q_A, b, A) \\ [\widehat{A}] &= \text{CONSTRAINT-A}(\widehat{\mathcal{T}}, \alpha_1, \alpha_2, L, \widetilde{A}) \end{aligned}$$

where GREEDY is applied to the  $d(d+1)/2$  distinct components of  $A$ , marks  $N$  elements of  $\mathcal{T}$  for refinement with

$$N \leq C |A|_{X_{p_A}^{s_A}(\Omega; \mathcal{T}_0)}^{\frac{d}{s_A}} \tau^{-\frac{d}{s_A}} \quad (7.20)$$

and  $C = C(p_A, q_A, s_A, b, d, n_A, \alpha_2/\alpha_1, L, \Omega, \mathcal{T}_0)$ . Moreover,  $\widehat{A} \in (\mathbb{S}_{\mathcal{T}}^{n_A, -1})^{d \times d}$  satisfies

$$\widehat{A} \in M(\widehat{\alpha}_1, \widehat{\alpha}_2) : \quad \widehat{\alpha}_1 = \frac{\alpha_1}{2}, \quad \widehat{\alpha}_2 = (1 + 4L) \frac{\alpha_2}{2}, \quad (7.21)$$

and there is a constant  $C_{data} > 0$  such that

$$\|A - \widehat{A}\|_{L^q(\Omega)} \leq C_{data} \tau.$$

*Proof.* This result follows upon invoking Corollary 7.6 (performance of greedy) and Corollary 7.11 (locally enforcing constraints for matrices).  $\square$

**Remark 7.13 (constrained approximation class of matrices)** As a consequence of Corollary 7.12, we realize that for  $A \in M(\alpha_1, \alpha_2)$ ,

$$\delta_{\mathcal{T}}(A)_r \leq \tilde{\delta}_{\mathcal{T}}(A)_r \leq C_{\text{data}} \delta_{\mathcal{T}}(A)_r,$$

where the best approximation error  $\delta_{\mathcal{T}}(A)_r$  and best constrained approximation error  $\tilde{\delta}_{\mathcal{T}}(A)_r$  are defined in (6.9) and (6.10).

### 7.2.3 Constrained approximation of the reaction coefficients

If the reaction coefficient  $c \in R(c_1, c_2)$  is strictly positive ( $c_1 > 0$ ), then Corollary 7.9 (locally enforcing constraints) with  $m = n_c$  directly applies to  $v_T = c|_T$ ,  $T \in \mathcal{T}$ , and guarantees that the approximate coefficient  $\hat{c} \in \mathbb{S}_{\mathcal{T}}^{n_c, -1}$  defined on  $T \in \mathcal{T}$  by  $\hat{c}|_T := \hat{v}_T$  satisfies

$$\hat{c} \in R(\hat{c}_1, \hat{c}_2) : \quad \hat{c}_1 = \frac{c_1}{2}, \quad \hat{c}_2 = (1 + 4L) \frac{c_2}{2}.$$

However, reaction coefficients are not necessarily strictly positive on  $\bar{\Omega}$  and Corollary 7.9 cannot be invoked directly. Instead, we take advantage of the fact that the perturbed problem (5.5) is still well-posed provided  $\hat{c} \geq -\hat{\alpha}_1/(2C_P^2)$  and the approximate diffusion coefficient  $\hat{A} \in M(\hat{\alpha}_1, \hat{\alpha}_2)$  of  $A \in M(\alpha_1, \alpha_2)$  satisfies  $\hat{\alpha}_1 \geq \alpha_1/2$  according with (5.52); hence  $\hat{c} \geq -\alpha_1/(4C_P^2)$ . Therefore, we apply Corollary 7.9 to the shifted reaction coefficient  $v = c + \frac{\hat{\alpha}_1}{C_P^2}$ , which satisfies

$$v_1 := c_1 + \frac{\hat{\alpha}_1}{C_P^2} \leq v \leq c_2 + \frac{\hat{\alpha}_1}{C_P^2} =: v_2. \quad (7.22)$$

Below is the proposed algorithm for the construction of  $\hat{c}$  in the general case  $c_1 \geq 0$ .

**Algorithm 7.14 (CONSTRAINT-c)** Given  $L > 2$ ,  $\hat{\alpha}_1 > 0$ , a conforming refinement  $\mathcal{T} \in \mathbb{T}$  of  $\mathcal{T}_0$ , and  $\tilde{c} \in \mathbb{S}_{\mathcal{T}}^{n_c, -1}$ , this routine constructs  $\hat{c} \in \mathbb{S}_{\mathcal{T}}^{n_c, -1}$  as follows:

```

[ $\hat{c}$ ] = CONSTRAINT-c( $\mathcal{T}, \hat{\alpha}_1, L, \tilde{c}$ )
 $\tilde{v} = \tilde{c} + \frac{\hat{\alpha}_1}{C_P^2}$ 
For  $T \in \mathcal{T}$ 
  if  $\|\tilde{v}\|_{L^\infty(T)} \geq Lv_2$ 
     $\hat{v}|_T = v_2$ 
  else if  $\min_{x \in T} \tilde{v}(x) < \frac{v_1}{2}$ 
     $\hat{v}|_T = \tilde{v}|_T - \min_{x \in T} \tilde{v}(x) + \frac{v_1}{2}$ 
  else
     $\hat{v}|_T = \tilde{v}|_T$ 
 $\hat{c} = \hat{v} - \frac{\hat{\alpha}_1}{C_P^2}$ 
return  $\hat{c}$ 

```

We note that if  $n_c = 0$ , then  $\tilde{c}$  is the piecewise average of  $c$  and CONSTRAINT-c does not modify  $\tilde{c}$  which already satisfies the structural assumption (7.11).

The next result shows that the output  $\widehat{c}$  of CONSTRAINT-c is a modification of  $\widetilde{c}$  which satisfies  $\widehat{c} \in R(\widehat{c}_1, \widehat{c}_2)$ , with

$$\widehat{c}_1 := \frac{c_1}{2} - \frac{\widehat{\alpha}_1}{2C_P^2} \quad \text{and} \quad \widehat{c}_2 := (1 + 4L)\frac{c_2}{2} + (4L - 1)\frac{\widehat{\alpha}_1}{2C_P^2} \quad (7.23)$$

without affecting the approximation of  $c$  in  $L^q$ ,  $1 \leq q \leq \infty$  (up to a multiplicative constant). In particular  $\widehat{c}_1 \geq -\frac{\widehat{\alpha}_1}{2C_P^2}$ , which is necessary for the well-posedness of the perturbed problem (5.5) when  $\widehat{A} \in M(\widehat{\alpha}_1, \widehat{\alpha}_2)$ .

**Corollary 7.15 (locally enforcing constraints for nonnegative scalar functions)**

Let  $A \in M(\widehat{\alpha}_1, \widehat{\alpha}_2)$  with  $0 < \widehat{\alpha}_1 \leq \widehat{\alpha}_2$ , and  $c \in R(c_1, c_2)$  with  $0 \leq c_1 \leq c_2$ . Let  $L > 2$  and  $v_1 \leq v_2$  be defined in (7.22). Let  $\mathcal{T} \in \mathbb{T}$  be any conforming refinement of  $\mathcal{T}_0$  and  $\widetilde{c} \in \mathbb{S}_{\mathcal{T}}^{n_c, -1}$ . Then the output  $[\widehat{c}] = \text{CONSTRAINT-c}(\mathcal{T}, \widehat{\alpha}_1, L, \widetilde{c})$  satisfies

$$\widehat{c}_1 \leq \widehat{c} \leq \widehat{c}_2 \quad \text{a.e. in } \Omega,$$

where  $\widehat{c}_1$  and  $\widehat{c}_2$  are given by (7.23). Moreover for  $0 < q \leq \infty$ , we have

$$\|c - \widehat{c}\|_{L^q(T)} \leq C_{\text{data}} \|c - \widetilde{c}\|_{L^q(T)} \quad \forall T \in \mathcal{T},$$

where  $C_{\text{data}}$  is a constant only depending on  $d, n, v_2/v_1, \Omega, L$ , and the shape regularity of  $\mathbb{T}$ .

*Proof.* Set  $\kappa := \frac{\widehat{\alpha}_1}{C_P^2}$  and  $v := c + \kappa \in R(c_1 + \kappa, c_2 + \kappa)$  so that  $c_1 + \kappa > 0$ . On each  $T \in \mathcal{T}$ , we invoke Corollary 7.9 (locally enforcing constraints) with  $m = n_c$ ,  $\widetilde{v}_T = \widetilde{c}|_T + \kappa$  and where  $c_1, c_2$  are replaced by  $c_1 + \kappa, c_2 + \kappa$  respectively. Hence, we deduce that the function  $\widehat{v}$  constructed within CONSTRAINT-c satisfies

$$\frac{c_1 + \kappa}{2} \leq \widehat{v} \leq (1 + 4L)\frac{c_2 + \kappa}{2}.$$

and

$$\|v - \widehat{v}\|_{L^q(T)} \leq C_{\text{data}} \|v - \widetilde{v}\|_{L^q(T)} \quad \forall T \in \mathcal{T}, \quad (7.24)$$

with a constant  $C_{\text{data}}$  depending on  $d, n, v_2/v_1, L$ , and the shape regularity of  $\mathbb{T}$ . Shifting back,  $c = v - \kappa$  and  $\widehat{c} := \widehat{v} - \kappa$ , we find that the approximation  $\widehat{c}$  constructed by CONSTRAINT-c satisfies

$$\frac{c_1 + \kappa}{2} - \kappa \leq \widehat{c} \leq (1 + 4L)\frac{c_2 + \kappa}{2} - \kappa$$

or equivalently

$$\frac{c_1}{2} - \frac{\kappa}{2} \leq \widehat{c} \leq (1 + 4L)\frac{c_2}{2} + (4L - 1)\frac{\kappa}{2}.$$

In view of (7.23) and  $\kappa = \frac{\widehat{\alpha}_1}{C_P^2}$ , this is the first desired inequality in disguised.

Furthermore, the second desired inequality follows from (7.24) because for  $T \in \mathcal{T}$  we have  $c - \widehat{c} = v - \widehat{v}$  and  $c - \widetilde{c} = v - \widetilde{v}$ .  $\square$

The next corollary combines the linear approximation of GREEDY together with the nonlinear correction into the constraint of CONSTRAINT-c. We recall from Corollary 6.37 (approximation class of  $c$ ) that the admissible set of parameters of  $c$  for GREEDY are  $n_c \leq n-1, s_c \in (0, n_c], p_c \in (0, \infty]$

$$\begin{aligned} n_c > 0 &\Rightarrow q_c > \frac{d}{2}, \quad s_c - \frac{d}{p_c} + \frac{d}{q_c} > 0, \quad t_c = 0; \\ n_c = 0 &\Rightarrow q_c = 2, \quad s_c - \frac{d}{p_c} + \frac{d}{2} > 0, \quad 0 < t_c < 2 - \frac{d}{2}. \end{aligned}$$

**Corollary 7.16 (complexity of constrained GREEDY for  $c$ )** *Let the initial subdivision  $\mathcal{T}_0$  of  $\Omega \subset \mathbb{R}^d$  satisfy Assumption 6.19 (initial labeling) and  $\mathcal{T} \in \mathbb{T}$  be any admissible refinement of  $\mathcal{T}_0$ . Let  $\tau > 0$  be the target tolerance,  $b \geq 1$  be the number of bisections performed on each marked element,  $L > 2$  be the threshold parameter, and  $\hat{\alpha}_1 > 0$ . Furthermore, assume that  $(c, s_c, t_c, p_c, q_c)$  satisfies Assumption 7.2 (admissible set of parameters for GREEDY) with local oscillations  $\{\|c - \tilde{c}\|_{L^{q_c}(T)}\}_{T \in \mathcal{T}}$  and that  $c \in R(c_1, c_2)$  for some  $0 \leq c_1 \leq c_2$ . The algorithm*

$$\begin{aligned} [\hat{\mathcal{T}}, \tilde{c}] &= \text{GREEDY}(\mathcal{T}, \tau, q_c, b, c) \\ [\hat{c}] &= \text{CONSTRAINT-c}(\hat{\mathcal{T}}, \hat{\alpha}_1, L, \tilde{c}) \end{aligned}$$

marks  $N$  elements of  $\mathcal{T}$  for refinement with

$$N \leq C |c|_{X_{p_c}^{s_c}(\Omega; \mathcal{T}_0)}^{\frac{d}{s_c + t_c}} \tau^{-\frac{d}{s_c + t_c}} \quad (7.25)$$

and a constant  $C = C(p_c, q_c, s_c, b, d, n_c, v_2/v_1, L, \Omega, \mathcal{T}_0)$  with  $v_1 \leq v_2$  defined in (7.22) to construct  $\hat{\mathcal{T}}$ . The function  $\hat{c} \in \mathbb{S}_{\hat{\mathcal{T}}}^{n_c, -1}$  is a piecewise polynomial of degree  $\leq n_c$  over  $\hat{\mathcal{T}}$  and satisfies

$$\hat{c} \in R(\hat{c}_1, \hat{c}_2),$$

where  $\hat{c}_1 \leq \hat{c}_2$  are given by (7.23). Moreover, for  $1 < q_c \leq \infty$ , there is a constant  $C_{\text{data}}$  only depending on  $d, n_c, v_2/v_1, \Omega, L$ , and the shape regularity of  $\mathbb{T}$  such that

$$\|c - \hat{c}\|_{L^{q_c}(\Omega)} \leq C_{\text{data}} \tau.$$

*Proof.* Simply apply Corollary 7.6 (performance of greedy) and Corollary 7.15 (locally enforcing constraints for nonnegative scalar functions).  $\square$

**Remark 7.17 (constrained approximation class of scalars)** Corollary 7.16 implies that for  $c \in R(c_1, c_2)$ ,

$$\delta_{\mathcal{T}}(c)_q \leq \tilde{\delta}_{\mathcal{T}}(c)_q \leq C_{\text{data}} \delta_{\mathcal{T}}(c)_q,$$

where the best approximation error  $\delta_{\mathcal{T}}(c)_q$  and best constrained approximation error  $\tilde{\delta}_{\mathcal{T}}(c)_q$  are defined in (6.9) and (6.10).

### 7.3 Approximation of the load term $f$

We now turn our attention to the question of designing a practical algorithm for reducing the global oscillation

$$E_{\mathcal{T}}(f)_{-1}^2 := \sum_{T \in \mathcal{T}} \|f - P_{\mathcal{T}}f\|_{H^{-1}(\omega_T)}^2 \approx \sum_{z \in \mathcal{V}} \|f - P_{\mathcal{T}}f\|_{H^{-1}(\omega_z)}^2, \quad (7.26)$$

where the projection  $P_{\mathcal{T}}$  is defined in (4.34). The approximation of functionals in  $H^{-1}(\Omega)$  is rather intricate and out of reach without assuming additional structure enabling practical evaluation of their actions on polynomial functions.

We examine three cases of independent interest. In Section 7.3.1 we consider  $f \in L^q(\Omega)$  for  $q$  satisfying  $2d/(d+2) < q \leq \infty$ , which includes the most common setting  $f \in L^2(\Omega)$ . Sections 7.3.2 and 7.3.3 present examples of right-hand sides not in  $L^1$ . In Section 7.3.2 we treat the case  $f = g\delta_{\Gamma}$ , where  $\Gamma$  is an hyper-surface not necessarily captured by the faces of the subdivisions and  $g \in L^q(\Gamma)$ ,  $q \geq 2$ , while in Section 7.3.3 we consider  $f = \operatorname{div} \mathbf{g}$  for some  $\mathbf{g} \in L^2(\Omega; \mathbb{R}^d)$ . In all cases, the total error  $E_{\mathcal{T}}(f)_{-1}$  is estimated by a surrogate  $\tilde{E}_{\mathcal{T}}(f)_{-1}$ , namely  $E_{\mathcal{T}}(f)_{-1} \leq C_{\text{data}} \tilde{E}_{\mathcal{T}}(f)_{-1}$

$$\tilde{E}_{\mathcal{T}}(f)_{-1}^2 := \sum_{T \in \mathcal{T}} \widetilde{\text{osc}}_{\mathcal{T}}(f, T)_q^2$$

with a definition of  $\widetilde{\text{osc}}_{\mathcal{T}}(f, T)_q$  depending on the situation but local to  $T \in \mathcal{T}$  (and not on stars). This allows Algorithm 7.1 (GREEDY) to reduce  $\tilde{E}_{\mathcal{T}}(f)_{-1}$ .

Before starting, we recall relevant definitions and results from Section 4 (a posteriori error analysis). For  $z \in \mathcal{V}$ , we denote by  $\mathcal{T}_z \subset \mathcal{T}$  all the elements in  $\omega_z$  and  $\mathcal{F}_z \subset \mathcal{F}$  all the faces in  $\omega_z$ . For  $\ell \in H^{-1}(\Omega)$ , the restriction  $P_{\mathcal{T}}\ell|_{\omega_z}$  belongs to the space  $\mathbb{F}(\mathcal{T}_z) = \mathbb{F}_{m_1, m_2}(\mathcal{T}_z)$  made of functional whose action against  $w \in H_0^1(\omega_z)$  reads

$$\langle \ell, w \rangle = \sum_{T \in \mathcal{T}_z} \int_T q_T w + \sum_{F \in \mathcal{F}_z} \int_F q_F w \quad (7.27)$$

for some  $q_F \in P_{m_1}(F)$ ,  $F \in \mathcal{F}_z$  and  $q_T \in P_{m_2}(T)$ ,  $T \in \mathcal{T}_z$ . The polynomial degrees are chosen to be  $m_1 = n - 1$  and  $m_2 = n - 2$  but can be general in this discussion.

Corollary 4.31 (local near-best approximation) guarantees  $P_{\mathcal{T}}\ell|_{\omega_z}$  is the quasi-best discrete functional in  $\mathbb{F}(\mathcal{T}_z)$ , namely

$$\|\ell - P_{\mathcal{T}}\ell\|_{H^{-1}(\omega_z)} \leq C_P \inf_{\chi \in \mathbb{F}(\mathcal{T}_z)} \|\ell - \chi\|_{H^{-1}(\omega_z)}. \quad (7.28)$$

This will be used repeatedly to replace  $P_{\mathcal{T}}\ell$  by more tractable quantities and justify the use of GREEDY algorithms to reduce (7.26).

#### 7.3.1 The case $f \in L^q(\Omega)$

In this section, we show how to reduce the oscillation error (7.26) when  $f \in L^q(\Omega)$ , with  $q > \frac{2d}{d+2}$  to guarantee that  $L^q(\Omega)$  compactly embeds in  $H^{-1}(\Omega)$ . Note that this not only includes the most treated case in the literature  $f \in L^2(\Omega)$  but also the more intricate cases  $q < 2$  originally analyzed in [Cohen et al. 2012].

If  $\Pi_{\mathcal{T}}f$  is the  $L^2$ -projection of  $f$  into the space  $\mathbb{S}_{\mathcal{T}}^{n_f, -1}$  of discontinuous piecewise polynomials of degree  $n_f$ , let  $\widehat{f} \in \mathbb{S}_{\mathcal{T}}^{n_f, -1}$  be defined by (5.70);  $n_f = n - 1$  in some applications but not always. Since  $\widehat{f}|_{\omega_z} \in \mathbb{F}(\mathcal{T}_z)$  by taking  $q_F = 0$  and  $q_T = \widehat{f}|_T$  in (7.27), the local near-best approximation property (7.28) of  $P_{\mathcal{T}}$  implies

$$\|f - P_{\mathcal{T}}f\|_{H^{-1}(\omega_z)} \leq C_P \|f - \widehat{f}\|_{H^{-1}(\omega_z)}.$$

Furthermore, for  $v \in H_0^1(\omega_z)$  one has

$$\langle f - \widehat{f}, v \rangle \leq \|f - \widehat{f}\|_{L^q(\omega_z)} \|v\|_{L^{\tilde{q}}(\omega_z)}$$

where  $\frac{1}{q} + \frac{1}{\tilde{q}} = 1$ . Note that the restriction  $q > \frac{2d}{d+2}$  guarantees that  $1 \leq \tilde{q} < \frac{2d}{d-2}$  and thus  $\text{sob}(H^1) > \text{sob}(L^{\tilde{q}})$ . Therefore, Lemma 2.2 (first Poincaré inequality) yields

$$\|f - \widehat{f}\|_{H^{-1}(\omega_z)} \lesssim \text{diam}(\omega_z)^{1+d(\frac{1}{2}-\frac{1}{q})} \|f - \widehat{f}\|_{L^q(\omega_z)}.$$

Returning to (7.26), after rearranging the terms element-wise and invoking the shape-regularity of  $\mathbb{T}$ , we obtain  $E_{\mathcal{T}}(f)_{-1} \leq C_{\text{data}} \widetilde{E}_{\mathcal{T}}(f)_{-1}$ , where

$$\widetilde{E}_{\mathcal{T}}(f)_{-1}^2 := \sum_{T \in \mathcal{T}} \widetilde{\text{osc}}_{\mathcal{T}}(f, T)_q^2, \quad (7.29)$$

and  $\widetilde{\text{osc}}_{\mathcal{T}}(f, T)_q := h_T^t \|f - \widehat{f}\|_{L^q(T)}$  with  $t := 1 + d(\frac{1}{2} - \frac{1}{q}) > 0$ .

In view of the definition (5.70), the local oscillations  $\widetilde{\text{osc}}(f, T)_q$  satisfy Assumption 7.3 (monotonicity of local oscillations) in  $\ell^2$  and we can now employ Algorithm 7.1 (GREEDY) with local errors  $\widetilde{\text{osc}}_{\mathcal{T}}(f, T)_q$  accumulating in  $\ell^2$ . Recall that we use the convention  $X_q^0(\Omega; \mathcal{T}_0) = L^q(\Omega)$ .

**Corollary 7.18 (approximation class of  $f \in L^q(\Omega)$ )** *Let the initial subdivision  $\mathcal{T}_0$  of  $\Omega \subset \mathbb{R}^d$  satisfy Assumption 6.19 (initial labeling) and  $\mathcal{T} \in \mathbb{T}$  be any admissible refinement of  $\mathcal{T}_0$ . Let  $\tau > 0$  be the target tolerance and  $b \geq 1$  be the number of bisections performed on each marked element. Let  $2d/(d+2) < q \leq \infty$  and set  $t = 1 + d(\frac{1}{2} - \frac{1}{q})$ . Let  $(f, s, t, p, 2)$  satisfy Assumption 7.2 (admissible set of parameters for GREEDY) with local oscillations  $\{\widetilde{\text{osc}}(f, T)_q\}_{T \in \mathcal{T}}$ . Then  $[\widehat{\mathcal{T}}, \widehat{f}] = \text{GREEDY}(\mathcal{T}, \tau, 2, b, f)$  terminates in a finite number of steps with  $E_{\widehat{\mathcal{T}}}(f)_{-1} \leq \tau$ , whence*

$$E_{\widehat{\mathcal{T}}}(f)_{-1} \leq C_{\text{data}} \tau.$$

Moreover, the number  $N$  of marked elements by GREEDY satisfies

$$N \lesssim |f|_{X_p^s(\Omega; \mathcal{T}_0)}^{\frac{d}{s+t}} \tau^{-\frac{d}{s+t}}. \quad (7.30)$$

In particular,  $f \in \mathbb{F}_{\frac{d}{s+t}}$  with  $|f|_{\mathbb{F}_{\frac{d}{s+t}}} \lesssim \|f\|_{X_p^s(\Omega; \mathcal{T}_0)}$ .

*Proof.* Directly apply Corollary 7.6 (performance of GREEDY). □



### 7.3.2 The case $f = g\delta_C$

We now consider the case where the right-hand side data  $f$  is a density supported on a Lipschitz hyper-surface  $C \subset \Omega$  in  $\mathbb{R}^d$  with  $(d-1)$ -measure  $|C| < \infty$ .

The intricate interactions between bulk and interface contributions on  $P_{\mathcal{T}}$  makes it difficult to analyze when  $f = g\delta_C$  with density  $g \in L^q(C)$ . We take a simpler approach, likely suboptimal when  $n > 1$  and  $d > 2$ , which discards  $P_{\mathcal{T}}$  in view of the near-best approximation property (7.28)

$$\|f - P_{\mathcal{T}}f\|_{H^{-1}(\omega_z)} \lesssim \|f\|_{H^{-1}(\omega_z)}. \quad (7.31)$$

The right-hand side of the above estimate is the starting point of the analysis in [Cohen *et al.* 2012] assuming  $n = 1$  and  $d = 2$ .

We start with the derivation of a first upper bound for the local error  $\|f\|_{H^{-1}(\omega_z)}$ .

**Lemma 7.19 (local oscillation)** *Let  $\mathcal{T} \in \mathbb{T}$ ,  $z \in \mathcal{N}$ , and  $q > \frac{2(d-1)}{d}$ . If  $g \in L^q(C)$  and  $t := \frac{d}{2} - \frac{1}{q}(d-1) > 0$ , then there holds*

$$\|f\|_{H^{-1}(\omega_z)} \lesssim |\omega_z \cap C|^{\frac{t}{d-1}} \|g\|_{L^q(\omega_z \cap C)} \lesssim \sum_{T \subset \omega_z} h_T^t \|g\|_{L^q(T \cap C)}. \quad (7.32)$$

*Proof.* For  $v \in H_0^1(\omega_z)$  and  $\frac{1}{q} + \frac{1}{\tilde{q}} = 1$ , we have

$$\langle f, v \rangle = \int_{\omega_z \cap C} gv \leq \|g\|_{L^q(\omega_z \cap C)} \|v\|_{L^{\tilde{q}}(\omega_z \cap C)}. \quad (7.33)$$

We realize that  $H^{1/2}(\omega_z \cap C)$  compactly embeds in  $L^{\tilde{q}}(\omega_z \cap C)$  because

$$\begin{aligned} t := \text{sob}(H^{1/2}(\omega_z \cap C)) - \text{sob}(L^{\tilde{q}}(\omega_z \cap C)) &= \frac{1}{2} - (d-1) \left( \frac{1}{2} - \frac{1}{\tilde{q}} \right) \\ &= \frac{d}{2} - \frac{1}{q}(d-1) > 0, \end{aligned}$$

provided  $q > \frac{2(d-1)}{d}$ . Consequently, we find that

$$\|v\|_{L^{\tilde{q}}(\omega_z \cap C)} \lesssim |\omega_z \cap C|^{\frac{t}{d-1}} \|v\|_{H^{1/2}(\omega_z \cap C)}.$$

It remains to invoke the continuity (2.4) of the trace operator to write

$$\|v\|_{L^{\tilde{q}}(\omega_z \cap C)} \lesssim |\omega_z \cap C|^{\frac{t}{d-1}} \|v\|_{H^1(\omega_z)},$$

which, together with (7.33), yields the first estimate in (7.32). To deduce the second estimate, it suffices to note that  $|\omega_z \cap C| \lesssim \text{diam}(\omega_z)^{d-1} \lesssim h_T^{d-1}$  for  $T \subset \omega_z$  and that  $\|g\|_{L^q(\omega_z \cap C)} \leq \sum_{T \subset \omega_z} \|g\|_{L^q(T \cap C)}$ .  $\square$

Estimate (7.31) and Lemma 7.19 provide a surrogate for data oscillation

$$\tilde{E}_{\mathcal{T}}(f)_{-1}^2 := \sum_{T \in \mathcal{T}} \widetilde{\text{osc}}_{\mathcal{T}}(g, T)_q^2, \quad \widetilde{\text{osc}}_{\mathcal{T}}(g, T)_q := h_T^t \|g\|_{L^q(T \cap C)}, \quad (7.34)$$

where  $t = \frac{d}{2} - \frac{1}{q}(d-1)$ . The quantity  $\widetilde{\text{osc}}_{\mathcal{T}}(g, T)_q$  verifies Assumption 7.3 (monotonicity of local oscillations) with  $\Omega$  replaced by  $C$  because of its element-wise structure. Therefore, Proposition 7.4 (performance of GREEDY) states that Algorithm 7.1 (GREEDY) can reduce  $\widetilde{E}_{\mathcal{T}}(f)_{-1}$ . This is in contrast with the star-wise GREEDY algorithm analyzed in [Cohen et al. 2012], which requires that all marked stars are refined  $d$  times to ensure all the faces in the marked stars are refined.

We now discuss the performance of GREEDY with local indicators  $\widetilde{\text{osc}}_{\mathcal{T}}(g, T)_q$ .

**Lemma 7.20 (approximation class of  $f = g\delta_C$ )** *Let  $C \subset \Omega$  be Lipschitz hypersurface. Let the initial subdivision  $\mathcal{T}_0$  of  $\Omega \subset \mathbb{R}^d$  satisfy Assumption 6.19 (initial labeling) and  $\mathcal{T} \in \mathbb{T}$  be any admissible refinement of  $\mathcal{T}_0$ . Let  $\tau > 0$  be the target tolerance and  $b \geq 1$  be the number of bisections performed on each marked element, and  $2(d-1)/d < q \leq \infty$ . Then  $[\widehat{\mathcal{T}}, \widehat{f}] = \text{GREEDY}(\mathcal{T}, \tau, 2, b, f)$  terminates in a finite number of steps with surrogate estimator  $\widetilde{E}_{\widehat{\mathcal{T}}}(f)_{-1} \leq \tau$  defined in (7.34), whence*

$$E_{\widehat{\mathcal{T}}}(f)_{-1} \leq C_{\text{data}}\tau.$$

Moreover, the number  $N$  of marked elements by GREEDY satisfies

$$N \lesssim \|g\|_{L^q(C)}^{2(d-1)} \tau^{-2(d-1)}. \quad (7.35)$$

In particular,  $f = g\delta_C \in \mathbb{F}_{\frac{1}{2(d-1)}}$  with  $|f|_{\mathbb{F}_{\frac{1}{2(d-1)}}} \lesssim \|g\|_{L^q(C)}$

*Proof.* This proof mainly follows the proof of Proposition 7.4 (performance of GREEDY) but requires a few modifications to account for the geometry of the problem. Since in turn the proof of Proposition 7.4 describes modifications to the proof of Proposition 3.19 (abstract greedy), we now provide a complete proof. We proceed in several steps. We first consider the call  $\text{GREEDY}(\mathcal{T}_0, \tau, 2, 1, f)$  from  $\mathcal{T}_0$  with one bisection  $b = 1$  and accumulation in  $\ell^2$ , and discuss the general call from  $\mathcal{T}$  with  $b \geq 1$  in the last step of this proof.

**[1] Termination.** Since  $h_{\mathcal{T}}$  decreases monotonically to 0 with bisection, so does  $\widetilde{\text{osc}}_{\mathcal{T}}(g, T)_q$ . Consequently, GREEDY terminates in finite number  $k \geq 1$  of iterations. Let  $T_1, \dots, T_k$  be the sequence of marked elements, with  $\mathcal{M} = \{T_1, \dots, T_k\}$  and  $\mathcal{T}_1, \dots, \mathcal{T}_k$  be the sequence of refinements produced by GREEDY starting from  $\mathcal{T}_0$ . Upon termination, the surrogate error satisfies  $\widetilde{E}_{\mathcal{T}_k}(f)_{-1} \leq \tau$ , whence  $E_{\mathcal{T}_k}(f)_{-1} \leq C_{\text{data}}\tau$ .

**[2] Counting.** To estimate the cardinality of  $\mathcal{T}_k$ , we need to count  $\#\mathcal{M}$ . Set

$$\delta_i := \widetilde{\text{osc}}_{\mathcal{T}_i}(g, T_i)_q, \quad 1 \leq i \leq k, \quad \text{and} \quad \delta := \delta_{k-1}.$$

Then, there holds

$$\widetilde{E}_{\mathcal{T}_k}(f)_{-1} \leq \tau < \widetilde{E}_{\mathcal{T}_{k-1}}(f)_{-1} \leq \delta (\#\mathcal{T}_{k-1})^{\frac{1}{2}} \leq \delta (\#\mathcal{T}_k)^{\frac{1}{2}}. \quad (7.36)$$

We organize the elements in  $\mathcal{M}$  by size in such a way that allows for a counting

argument. Let  $\mathcal{P}_j$  be the set of elements  $T$  of  $\mathcal{M}$  with size

$$2^{-(j+1)} \leq |T| < 2^{-j} \quad \Rightarrow \quad 2^{-\frac{j+1}{d}} \leq h_T < 2^{-\frac{j}{d}}.$$

We first observe that all  $T$ 's in  $\mathcal{P}_j$  are *disjoint*. This is because if  $T_1, T_2 \in \mathcal{P}_j$  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 which works in any dimension  $d \geq 1$ ; see Section 3.5. Hence,

$$|T_1| \leq \frac{1}{2} |T_2|$$

contradicting the definition of  $\mathcal{P}_j$ . On the one hand, this implies the first bound

$$2^{-\frac{(j+1)(d-1)}{d}} \#\mathcal{P}_j \lesssim |C| \quad \Rightarrow \quad \#\mathcal{P}_j \lesssim |C| 2^{\frac{(j+1)(d-1)}{d}}, \quad (7.37)$$

where we used that  $h_T^{d-1} \approx |\omega_T \cap C|$  since  $T \cap C \neq \emptyset$  for all marked elements. Recall that  $\omega_T$  stands for the patch of elements around  $T$ .

On the other hand, the monotonicity of the local error indicators  $\widetilde{\text{osc}}_{\mathcal{T}_i}(g, T)_q = h_T^t \|g\|_{L^q(T \cap C)}$ , implies that REFINES does not increase  $\widetilde{\text{osc}}_{\mathcal{T}_i}(g, T)_q$  and thus

$$\delta \leq \delta_i = \widetilde{\text{osc}}_{\mathcal{T}_i}(g, T_i)_q, \quad 1 \leq i \leq k-1,$$

where  $t = \frac{d}{2} - \frac{1}{q}(d-1)$ . In view of (7.34), if  $T_i \in \mathcal{P}_j$ , then we obtain

$$\delta \leq \widetilde{\text{osc}}_{\mathcal{T}_i}(g, T_i)_q \lesssim 2^{-\frac{jt}{d}} \|g\|_{L^q(T_i \cap C)}.$$

Therefore, accumulating these quantities in  $\ell^q$  yields

$$\delta^q \#\mathcal{P}_j \lesssim 2^{-\frac{jtq}{d}} \|g\|_{L^q(C)}^q$$

and gives rise to the second bound

$$\#\mathcal{P}_j \lesssim \delta^{-q} 2^{-\frac{jtq}{d}} \|g\|_{L^q(C)}^q. \quad (7.38)$$

3 *Cardinality.* The two bounds for  $\#\mathcal{P}$  in (7.37) and (7.38) are complementary. The first one 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

$$2^{\frac{(j_0+1)(d-1)}{d}} |C| \approx \delta^{-q} 2^{-\frac{j_0 tq}{d}} \|g\|_{L^q(C)}^q \quad \Rightarrow \quad 2^{j_0} \approx |C|^{-\frac{2}{q}} \delta^{-2} \|g\|_{L^q(C)}^2,$$

upon using the expression for  $t$ . We now compute

$$k = \#\mathcal{M} = \sum_j \#\mathcal{P}_j \lesssim \sum_{j \leq j_0} 2^{\frac{j(d-1)}{d}} |C| + \delta^{-q} \|g\|_{L^q(C)}^q \sum_{j > j_0} 2^{-\frac{tq}{d}j}.$$

Since

$$\sum_{j \leq j_0} 2^{\frac{j(d-1)}{d}} \approx 2^{\frac{j_0(d-1)}{d}}, \quad \sum_{j > j_0} (2^{-\frac{tq}{d}})^j \lesssim 2^{-\frac{tq j_0}{d}},$$

we can write

$$\#\mathcal{M} \lesssim |C|^{1-\frac{2(d-1)}{qd}} (\delta^{-1} \|g\|_{L^q(C)})^{\frac{2(d-1)}{d}}.$$

We finally apply Theorem 3.16 (complexity of REFINE) to arrive at

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \lesssim \#M \lesssim |C|^{1-\frac{2(d-1)}{qd}} (\delta^{-1} \|g\|_{L^q(C)})^{\frac{2(d-1)}{d}},$$

or equivalently

$$\delta \lesssim |C|^{\frac{d}{d-1}-\frac{2}{q}} \|g\|_{L^q(C)} (\#\mathcal{T} - \#\mathcal{T}_0)^{-\frac{d}{2(d-1)}}.$$

We deduce from (7.36) that

$$\tau \lesssim \delta(\#\mathcal{T})^{\frac{1}{2}} \lesssim |C|^{\frac{d}{d-1}-\frac{2}{q}} \|g\|_{L^q(C)} (\#\mathcal{T} - \#\mathcal{T}_0)^{-\frac{d}{2(d-1)}+\frac{1}{2}}$$

or equivalently

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \lesssim \|g\|_{L^q(C)}^{2(d-1)} \tau^{-2(d-1)}, \quad (7.39)$$

From this we conclude that  $f = g\delta_C \in \mathbb{F}_{\frac{1}{2(d-1)}}$  with  $|f|_{\mathbb{F}_{\frac{1}{2(d-1)}}} \lesssim \|g\|_{L^q(C)}$  as desired.

<sup>[4]</sup> *Starting from  $\mathcal{T}$ .* To derive similar properties for GREEDY starting from  $\mathcal{T} \in \mathbb{T}$ , we proceed as in the proof of Corollary 7.6 (performance of greedy). We distinguish the output  $[\tilde{\mathcal{T}}, f] = \text{GREEDY}(\mathcal{T}_0, \tau, 2, 1, f)$  starting from  $\mathcal{T}_0$  and performing  $b = 1$  bisection per marked element with  $[\hat{\mathcal{T}}, \hat{f}] = \text{GREEDY}(\mathcal{T}, \tau, 2, b, f)$  starting from  $\mathcal{T} \in \mathbb{T}$  and performing  $b \geq 1$  bisections per marked element. Lemma 7.5 (GREEDY starting from  $\mathcal{T}$ ) guarantees that  $\text{GREEDY}(\mathcal{T}, \tau, 2, b, f)$  terminates with  $\tilde{E}_{\hat{\mathcal{T}}}(f)_{-1} \leq \tau$ . Moreover, Lemma 7.5 also ensures that the number of marked elements satisfies

$$N \leq \#\tilde{\mathcal{T}} - \#\mathcal{T}_0 \lesssim \|g\|_{L^q(C)}^{2(d-1)} \tau^{-2(d-1)},$$

where we used (7.39) to derive the last inequality. This ends the proof.  $\square$

### 7.3.3 The case $f = \text{div } \mathbf{g}$ with $\mathbf{g} \in L^2(\Omega; \mathbb{R}^d)$

A characterization of distributions in  $H^{-1}(\Omega)$  is given in [Evans 2010, Section 5.9.1]: they are of the form

$$f = f_0 + \text{div } \mathbf{g}$$

with  $f_0 \in L^2(\Omega)$ ,  $\mathbf{g} \in L^2(\Omega; \mathbb{R}^d)$ . Since we have already treated separately the ubiquitous case  $\mathbf{g} = \mathbf{0}$  in Section 7.3.1, we consider now the case  $f_0 = 0$ . Therefore,

$$\langle f, v \rangle = - \int_{\Omega} \mathbf{g} \cdot \nabla v \quad \forall v \in H_0^1(\Omega) \quad (7.40)$$

gives the action of  $f$  on  $v$  and its norm is [Evans 2010, Section 5.9.1]

$$\|f\|_{H^{-1}(\Omega)} = \inf \{ \|\mathbf{g}\|_{L^2(\Omega)} : \mathbf{g} \in L^2(\Omega; \mathbb{R}^d) \text{ satisfies (7.40)} \}. \quad (7.41)$$

Since adding the curl of a smooth vector field to  $\mathbf{g}$  does not change (7.40), we realize that the actual computation of (7.41) is problematic. We assume here that

$\mathbf{g}$  is given and simply deal directly with  $\mathbf{g}$  thereby exploiting the relation

$$\|f\|_{H^{-1}(\Omega)} \leq \|\mathbf{g}\|_{L^2(\Omega)}; \quad (7.42)$$

this leads to a surrogate estimator. We first approximate  $\mathbf{g}$  by discontinuous piecewise polynomials of degree  $n_f \leq n - 1$ , namely we compute the  $L^2$ -projection  $\mathbf{g}_{\mathcal{T}} = \Pi_{\mathcal{T}} \mathbf{g}$  onto  $[\mathbb{S}_{\mathcal{T}}^{n_f, -1}]^d$ , then we let  $f_{\mathcal{T}} := \operatorname{div} \mathbf{g}_{\mathcal{T}} \in \mathbb{F}_{\mathcal{T}} \subset H^{-1}(\Omega)$  be the approximation of  $f$ :

$$\langle f_{\mathcal{T}}, v \rangle = - \sum_{T \in \mathcal{T}} \int_T \operatorname{div} \mathbf{g}_{\mathcal{T}} v - \sum_{F \in \mathcal{F}} \int_F [[\mathbf{g}_{\mathcal{T}}]] \cdot \mathbf{n}_F v \quad \forall v \in H_0^1(\Omega).$$

We see that for  $z \in \mathcal{V}$ ,  $f_{\mathcal{T}}|_{\omega_z}$  has the form of a functional in  $\mathbb{F}(\mathcal{T}_z)$  (see (7.27)) with  $q_T = \operatorname{div} \mathbf{g}_{\mathcal{T}}|_T \in \mathbb{P}_{n_f-1}$ ,  $q_F = [[\mathbf{g}_{\mathcal{T}}]] \cdot \mathbf{n}_F \in \mathbb{P}_{n_f}$  for all  $T \in \mathcal{T}$ ,  $F \in \mathcal{F}$ , but with smaller polynomial degree than functions in  $\mathbb{F}(\mathcal{T}_z)$ . We next exploit the local near-best approximation (7.28) to replace  $P_{\mathcal{T}} f$  by  $f_{\mathcal{T}}$

$$\|f - P_{\mathcal{T}} f\|_{H^{-1}(\omega_z)} \leq C_P \|f - f_{\mathcal{T}}\|_{H^{-1}(\omega_z)} \leq C_P \|\mathbf{g} - \mathbf{g}_{\mathcal{T}}\|_{L^2(\omega_z)} \quad (7.43)$$

by virtue of (7.42) with  $\Omega$  replaced by  $\omega_z$ . This leads to the surrogate element-wise oscillation  $\widetilde{\operatorname{osc}}_{\mathcal{T}}(\mathbf{g}, T)_2 := \|\mathbf{g} - \mathbf{g}_{\mathcal{T}}\|_{L^2(T)}$ , which satisfies Assumption 7.3 (monotonicity of local oscillation). We thus have the global surrogate

$$\widetilde{E}_{\mathcal{T}}(f)_{-1}^2 := \sum_{T \in \mathcal{T}} \widetilde{\operatorname{osc}}_{\mathcal{T}}(\mathbf{g}, T)_2^2.$$

**Corollary 7.21 (approximation class of  $\operatorname{div} \mathbf{g}$ )** *Let the initial subdivision  $\mathcal{T}_0$  of  $\Omega \subset \mathbb{R}^d$  satisfy Assumption 6.19 (initial labeling) and  $\mathcal{T} \in \mathbb{T}$  be any admissible refinement of  $\mathcal{T}_0$ . Let  $\tau > 0$  be the target tolerance and  $b \geq 1$  be the number of bisections performed on each marked element. Let  $(\mathbf{g}, s, 0, p, 2)$  satisfy Assumption 7.2 (admissible set of parameters for GREEDY) with local oscillations  $\{\widetilde{\operatorname{osc}}(\mathbf{g}, T)_2\}_{T \in \mathcal{T}}$ . Then  $[\widehat{\mathcal{T}}, \widehat{f}] = \text{GREEDY}(\mathcal{T}, \tau, 2, b, f)$  terminates in a finite number of steps with  $\widetilde{E}_{\widehat{\mathcal{T}}}(f)_{-1} \leq \tau$ , whence  $E_{\widehat{\mathcal{T}}}(f)_{-1} \leq C_{\text{data}} \tau$ . Moreover, the number  $N$  of marked elements by GREEDY satisfies*

$$N \lesssim \|\mathbf{g}\|_{X_p^s(\Omega)}^{\frac{d}{s}} \tau^{-\frac{d}{s}}.$$

In particular  $f = \operatorname{div} \mathbf{g} \in \mathbb{F}_{\frac{s}{d}}$  with

$$|f|_{\mathbb{F}_{\frac{s}{d}}} \lesssim \|\mathbf{g}\|_{X_p^s(\Omega)}.$$

*Proof.* Apply Corollary 7.6 (performance of GREEDY) with  $q = 2$  to  $\mathbf{g}$ .  $\square$

#### 7.4 DATA module

We summarize now in one single algorithm, called DATA, all the developments in Sections 7.2.2, 7.2.3, and 7.3. We first recall that Corollaries 7.12 (complexity of constrained GREEDY for  $A$ ) and 7.16 (complexity of constrained GREEDY for  $c$ )

deliver piecewise polynomial approximations  $(\widehat{A}, \widehat{c})$  of the coefficients  $(A, c)$  over an admissible mesh  $\widehat{\mathcal{T}}$  that satisfy both the global errors estimates

$$E_{\widehat{\mathcal{T}}}(A)_{q_A} \leq C_{\text{data}} \tau, \quad E_{\widehat{\mathcal{T}}}(c)_{q_c} \leq C_{\text{data}} \tau,$$

where  $2 \leq q_A, q_c \leq \infty$  are the corresponding integrability indices, as well as the structural constraint (5.51).

The situation for the load  $f$  is more intricate due to the evaluation of the nonlocal norm  $H^{-1}(\Omega)$ , which requires further structure of  $f$  besides regularity. Section 7.3 provides three examples of practical significance that allow for computable surrogate errors  $\widetilde{E}_{\mathcal{T}}(f)_{-1}$  larger than the desired oscillations  $E_{\mathcal{T}}(f)_{-1}$ . Since these examples have different requirements for the approximation procedure to work, we gather the salient structural points in the following assumption.

**Assumption 7.22 (structure of  $f$ )** Let  $(s_f, p_f)$  denote the additional regularity-integrability indices of  $f$  beyond the basic  $H^{-1}$ -regularity, which are required by Assumption 7.2 (admissible set of parameters for GREEDY). Let  $|f|_{\widetilde{X}_{p_f}^{s_f}(\Omega; \mathcal{T}_0)}$  be a measure of piecewise regularity of  $f$  in  $\mathcal{T}_0$  expressed below in terms of surrogates. Assume that either one of the following cases holds and note that all accumulate local oscillations in  $\ell^2$ .

- $f \in L^q(\Omega)$ , with  $\frac{2d}{d+2} < q \leq \infty$ . Let  $\widetilde{\text{osc}}_{\mathcal{T}}(f, T)_q = h_T^{t_f} \|f - \widehat{f}\|_{L^q(T)}$  be the local oscillation with  $t_f = 1 + d(\frac{1}{2} - \frac{1}{q}) \geq 0$  and  $(f, s_f, t_f, p_f, 2)$  satisfy Assumption 7.2, and set  $|f|_{\widetilde{X}_{p_f}^{s_f}(\Omega; \mathcal{T}_0)} := |f|_{X_{p_f}^{s_f}(\Omega; \mathcal{T}_0)}$ .
- $f = g\delta_C$  where  $C \subset \Omega$  is a Lipschitz hyper-surface and  $g \in L^q(C)$  with  $\frac{2(d-1)}{d} < q \leq \infty$ . Let  $\widetilde{\text{osc}}_{\mathcal{T}}(g, T)_q = h_T^r \|g\|_{L^q(T \cap C)}$  be the local oscillation with  $r = \frac{d}{2} - \frac{1}{q}(d-1) > 0$ . Set  $s_f = 0$ ,  $t_f = d/(2(d-1))$ ,  $p_f = q$ , and  $|f|_{\widetilde{X}_{p_f}^{s_f}(\Omega; \mathcal{T}_0)} := \|g\|_{L^q(C)}$ .
- $f = \text{div } \mathbf{g}$  with  $\mathbf{g} \in L^2(\Omega; \mathbb{R}^d)$ . Let  $\widetilde{\text{osc}}_{\mathcal{T}}(f, T)_2 = \|\mathbf{g} - \Pi_{\mathcal{T}} \mathbf{g}\|_{L^2(T)}$  be the local oscillation,  $t_f = 0$ , and  $(\mathbf{g}, s_f, t_f, p_f, 2)$  satisfy Assumption 7.3, and set  $|f|_{\widetilde{X}_{p_f}^{s_f}(\Omega; \mathcal{T}_0)} := \|\mathbf{g}\|_{X_{p_f}^{s_f}(\Omega; \mathcal{T}_0)}$ .

In all these cases, GREEDY algorithms with tolerance  $\tau > 0$  reduce the surrogate error  $\widetilde{E}_{\mathcal{T}}(f)_{-1}^2$  and eventually guarantee that

$$E_{\mathcal{T}}(f)_{-1} \leq C_{\text{data}} \tau,$$

where  $C_{\text{data}} \geq 1$  is the constant appearing in Corollary 7.18, Lemma 7.20, or Corollary 7.21 depending on Assumption 7.22 (structure of  $f$ ).

**Algorithm 7.23 (DATA)** Given a tolerance  $\tau > 0$  and an arbitrary conforming grid  $\mathcal{T} \in \mathbb{T}$ , not necessarily  $\mathcal{T}_0$ , DATA finds a conforming refinement  $\widehat{\mathcal{T}} \geq \mathcal{T}$  of  $\mathcal{T}$  and approximate data  $\widehat{\mathcal{D}} = (\widehat{A}, \widehat{c}, \widehat{f}) \in \mathbb{D}_{\widehat{\mathcal{T}}}$  over  $\widehat{\mathcal{T}}$  such that

$$\|\mathcal{D} - \widehat{\mathcal{D}}\|_{D(\Omega)} = E_{\widehat{\mathcal{T}}}(A)_{q_A} + E_{\widehat{\mathcal{T}}}(c)_{q_c} + E_{\widehat{\mathcal{T}}}(f)_{-1} \leq C_{\text{data}} \tau.$$

```

 $[\widehat{\mathcal{T}}, \widehat{\mathcal{D}}] = \text{DATA}(\mathcal{T}, \tau, \mathcal{D})$ 
 $[\mathcal{T}_A, \widehat{\mathbf{A}}] = \text{GREEDY}(\mathcal{T}, \tau/3, q_A, b, \mathbf{A})$ 
 $\widehat{\mathbf{A}} = \text{CONSTRAINT-A}(\mathcal{T}_A, \alpha_1, \alpha_2, L, \widehat{\mathbf{A}})$ 
Set  $\widehat{\alpha}_1 = \frac{1}{2}\alpha_1$  and  $\widehat{\alpha}_2 = (1 + 4L)\frac{\alpha_2}{2}$ 
 $[\mathcal{T}_c, \widehat{c}] = \text{GREEDY}(\mathcal{T}_A, \tau/3, q_c, b, c)$ 
 $\widehat{c} = \text{CONSTRAINT-c}(\mathcal{T}_c, \widehat{\alpha}_1, L, \widehat{c})$ 
 $[\widehat{\mathcal{T}}, \widehat{f}] = \text{GREEDY}(\mathcal{T}_c, \tau/3, 2, b, f)$ 
return  $\widehat{\mathcal{T}}, \widehat{\mathcal{D}}$ 

```

Note that DATA depends on the threshold parameter  $L > 2$  used in CONSTRAINT-A and CONSTRAINT-c, although for simplicity it is not listed among the input parameters.

The next result summarizes the properties of DATA.

**Corollary 7.24 (performance of DATA)** *Let the initial subdivision  $\mathcal{T}_0$  of  $\Omega \subset \mathbb{R}^d$  satisfy Assumption 6.19 (initial labeling) and  $\mathcal{T} \in \mathbb{T}$  be any admissible refinement of  $\mathcal{T}_0$ . Let  $b \geq 1$  be the number of bisections performed on each marked element. Let the assumptions of Corollaries 7.12 and 7.16 for the coefficients  $(\mathbf{A}, c)$  be valid, and let  $f$  satisfy Assumption 7.22.*

*For any target tolerance  $\tau > 0$  and any threshold parameter  $L > 2$ ,  $[\widehat{\mathcal{T}}, \widehat{\mathcal{D}}] = \text{DATA}(\mathcal{T}, \tau, \mathcal{D})$  terminates in a finite number of iterations and outputs  $\widehat{\mathcal{D}}, \widehat{\mathcal{T}} \in \mathbb{T}$  such that  $\widehat{\mathbf{A}}$  is symmetric and*

$$\widehat{\mathbf{A}} \in M(\widehat{\alpha}_1, \widehat{\alpha}_2), \quad \widehat{c} \in R(\widehat{c}_1, \widehat{c}_2),$$

where  $\widehat{\alpha}_1, \widehat{\alpha}_2$  are given by (7.21) while  $\widehat{c}_1, \widehat{c}_2$  are given by (7.23). Moreover, there is a constant  $C_{\text{data}} \geq 1$  such that DATA terminates with

$$\|\mathcal{D} - \widehat{\mathcal{D}}\|_{D(\Omega)} \leq C_{\text{data}}\tau,$$

and the number  $N$  of elements marked to construct  $\widehat{\mathcal{T}}$  satisfies

$$N \lesssim |\mathcal{D}|_{\mathbb{X}_{\frac{s_{\mathcal{D}}}{d}}}^{\frac{d}{s_{\mathcal{D}}}} \tau^{-\frac{d}{s_{\mathcal{D}}}} \quad (7.44)$$

with  $s_{\mathcal{D}} := \min\{s_A, s_c + t_c, s_f + t_f\}$ , and

$$|\mathcal{D}|_{\mathbb{X}_{\frac{s_{\mathcal{D}}}{d}}} = \left( |\mathbf{A}|_{X_{p_A}^{s_A}(\Omega, \mathcal{T}_0)}^{\frac{d}{s_A}} + |c|_{X_{p_c}^{s_c}(\Omega, \mathcal{T}_0)}^{\frac{d}{s_c}} + |f|_{\widetilde{X}_{p_f}^{s_f}(\Omega, \mathcal{T}_0)}^{\frac{d}{s_f}} \right)^{\frac{s_{\mathcal{D}}}{d}}.$$

*Proof.* Since the local oscillations for  $\mathbf{A}$  and  $c$  satisfy Assumption 7.3 (monotonicity of local oscillations), we deduce that global oscillations do not increase upon refinement, namely for  $\widehat{\mathcal{T}} \geq \mathcal{T}_c \geq \mathcal{T}_A$

$$E_{\widehat{\mathcal{T}}}(\mathbf{A})_{q_A} + E_{\widehat{\mathcal{T}}}(c)_{q_c} \leq E_{\mathcal{T}_A}(\mathbf{A})_{q_A} + E_{\mathcal{T}_c}(c)_{q_c}.$$

In view of Corollaries 7.12 and 7.16, this in turn implies

$$E_{\hat{\mathcal{T}}}(A)_{q_A} + E_{\hat{\mathcal{T}}}(c)_{q_c} \leq C_{\text{data}} \frac{2}{3} \tau.$$

For the load term  $f$ , we invoke Corollary 7.18, Lemma 7.20, or Corollary 7.21, depending on Assumption 7.22 (structure of  $f$ ), to infer that

$$E_{\hat{\mathcal{T}}}(f)_{-1} \leq C_{\text{data}} \frac{1}{3} \tau.$$

Hence,

$$\|\mathcal{D} - \hat{\mathcal{D}}\|_{D(\Omega)} = E_{\hat{\mathcal{T}}}(A)_{q_A} + E_{\hat{\mathcal{T}}}(c)_{q_c} + E_{\hat{\mathcal{T}}}(f)_{-1} \leq C_{\text{data}} \tau$$

as desired. The complexity estimate (7.44) directly follows from the complexity estimates given in Corollaries 7.12 and 7.16 for  $(A, c)$ , and Corollary 7.18, Lemma 7.20, or Corollary 7.21 for  $f$  depending on its structure.  $\square$

Similar ideas apply to approximate non-vanishing Dirichlet data or boundary flux condition for Robin or Neumann problems, but we do not elaborate on this.

## 8 Mesh Refinement: The Bisection Method

This section is devoted to the complexity analysis of REFINES for  $\Lambda$ -admissible triangulations. Precisely, we prove the existence of a constant  $D > 0$  such that

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \leq D \sum_{j=0}^{k-1} \#\mathcal{M}_j, \quad k \geq 0.$$

This kind of result holds for conforming meshes ( $\Lambda = 0$ ) and was stated in Theorem 3.16, and for nonconforming meshes ( $\Lambda > 1$ ) as anticipated in Theorem 3.29. The results of Sections 8.1 and 8.2 are valid for  $d = 2$  but the proofs of the cited theorems extend easily to  $d > 2$ . We refer to the survey [Nochetto *et al.* 2009] for a full discussion for  $d \geq 2$ .

### 8.1 Conforming meshes

#### 8.1.1 Chains and labeling for $d = 2$

In order to study nonlocal effects of bisection for  $d = 2$  we introduce now the concept of chain [Binev *et al.* 2004]; this concept is not adequate for  $d > 2$  [Nochetto *et al.* 2009, Stevenson 2008]. Recall that  $E(T)$  denotes the edge of  $T$  assigned for refinement. To each  $T \in \mathcal{T}$  we associate the element  $F(T) \in \mathcal{T}$  sharing the edge  $E(T)$  if  $E(T)$  is interior and  $F(T) = \emptyset$  if  $E(T)$  is on  $\partial\Omega$ . A *chain*  $C(T, \mathcal{T})$ , with starting element  $T \in \mathcal{T}$ , is a sequence  $\{T, F(T), \dots, F^m(T)\}$  with no repetitions of elements and with

$$F^{m+1}(T) = F^k(T) \text{ for some } k \in \{0, \dots, m-1\}, \text{ or } F^{m+1}(T) = \emptyset;$$

see Figure 8.1. We observe that if an element  $T$  belongs to two different grids, then the corresponding chains may be different as well. Two adjacent elements



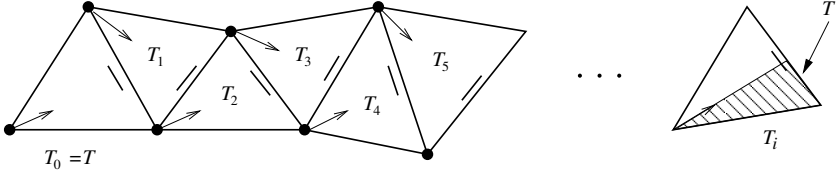


Figure 8.1. Typical chain  $C(T, \mathcal{T}) = \{T_j\}_{j=0}^i$  emanating from  $T = T_0 \in \mathcal{T}$  with  $T_j = F(T_{j-1})$ ,  $j \geq 1$ .

$T, T' = F(T)$  are *compatibly divisible* (or equivalently  $T, T'$  form a *compatible bisection patch*) if  $F(T') = T$ . Hence,  $C(T, \mathcal{T}) = \{T, T'\}$  and a bisection of either  $T$  or  $T'$  does not propagate outside the patch.

**Example** (chains): Let  $\mathcal{F} = \{T_i\}_{i=1}^{12}$  be the forest of Figure 3.5. Then  $C(T_6, \mathcal{T}) = \{T_6, T_7\}$ ,  $C(T_9, \mathcal{T}) = \{T_9\}$ , and  $C(T_{10}, \mathcal{T}) = \{T_{10}, T_8, T_2\}$  are chains, but only  $C(T_6, \mathcal{T})$  is a compatible bisection patch.

To study the structure of chains we rely on the initial labeling (3.35) and the bisection rule of Section 3.5 (see Figure 3.7):

every triangle  $T \in \mathcal{T}$  with generation  $g(T) = i$  receives the label  $(i + 1, i + 1, i)$  with  $i$  corresponding to the refinement edge  $E(T)$ , its side  $i$  is bisected and both new sides as well as the bisector are labeled  $i + 2$  whereas the remaining labels do not change. (8.1)

We first show that once the initial labeling and bisection rule are set, the resulting master forest  $\mathbb{F}$  is uniquely determined: the label of an edge is independent of the elements sharing this edge and no ambiguity arises in the recursion process.

**Lemma 8.1 (labeling)** *Let the initial labeling (3.35) for  $\mathcal{T}_0$  and above bisection rule be enforced. If  $\mathcal{T}_0 \leq \mathcal{T}_1 \leq \dots \leq \mathcal{T}_n$  are generated according to (8.1), then each side in  $\mathcal{T}_k$  has a unique label independent of the two triangles sharing this edge.*

*Proof.* We argue by induction over  $\mathcal{T}_k$ . For  $k = 0$  the assertion is valid due to the initial labeling. Suppose the statement is true for  $\mathcal{T}_k$ . An edge  $S$  in  $\mathcal{T}_{k+1}$  can be obtained in two ways. The first is that  $S$  is a bisector, and so a new edge, in which case there is nothing to prove about its label being unique. The second possibility is that  $S$  was obtained by bisecting an edge  $S' \in \mathcal{S}_k$ . Let  $T, T' \in \mathcal{T}_k$  be the elements sharing  $S'$ , and let us assume that  $E(T') = S'$ . Let  $(i + 1, i + 1, i)$  be the label of  $T'$ , which means that  $S$  is assigned the label  $i + 2$ . By induction assumption over  $\mathcal{T}_k$ , the label of  $S'$  as an edge of  $T$  is also  $i$ . There are two possible cases for the label of  $T$ :

- Label  $(i + 1, i + 1, i)$ : this situation is symmetric,  $E(T) = S'$ , and  $S'$  is bisected with both halves getting label  $i + 2$ . This is depicted in Fig. 8.2.
- Label  $(i, i, i - 1)$ : a bisection of side  $E(T)$  with label  $i - 1$  creates a child  $T''$  with label  $(i + 1, i + 1, i)$  that is compatibly divisible with  $T'$ . Joining the new node of

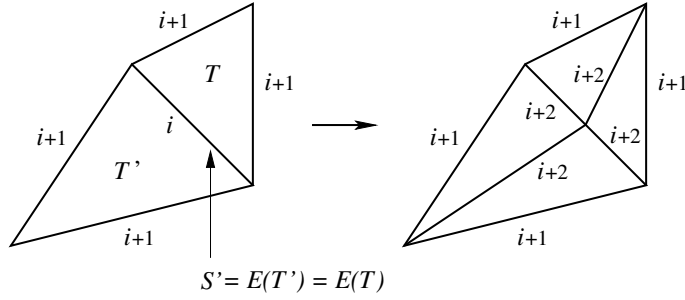


Figure 8.2.  $T$  and  $T'$  form a compatible patch, as they share the generation.

$T$  with the midpoint of  $S'$  creates a conforming partition with level  $i + 2$  assigned to  $S$ . This is depicted in Fig. 8.3.

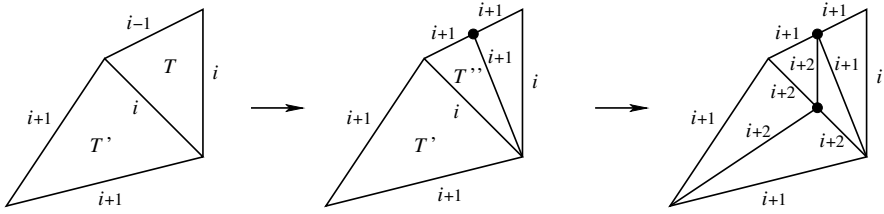


Figure 8.3.  $T'$  form a compatible patch with the child  $T''$  of  $T$ , indeed  $T$  has a lower generation than  $T''$ .

Therefore, in both cases the label  $i + 2$  assigned to  $S$  is the same from both sides, as asserted.  $\square$

The two possible configurations displayed in the two figures above lead readily to the following statement about generations.

**Corollary 8.2 (generation of consecutive elements)** *For any  $\mathcal{T} \in \mathbb{T}$  and  $T, T' \in \mathcal{T}$  with  $T = F(T')$  we either have:*

- (a)  $g(T) = g(T')$  and  $T, T'$  are compatibly divisible, or
- (b)  $g(T) = g(T') - 1$  and  $T'$  is compatibly divisible with a child of  $T$ .

**Corollary 8.3 (generations within a chain)** *For all  $\mathcal{T} \in \mathbb{T}$  and  $T \in \mathcal{T}$ , its chain  $C(T, \mathcal{T}) = \{T_k\}_{k=0}^m$  with  $T_k = F^k(T)$  have the property*

$$g(T_k) = g(T) - k \quad 0 \leq k \leq m - 1$$

*and  $T_m = F^m(T)$  has generation  $g(T_m) = g(T_{m-1})$  or it is a boundary element with lowest labeled edge on  $\partial\Omega$ . In the first case,  $T_{m-1}$  and  $T_m$  are compatibly divisible.*

*Proof.* Apply Corollary 8.2 repeatedly to consecutive elements of  $C(T, \mathcal{T})$ .  $\square$

### 8.1.2 Recursive bisection

Given an element  $T \in \mathcal{M}$  to be refined, the routine `REFINE_RECURSIVE` ( $\mathcal{T}, T$ ) recursively refines the chain  $C(T, \mathcal{T})$  of  $T$ , from the end back to  $T$ , and creates a minimal conforming partition  $\mathcal{T}_* \geq \mathcal{T}$  such that  $T$  is bisected once. This procedure reads as follows:

```

 $[\mathcal{T}_*] = \text{REFINE\_RECURSIVE}(\mathcal{T}, T)$ 
  if  $g(F(T)) < g(T)$ 
     $[\mathcal{T}] = \text{REFINE\_RECURSIVE}(\mathcal{T}, F(T))$ 
  else
    bisect the compatible bisection patch  $C(T, \mathcal{T})$ 
    update  $\mathcal{T}$ 
  return  $\mathcal{T}$ 

```

We denote by  $C_*(T, \mathcal{T}) \subset \mathcal{T}_*$  the recursive refinement of  $C(T, \mathcal{T})$  (or completion of  $C(T, \mathcal{T})$ ) caused by bisection of  $T$ . Since `REFINE_RECURSIVE` refines solely compatible bisection patches, intermediate meshes are always conforming.

We refer to Figure 8.4 for an example of recursive bisection  $C_*(T_{10}, \mathcal{T})$  of  $C(T_{10}, \mathcal{T}) = \{T_{10}, T_8, T_2\}$  in Figure 3.4: `REFINE_RECURSIVE` starts bisecting from the end of  $C(T_{10}, \mathcal{T})$ , namely  $T_2$ , which is a boundary element, and goes back the chain bisecting elements twice until it gets to  $T_{10}$ .

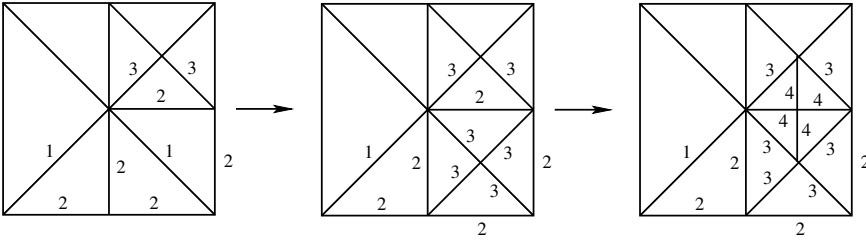


Figure 8.4. Recursive refinement of  $T_{10} \in \mathcal{T}$  in Figure 3.4 by `REFINE_RECURSIVE`. This entails refining the chain  $C(T_{10}, \mathcal{T}) = \{T_{10}, T_8, T_2\}$ , starting from the last element  $T_2 \in \mathcal{T}$ , which form alone a compatible bisection patch because its refinement edge is on the boundary, and continuing with  $T_8 \in \mathcal{T}$  and finally  $T_{10} \in \mathcal{T}$ . Note that the successive meshes are always conforming and that `REFINE_RECURSIVE` bisects elements in  $C(T_{10}, \mathcal{T})$  twice before getting back to  $T_{10}$ .

We now establish a fundamental property of `REFINE_RECURSIVE` ( $\mathcal{T}, T$ ) relating the generation of elements within  $C_*(T, \mathcal{T})$  [Binev *et al.* 2004].

**Lemma 8.4 (recursive refinement)** *Let  $\mathcal{T}_0$  satisfy the labeling (3.35), and let  $\mathcal{T} \in \mathbb{T}$  be a conforming refinement of  $\mathcal{T}_0$ . A call to `REFINE_RECURSIVE` ( $\mathcal{T}, T$ ) terminates, for all  $T$  in the set  $\mathcal{M}$  of marked elements, and outputs the smallest*

conforming refinement  $\mathcal{T}_*$  of  $\mathcal{T}$  such that  $T$  is bisected. In addition, all newly created  $T' \in C_*(T, \mathcal{T})$  satisfy

$$g(T') \leq g(T) + 1. \quad (8.2)$$

*Proof.* We first observe that  $T$  has maximal generation within  $C(T, \mathcal{T})$ . So recursion is applied to elements with generation  $\leq g(T)$ , whence the recursion terminates. We also note that this procedure creates children of  $T$  and either children or grandchildren of triangles  $T_k \in C(T, \mathcal{T}) = \{T_i\}_{i=0}^m$  with  $k \geq 1$ . If  $T'$  is a child of  $T$  there is nothing to prove. If not, we consider first  $m = 1$ , in which case  $T'$  is a child of  $T_1$  because  $T_0$  and  $T_1$  are compatibly divisible and so have the same generation; thus  $g(T') = g(T_1) + 1 = g(T_0) + 1$ . Finally, if  $m > 1$ , then  $g(T_k) < g(T)$  and we apply Corollary 8.3 to deduce

$$g(T') \leq g(T_k) + 2 \leq g(T) + 1,$$

as asserted.  $\square$

The following crucial lemma links generation and distance between  $T$  and  $T' \in C_*(T, \mathcal{T})$ , the latter being defined as [Binev et al. 2004]

$$\text{dist}(T', T) := \inf_{x' \in T', x \in T} |x' - x|.$$

**Lemma 8.5 (distance and generation)** *Let  $T \in \mathcal{M}$ . Any newly created  $T' \in C_*(T, \mathcal{T})$  by `REFINE_RECURSIVE` ( $\mathcal{T}, T$ ) satisfies*

$$\text{dist}(T', T) \leq D_2 \frac{2}{\sqrt{2} - 1} 2^{-g(T')/2}, \quad (8.3)$$

where  $D_2 > 0$  is the constant in (3.34).

*Proof.* Suppose  $T' \subset T_i \in C(T, \mathcal{T})$  has been created by subdividing  $T_i$  (see Figure 8.1). If  $i \leq 1$  then  $\text{dist}(T', T) = 0$  and there is nothing to prove. If  $i > 1$ , then we observe that  $\text{dist}(T', T_{i-1}) = 0$ , whence

$$\begin{aligned} \text{dist}(T', T) &\leq \text{dist}(T_{i-1}, T) + \text{diam}(T_{i-1}) \leq \sum_{k=1}^{i-1} \text{diam}(T_k) \\ &\leq D_2 \sum_{k=1}^{i-1} 2^{-g(T_k)/2} < D_2 \frac{1}{1 - 2^{-1/2}} 2^{-g(T_{i-1})/2}, \end{aligned}$$

because the generations decrease exactly by 1 along the chain  $C(T)$  according to Corollary 8.2(b). Since  $T'$  is a child or grandchild of  $T_i$ , we deduce

$$g(T') \leq g(T_i) + 2 = g(T_{i-1}) + 1,$$

whence

$$\text{dist}(T', T) < D_2 \frac{2^{1/2}}{1 - 2^{-1/2}} 2^{-g(T')/2}.$$

This is the desired estimate.  $\square$

The recursive procedure `REFINE_RECURSIVE` is the core of the routine `REFINE` of Section 3.5: given a conforming mesh  $\mathcal{T} \in \mathbb{T}$  and a subset  $\mathcal{M} \subset \mathcal{T}$  of marked elements, `REFINE` creates a conforming refinement  $\mathcal{T}_* \geq \mathcal{T}$  of  $\mathcal{T}$  such that all elements of  $\mathcal{M}$  are bisected at least once:

```

 $[\mathcal{T}_*] = \text{REFINE}(\mathcal{T}, \mathcal{M})$ 
  for all  $T \in \mathcal{M} \cap \mathcal{T}$  do
     $[\mathcal{T}] = \text{REFINE\_RECURSIVE}(\mathcal{T}, T)$ 
  return  $\mathcal{T}$ 

```

It may happen that an element  $T' \in \mathcal{M}$  is scheduled prior to  $T$  for refinement and  $T \in \mathcal{C}(T', \mathcal{T})$ . Since the call `REFINE_RECURSIVE` ( $\mathcal{T}, T'$ ) bisects  $T$ , its two children replace  $T$  in  $\mathcal{T}$ . This implies that  $T \notin \mathcal{M} \cap \mathcal{T}$ , which prevents further refinement of  $T$ .

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  $\mathcal{M} := \{T \in \mathcal{T} \mid b(T) > 0\}$ .

### 8.1.3 Complexity of bisection for conforming meshes

Figure 8.4 reveals that the issue of propagation of mesh refinement to keep conformity is rather delicate. In particular, an estimate of the form

$$\#\mathcal{T}_k - \#\mathcal{T}_{k-1} \leq C \#\mathcal{M}_{k-1}$$

is not valid with a constant  $C$  independent of  $k$ ; in fact the constant can be proportional to  $k$  according to Figure 8.4.

Binev, Dahmen, and DeVore [Binev *et al.* 2004] for  $d = 2$  and Stevenson [Stevenson 2008] for  $d > 2$  show that control of the propagation of refinement by bisection is possible when considering the collective effect:

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \leq D \sum_{j=0}^{k-1} \#\mathcal{M}_j. \quad (8.4)$$

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

$$\sum_{T \in \mathcal{T} \setminus \mathcal{T}_0} \lambda(T, T_*) \leq C_1 \quad \text{for all } T_* \in \mathcal{M}. \quad (8.5a)$$

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

$$\sum_{T_* \in \mathcal{M}} \lambda(T, T_*) \geq C_2 \quad \text{for all } T \in \mathcal{T} \setminus \mathcal{T}_0. \quad (8.5b)$$

Therefore, summing up (8.5b) and using the upper bound (8.5a) we readily obtain

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

which proves (8.4) for  $\mathcal{T}$  and  $\mathcal{M}$ . In the remainder of this section we design such an allocation function  $\lambda: \mathcal{T} \times \mathcal{M} \rightarrow \mathbb{R}^+$  in several steps and prove that recurrent refinement by bisection yields (8.5) provided  $\mathcal{T}_0$  satisfies (3.35), thereby establishing Theorem 3.16 (complexity of REFINE).

*Construction of the Allocation Function.* The function  $\lambda(T, T_*)$  is defined with the help of two sequences  $(a(\ell))_{\ell=-1}^\infty, (b(\ell))_{\ell=0}^\infty \subset \mathbb{R}^+$  of positive numbers satisfying

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

and  $b(0) \geq 1$ . Valid instances are  $a(\ell) = (\ell + 2)^{-2}$  and  $b(\ell) = 2^{\ell/3}$ .

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

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

where  $D_3 := D_2(1 + 2(\sqrt{2} - 1)^{-1})$ . Therefore, the investment of money by  $T_* \in \mathcal{M}$  is restricted to cells  $T$  that are sufficiently close and are of generation  $g(T) \leq g(T_*) + 1$ . Only elements of these generations can be created during refinement of  $T_*$  according to Lemma 8.4. We stress that except for the definition of  $B$ , this construction is multidimensional and we refer to [Nochetto *et al.* 2009, Stevenson 2008] for details.

The following lemma shows that the total amount of money spend by the allocation function  $\lambda(T, T_*)$  per marked element  $T_*$  is bounded.

**Lemma 8.6 (upper bound)** *There exists a constant  $C_1 > 0$  only depending on  $\mathcal{T}_0$  such that  $\lambda$  satisfies (8.5a), i. e.,*

$$\sum_{T \in \mathcal{T} \setminus \mathcal{T}_0} \lambda(T, T_*) \leq C_1 \quad \text{for all } T_* \in \mathcal{M}.$$

*Proof.* We proceed in two steps.

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

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

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

From (3.34) we learn that  $\text{diam}(T_*) \leq D_2 2^{-g_*/2} \leq 2D_2 2^{-(g_*+1)/2} \leq 2D_2 2^{-g/2}$  as well as  $\text{diam}(T) \leq D_2 2^{-g/2}$  for any  $T \in \mathcal{T}(T_*, g)$ . Hence, all elements of the set  $\mathcal{T}(T_*, g)$  lie inside a ball centered at the barycenter of  $T_*$  with radius  $(D_3 B + 3D_2) 2^{-g/2}$ . Again relying on (3.34) we thus conclude

$$\#\mathcal{T}(T_*, g) D_1 2^{-g} \leq \sum_{T \in \mathcal{T}(T_*, g)} |T| \leq c(D_3 B + 3D_2)^2 2^{-g},$$

whence  $\#\mathcal{T}(T_*, g) \leq c D_1^{-1} (D_3 B + 3D_2)^2 =: C$ .

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

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

which is the desired upper bound.  $\square$

The definition of  $\lambda$  also implies that each refined element receives a fixed amount of money. We show this next.

**Lemma 8.7 (lower bound)** *There exists a constant  $C_2 > 0$  only depending on  $\mathcal{T}_0$  such that  $\lambda$  satisfies (8.5b), i. e.,*

$$\sum_{T_* \in \mathcal{M}} \lambda(T, T_*) \geq C_2 \quad \text{for all } T \in \mathcal{T} \setminus \mathcal{T}_0.$$

*Proof.* We proceed in several steps.

[1] Fix an arbitrary  $T_0 \in \mathcal{T} \setminus \mathcal{T}_0$ . Then there is an iteration count  $1 \leq k_0 \leq k$  such that  $T_0 \in \mathcal{T}_{k_0}$  and  $T_0 \notin \mathcal{T}_{k_0-1}$ . Therefore there exists an  $T_1 \in \mathcal{M}_{k_0-1} \subset \mathcal{M}$  such that  $T_0$  is generated during  $\text{REFINE\_RECURSIVE}(\mathcal{T}_{k_0-1}, T_1)$ . Iterating this process we construct a sequence  $\{T_j\}_{j=1}^J \subset \mathcal{M}$  with corresponding iteration counts  $\{k_j\}_{j=1}^J$  such that  $T_j$  is created by  $\text{REFINE\_RECURSIVE}(\mathcal{T}_{k_j-1}, T_{j+1})$ . The sequence is finite since the iteration counts are strictly decreasing and thus  $k_J = 0$  for some  $J > 0$ , or equivalently  $T_J \in \mathcal{T}_0$ .

Since  $T_j$  is created during refinement of  $T_{j+1}$  we infer from (8.2) 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  $\text{dist}(T_0, T_j) \leq D_3 B g^{-g(T_0)/d}$ .

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

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

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

$$\text{dist}(T_0, T_j) \leq \text{dist}(T_0, T_1) + \text{diam}(T_1) + \text{dist}(T_1, T_j)$$

$$\leq \sum_{i=1}^j \text{dist}(T_{i-1}, T_i) + \sum_{i=1}^{j-1} \text{diam}(T_i).$$

We apply (8.3) for the terms of the first sum and (3.34) for the terms of the second sum to obtain

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

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

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

and the definition of  $\lambda$  then readily implies

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

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

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

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

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

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

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

For  $\ell^* > 0$  we see that  $T_0 \notin \mathcal{T}(T_0, \ell^*, j^*)$ , whence  $\mathcal{T}(T_0, \ell^*, j^*) \subset \mathcal{M}$ . In



addition,  $\lambda(T_0, T_i) = a(\ell^*)$  for all  $T_i \in \mathcal{T}(T_0, \ell^*, j^*)$ . From this we conclude

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

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

$$\sum_{T_* \in \mathcal{M}} \lambda(T_0, T_*) \geq \min\{a(-1), a(0), c_*\} =: C_2 > 0. \quad (8.6)$$

This completes the proof.  $\square$

**Remark 8.8 (complexity with  $b > 1$  bisections)** To show the complexity estimate when REFINE performs  $b > 1$  bisections, the set  $\mathcal{M}_k$  is to be understood as a sequence of *single* bisections recorded in sets  $\{\mathcal{M}_k(j)\}_{j=1}^b$ , which belong to intermediate triangulations between  $\mathcal{T}_k$  and  $\mathcal{T}_{k+1}$  with  $\#\mathcal{M}_k(j) \leq 2^{j-1} \#\mathcal{M}_k$ ,  $j = 1, \dots, b$ . Then we also obtain Theorem 3.16 because

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

In practice, it is customary to take  $b = d$  [Siebert 2012].

## 8.2 Nonconforming meshes

In this subsection, we consider two kinds of nonconforming meshes undergoing a refinement process: a) quadrilateral meshes with at most one hanging node per edge ( $\Lambda = 1$  in the definition of  $\Lambda$ -admissible meshes), and b) triangular meshes having global index bounded by a fixed, but arbitrary  $\Lambda > 1$ .

### 8.2.1 Complexity of bisection for nonconforming quadrilateral meshes

We examine briefly the refinement process for quadrilaterals with one hanging node per edge, which gives rise to the so-called *1-meshes*. The refinement of  $T \in \mathcal{T}$  might affect four elements of  $\mathcal{T}$  for  $d = 2$  (or  $2^d$  elements for any dimension  $d \geq 2$ ), all contained in the *refinement patch*  $R(T, \mathcal{T})$  of  $T$  in  $\mathcal{T}$ . The latter is defined as

$$R(T, \mathcal{T}) := \{T' \in \mathcal{T} \mid T' \text{ and } T \text{ share an edge and } g(T') \leq g(T)\},$$

and is called *compatible* provided  $g(T') = g(T)$  for all  $T' \in R(T, \mathcal{T})$ . The generation gap between elements sharing an edge, in particular those in  $R(T, \mathcal{T})$ , is always  $\leq 1$  for 1-meshes, and is 0 if  $R(T, \mathcal{T})$  is compatible. The element size satisfies

$$h_T = 2^{-g(T)} h_{T_0} \quad \forall T \in \mathcal{T}$$

where  $T_0 \in \mathcal{T}_0$  is the ancestor of  $T$  in the initial mesh  $\mathcal{T}_0$ . Lemma 3.15 is thus valid

$$h_T < \bar{h}_T \leq D 2^{-g(T)} \quad \forall T \in \mathcal{T}. \quad (8.7)$$

Given an element  $T \in \mathcal{M}$  to be refined, the routine `REFINE_RECURSIVE` ( $\mathcal{T}, T$ ) refines recursively  $R(T, \mathcal{T})$  in such a way that the intermediate meshes are always 1-meshes, and reads as follows:

```

[ $\mathcal{T}_*$ ] = REFINE_RECURSIVE ( $\mathcal{T}, T$ )
  if  $g = \min\{g(T'') : T'' \in R(T, \mathcal{T})\} < g(T)$ 
    let  $T' \in R(T, \mathcal{T})$  satisfy  $g(T') = g$ 
    [ $\mathcal{T}$ ] = REFINE_RECURSIVE ( $\mathcal{T}, T'$ )
  else
    subdivide  $T$ 
    update  $\mathcal{T}$  upon replacing  $T$  by its children
  return  $\mathcal{T}$ 

```

The conditional prevents the generation gap within  $R(T, \mathcal{T})$  from getting larger than 1. If it fails, then the refinement patch  $R(T, \mathcal{T})$  is compatible and refining  $T$  increases the generation gap from 0 to 1 without violating the 1-mesh structure. This implies a variant of Lemma 8.4: `REFINE_RECURSIVE` ( $\mathcal{T}, T$ ) creates a minimal 1-mesh  $\mathcal{T}_* \geq \mathcal{T}$  refinement of  $\mathcal{T}$  so that for all newly created elements  $T' \in \mathcal{T}_*$

$$g(T') \leq g(T) + 1 \quad (8.8)$$

and  $T$  is subdivided only *once*. This yields Lemma 8.5: there exist a geometric constant  $D_g > 0$  such that for all newly created elements  $T' \in \mathcal{T}_*$

$$\text{dist}(T, T') \leq D_g 2^{g(T')}. \quad (8.9)$$

The procedure `REFINE_RECURSIVE` is the core of `REFINE`, which is conceptually identical to that in Section 8.1.2. Suppose that each marked element  $T \in \mathcal{M}$  is to be subdivided  $b \geq 1$  times. We assign a flag  $q(T)$  to each element  $T$  which is initialized  $q(T) = b$  if  $T \in \mathcal{M}$  and  $q(T) = 0$  otherwise. The marked set  $\mathcal{M}$  is then the set of elements  $T$  with  $q(T) > 0$ , and every time  $T$  is subdivided it is removed from  $\mathcal{T}$  and replaced by its children, which inherit the flag  $q(T) - 1$ . This avoids the conflict of subdividing again an element that has been previously refined by `REFINE_RECURSIVE`. The procedure `REFINE` ( $\mathcal{T}, \mathcal{M}$ ) reads

```

[ $\mathcal{T}_*$ ] = REFINE ( $\mathcal{T}, \mathcal{M}$ )
  for all  $T \in \mathcal{M} \cap \mathcal{T}$  do
    [ $\mathcal{T}$ ] = REFINE_RECURSIVE ( $\mathcal{T}, T$ );
  end
  return  $\mathcal{T}$ 

```

and its output is a minimal 1-mesh  $\mathcal{T}_* \geq \mathcal{T}$ , refinement of  $\mathcal{T}$ , so that all marked elements of  $\mathcal{M}$  are refined at least  $b$  times. Since  $\mathcal{T}_*$  has one hanging node per side it is thus admissible in the sense of (3.47). However, the refinement may spread outside  $\mathcal{M}$  and the issue of complexity of `REFINE` again becomes non-trivial.

With the above ingredients in place, a statement similar to Theorem 3.16 (com-

plexity of REFINES) for nonconforming quadrilateral meshes follows along the lines of Section 8.1.3.

### 8.2.2 Complexity of bisection for $\Lambda$ -admissible triangular meshes

Let  $\mathcal{T} \in \mathbb{T}^\Lambda$  be a  $\Lambda$ -admissible simplicial mesh. Given any  $T \in \mathcal{T}$ , let us denote again by  $E(T)$  the edge of  $T$  assigned for refinement, i.e., the edge opposite to the newest vertex  $v(T)$ . Let us denote by  $x(T)$  the midpoint of the edge  $E(T)$ .

Two elements  $T', T'' \in \mathcal{T}$  are said *adjacent* if  $E = T' \cap T''$  is an edge for at least one element, and are said *compatible* if they are adjacent and both  $E(T')$  and  $E(T'')$  belong to the same line (see Fig. 8.5, cases A and B).

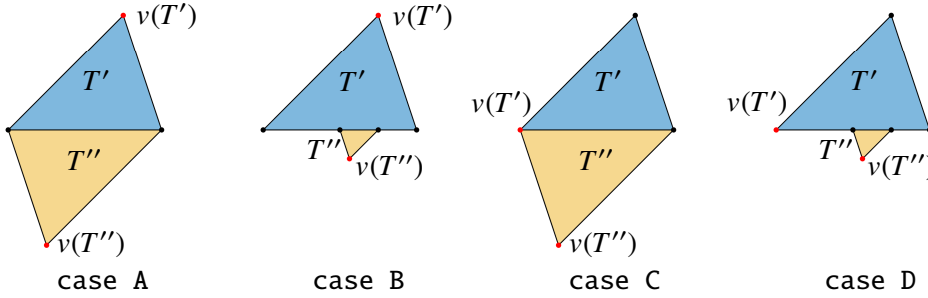


Figure 8.5. The elements  $T'$  and  $T''$  are adjacent in cases A to D. They are compatible in cases A and B, and non-compatible in cases C and D.

The following technical results will be helpful in the design of the refinement procedure.

**Lemma 8.9 (global index of a hanging node)** *Consider an edge  $E = [x', x'']$  of the partition  $\mathcal{T}$ . If  $x \in \mathcal{H} \cap \text{int } E$  is generated by  $m \geq 1$  bisections of  $E$ , then its global index  $\lambda(x)$  satisfies*

$$\lambda(x) = \max(\lambda(x'), \lambda(x'')) + m.$$

*Proof.* If  $m = 1$ , then  $x = x_M$  is the midpoint of  $E$ , and the formula is just the Definition 3.24 of global index. If  $m > 1$ , then  $x$  is generated by bisecting some interval  $[z', z''] \subset E$ , and  $\lambda(x) = \max(\lambda(z'), \lambda(z'')) + 1$ . Exactly one between  $z', z''$  has been generated by  $m - 1$  bisections, whereas the other one has been generated by less than  $m - 1$  bisections. Hence, one concludes by induction.  $\square$

**Lemma 8.10 (reducing the global index of hanging nodes)** *Let  $\mathcal{H} \cap \text{int } E$  contain at least the midpoint  $x_M$  of  $E$ . Assume that a bisection of some element in  $\mathcal{T}$  transforms  $x_M$  into a proper node, and let  $\lambda_{\text{new}}$  denote the new global-index mapping of the nodes in  $\mathcal{H} \cap \text{int } E$  after the bisection. Then there holds*

$$\lambda_{\text{new}}(x) \leq \lambda(x) - 1 \quad \forall x \in \mathcal{H} \cap \text{int } E.$$

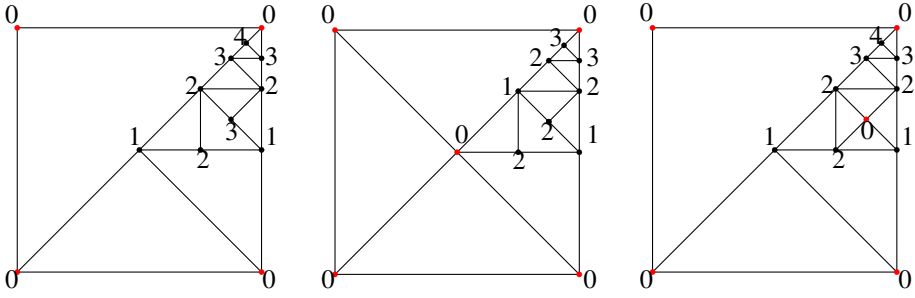


Figure 8.6. Three examples of distributions of proper nodes (red) and hanging nodes (black), with associated global indices  $\lambda$ . The bisection added in the middle picture converts the centered node into proper, and induces nonlocal changes of global indices on chains associated with it; if  $\Lambda = 3$ , the leftmost mesh is not admissible and this procedure is instrumental to restore admissibility. The right picture illustrates the creation of a proper node without nonlocal effects on global indices.

*Proof.* If  $x = x_M$ , then trivially  $\lambda_{\text{new}}(x) = 0 \leq \lambda(x) - 1$ . If  $x \in \mathcal{H} \cap \text{int } E$  is contained, say, in  $(x', x_M)$  and has been generated by  $m > 1$  successive bisections of  $E$ , then it is generated by  $m - 1$  successive bisections of  $[x', x_M]$ . Thus, by applying Lemma 8.9 we get

$$\begin{aligned} \lambda_{\text{new}}(x) &\leq \max(\lambda_{\text{new}}(x'), \lambda_{\text{new}}(x_M)) + m - 1 \\ &= \max(\lambda(x'), 0) + m - 1 = \lambda(x') + m - 1 \\ &\leq \max((\lambda(x'), \lambda(x'')) + m - 1 = \lambda(x) - 1. \end{aligned}$$

This gives the desired estimate.  $\square$

The result just established is the motivation for the proposed refinement strategy, introduced in [?]. Indeed, it assures that in order to reduce the global index of a hanging node sitting on an edge, it is enough to transform the midpoint of the edge into a proper node. The situation is well represented in Figure 8.6.

The following remark will be useful in the sequel.

**Remark 8.11 (facing element)** Given a  $\Lambda$ -admissible mesh  $\mathcal{T}$  and  $T \in \mathcal{T}$ , let  $x(T)$  be the midpoint of  $E(T)$ , and suppose that  $\lambda(x(T)) > \Lambda$ . Then  $x(T)$  is not a node of  $\mathcal{T}$ , whence the edge  $E(T)$  cannot contain any hanging node in its interior. We conclude that there exists a unique adjacent element  $\tilde{T} \in \mathcal{T}$ ,  $\tilde{T} \neq T$ , such that  $T \cap \tilde{T} = E(T)$ . This element will be called the element *facing*  $T$ , and denoted by  $F(T)$ .

Given an element  $T \in \mathcal{T}$  which has been marked for refinement, we are ready to identify those elements in  $\mathcal{T}$  that need be bisected with  $T$  in order to create a  $\Lambda$ -admissible refinement of  $\mathcal{T}$ . Figure 8.7 illustrates the possible situations.

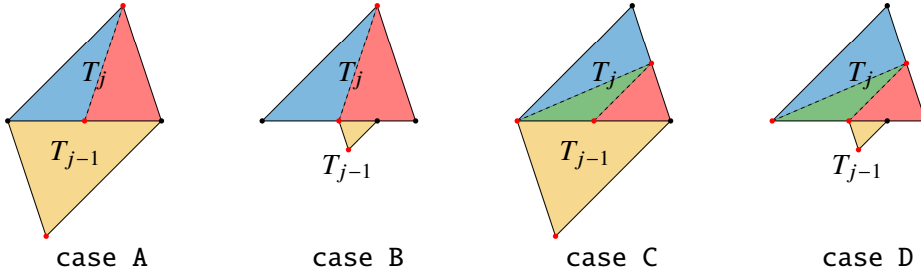


Figure 8.7. Two elements  $T_{j-1}$  and  $T_j$  in the chain  $C(\mathcal{T}, T)$ :  $T_{j-1}$  can be bisected in a  $\Lambda$ -admissible way, only after  $T_j$  is refined once (cases A and B), or twice (cases C and D)

**Definition 8.12 (chain of elements to be refined)** Define by recurrence the chain of elements starting at  $T$

$$C(T, \mathcal{T}) = \{T_0, T_1, \dots, T_k\}$$

for some  $k \geq 0$ , as follows: set first  $T_0 = T$  and, assuming to have defined  $T_j$  for  $j \geq 0$ , then

- (i) if  $\lambda(x(T_j)) \leq \Lambda$ , set  $k = j$  and stop;
- (ii) if  $\lambda(x(T_j)) = \Lambda + 1$  and the facing element  $F(T_j)$  is compatible with  $T_j$ , set  $T_{j+1} = F(T_j)$ ,  $k = j + 1$  and stop;
- (iii) if  $\lambda(x(T_j)) = \Lambda + 1$  and the facing element  $F(T_j)$  is not compatible with  $T_j$ , set  $T_{j+1} = F(T_j)$  and continue.

**Lemma 8.13 (properties of the chain of refinement)** *The chain  $C(T, \mathcal{T})$  has finite length, precisely it holds  $k \leq g(T) + 1$ , where  $g(T)$  is the generation of  $T$ , defined in Sect. 3.5. Furthermore, the sequence of element generations  $\{g(T_j)\}_{j=0}^k$  is not increasing.*

*Proof.* We claim that step (iii) in Definition 8.12 reduces the generation by at least one. In fact,  $T_j$  coincides with or is a refinement of a triangle  $\hat{T} \in \mathbb{T}$  sharing a full edge with  $T_{j+1}$ ; thus  $g(T_j) \geq g(\hat{T})$ . Such triangle  $\hat{T}$  satisfies  $g(\hat{T}) = g(T_{j+1}) + 1$ , whence

$$g(T_{j+1}) = g(\hat{T}) - 1 \leq g(T_j) - 1. \quad (8.10)$$

Therefore, for as long as case (iii) is active, i.e. for all  $j < k$ , we have  $g(T_j) \leq g(T_0) - j$  and

$$0 \leq g(T_{k-1}) \leq g(T_0) - (k - 1),$$

which gives the first statement of the lemma. The monotonicity of  $\{g(T_j)\}_{j=0}^k$  follows from (8.10) and the fact that  $g(T_{k-1}) = g(T_k)$  in case (ii).  $\square$

Once the chain  $C(\mathcal{T}, T)$  is defined, all its elements are refined, starting from

the last one and proceeding backwards. This is accomplished in the following procedure.

```

 $[\mathcal{T}_*] = \text{REFINE\_RECURSIVE}(\mathcal{T}, T, \Lambda)$ 
  if  $\lambda(x(T)) \leq \Lambda$ 
    bisect  $T$ 
    update  $\mathcal{T}$ 
  else if  $F(T)$  is compatible with  $T$ 
    bisect  $F(T)$  and  $T$ 
    update  $\mathcal{T}$ 
  else
     $[\mathcal{T}] = \text{REFINE\_RECURSIVE}(\mathcal{T}, F(T), \Lambda)$ 
  return  $\mathcal{T}$ 

```

**Proposition 8.14 (properties of REFINE\_RECURSIVE)** *If  $\mathcal{T}$  is  $\Lambda$ -admissible, the call  $[\mathcal{T}_*] = \text{REFINE\_RECURSIVE}(\mathcal{T}, T, \Lambda)$  outputs the smallest  $\Lambda$ -admissible refinement  $\mathcal{T}_*$  of  $\mathcal{T}$  such that  $T$  is bisected. In addition, every element  $T' \in \mathcal{T}_*$  generated by this call satisfies*

$$g(T') \leq g(T) + 1. \quad (8.11)$$

*Proof.* Let  $C(T, \mathcal{T}) = \{T_j\}_{j=0}^k$  and observe that, for  $j \geq 1$ , one or two bisections of  $T_j$  convert the midpoint of the edge  $E$  of  $T_j$  shared with  $T_{j-1}$  into a proper node. Therefore, Lemma 8.10 (reducing the global index of hanging nodes) implies that the global indices of all interior nodes to  $E$  decrease by at least 1, and makes the bisection of  $T_{j-1}$   $\Lambda$ -admissible as desired.

To prove (8.11) we take  $j \geq 1$  and consider the following two mutually exclusive cases. If  $T_j$  and  $T_{j-1}$  are compatible, then  $T_j$  is replaced by two elements  $T' \in \mathcal{T}_*$  of generation

$$g(T') = g(T_j) + 1 \leq g(T) + 1,$$

according to Lemma 8.13 (properties of the chain of refinement). On the other hand, if  $T_j$  and  $T_{j-1}$  are not compatible, then  $T_j$  is replaced by one element of generation  $g(T_j) + 1$  and two elements  $T' \in \mathcal{T}_*$  of generation

$$g(T') = g(T_j) + 2 \leq g(T_{j-1}) + 1 \leq g(T) + 1$$

because of (8.10). Finally, the element  $T_0 = T$  is replaced by two elements of generation  $g(T) + 1$ .  $\square$

If one considers the chains starting at any element  $T \in \mathcal{M}$ , one obtains the procedure  $\text{REFINE}(\mathcal{T}, \mathcal{M}, \Lambda)$ , which reads

```

 $[\mathcal{T}_*] = \text{REFINE}(\mathcal{T}, \mathcal{M}, \Lambda)$ 
  for all  $T \in \mathcal{M} \cap \mathcal{T}$  do
     $[\mathcal{T}] = \text{REFINE\_RECURSIVE}(\mathcal{T}, T, \Lambda)$ 

```

return  $\mathcal{T}$

and outputs a minimal  $\Lambda$ -admissible mesh  $\mathcal{T}_* \geq \mathcal{T}$ , refinement of  $\mathcal{T}$ , so that all marked elements of  $\mathcal{M}$  are refined.

*Proof of Theorem 3.29 (complexity of REFINES for  $\Lambda$ -admissible meshes).* The arguments given in Sect. 8.1.3 for the conforming case can be adapted to the current situation. The two crucial properties needed are the relation (8.3) between the distance of two elements in a chain and their generation, which is valid for bisection grids regardless of  $\Lambda$ -admissibility, and the relation (8.11) between generations of elements.  $\square$

### 8.2.3 Mesh overlay and $\Lambda$ -admissibility

Given two partitions  $\mathcal{T}_A$  and  $\mathcal{T}_B$ , denote by  $\mathcal{T}_A \oplus \mathcal{T}_B$  the *overlay* of  $\mathcal{T}_A$  and  $\mathcal{T}_B$ , i.e., the partition whose associated tree is the union of the trees of  $\mathcal{T}_A$  and  $\mathcal{T}_B$ . The following property holds.

**Proposition 8.15 (mesh overlay is  $\Lambda$ -admissible)** *If  $\mathcal{T}_A$  and  $\mathcal{T}_B$  are  $\Lambda$ -admissible, then  $\mathcal{T}_A \oplus \mathcal{T}_B$  remains  $\Lambda$ -admissible.*

*Proof.* Denote here by  $\mathcal{N}$  the set of all nodes obtained by newest-vertex bisection from the root partition  $\mathcal{T}_0$ . Let  $\mathcal{N}_0, \mathcal{N}_A, \mathcal{N}_B, \mathcal{N}_{A+B}$ , resp., be the set of nodes of the partitions  $\mathcal{T}_0, \mathcal{T}_A, \mathcal{T}_B, \mathcal{T}_A \oplus \mathcal{T}_B$ , resp.. It is easily seen that for each  $x \in \mathcal{N} \setminus \mathcal{N}_0$  there exists a unique set  $\mathcal{B}(x) = \{x', x''\} \subset \mathcal{N}$  such that  $x$  is generated by the bisection of the segment  $[x', x'']$ . Furthermore, if  $x \in \mathcal{N}_{A+B}$  is a proper node of  $\mathcal{T}_A$  (of  $\mathcal{T}_B$ , resp.), then it is also a proper node of  $\mathcal{T}_A \oplus \mathcal{T}_B$ .

Let us denote by  $\lambda_A, \lambda_B, \lambda_{A+B}$ , resp., the global-index mappings defined on  $\mathcal{N}_A, \mathcal{N}_B, \mathcal{N}_{A+B}$ , resp.. It is convenient to extend the definition of  $\lambda_A$  and  $\lambda_B$  to the whole  $\mathcal{N}_{A+B}$  by setting

$$\lambda_A(x) = +\infty \quad \text{if } x \in \mathcal{N}_{A+B} \setminus \mathcal{N}_A, \quad \lambda_B(x) = +\infty \quad \text{if } x \in \mathcal{N}_{A+B} \setminus \mathcal{N}_B.$$

With these notations at hand, we are going to prove the inequality

$$\lambda_{A+B}(x) \leq \min(\lambda_A(x), \lambda_B(x)) \quad \forall x \in \mathcal{N}_{A+B}, \quad (8.12)$$

from which the thesis immediately follows.

We proceed by induction on  $k = \lambda_{A+B}(x)$ ,  $x \in \mathcal{N}_{A+B}$ . If  $k = 0$ , the inequality is trivial since  $\lambda_A(x), \lambda_B(x) \geq 0$ . So suppose (8.12) hold up to some  $k \geq 0$ . If  $x \in \mathcal{N}_{A+B}$  satisfies  $\lambda_{A+B}(x) = k + 1 > 0$ , then it is a hanging node of  $\mathcal{T}_A \oplus \mathcal{T}_B$  by definition of global index, hence, it is a hanging node of  $\mathcal{T}_A$  or  $\mathcal{T}_B$ ; without loss of generality, suppose it is a hanging node of  $\mathcal{T}_A$ . If  $x$  is generated by the bisection of the segment  $[x', x'']$ , then again by definition of global index it holds

$$k + 1 = \lambda_{A+B}(x) = \max(\lambda_{A+B}(x'), \lambda_{A+B}(x'')) + 1,$$

which implies

$$\lambda_{A+B}(x') \leq k, \quad \lambda_{A+B}(x'') \leq k.$$

By induction,

$$\lambda_{A+B}(x') \leq \min(\lambda_A(x'), \lambda_B(x')), \quad \lambda_{A+B}(x'') \leq \min(\lambda_A(x''), \lambda_B(x'')),$$

from which we obtain

$$\lambda_{A+B}(x) \leq \max(\lambda_A(x'), \lambda_A(x'')) + 1 = \lambda_A(x)$$

since  $x$  is a hanging node of  $\mathcal{T}_A$ . On the other hand, either  $x \in \mathcal{N}_B$  or  $x \notin \mathcal{N}_B$ . In the latter case,  $\lambda_B(x) = +\infty$ , and (8.12) is proven. In the former case, necessarily  $x$  is a hanging node of  $\mathcal{T}_B$ , hence as above

$$\lambda_{A+B}(x) \leq \max(\lambda_B(x'), \lambda_B(x'')) + 1 = \lambda_B(x),$$

and the thesis is proven.  $\square$

## 9 Discontinuous Galerkin Methods

So far we have studied conforming finite element approximations. In this section we present and analyze a two-step AFEM for discontinuous Galerkin methods (dG). The core PDE routine GALERKIN is thereby replaced by GALERKIN-DG, which hinges on the interior penalty discontinuous FEM. We regard dG as a prototype non-conforming method of practical importance and thus the natural first step to investigate the effects of non-conformity within adaptivity.

Finite element functions, being discontinuous, allow for non-conforming meshes to support them. We consider  $\Lambda$ -admissible subdivisions, according to Definition 3.25, where  $\Lambda \geq 0$  restricts the level of non-conformity, and denote by  $\mathbb{T}^\Lambda$  the collection of all  $\Lambda$ -admissible refinements of an initial subdivision  $\mathcal{T}_0$ ; we refer to Section 8 for details. However, we further assume that  $\mathcal{T}_0$  is conforming to limit the level of technicalities.

There are several novel but characteristic aspects of dG. The most notable one is the appearance of jumps in its formulation, to compensate for the lack of  $H^1$ -conformity, as well as in the a posteriori upper bounds and the comparison of Galerkin solutions on different meshes. The lack of monotonicity of these jumps presents a formidable obstruction to the available proof techniques in adaptivity. However, we show in Lemma 9.11 that they are controlled by the residual estimator, thereby enabling us to loosely follow the roadmap of the conforming method, namely Sections 4, 5, and 6. Our approach is based on [Bonito and Nochetto 2010] for the one-step AFEM.

The extra flexibility provided by non-conforming meshes, and corresponding discontinuous functions, does not yield better asymptotic rate in  $H^1$ . An early manifestation of this fact, although written for conforming subdivisions, is Proposition 6.2 (equivalence of classes for  $u$ ). We extend this result below for general  $\Lambda$ -admissible partitions.

One advantage of the two-step AFEM is that its design and analysis allows for  $f \in H^{-1}(\Omega)$  without added difficulties: the function  $f$  is replaced by the discrete functional  $\hat{f} = P_{\mathcal{T}} f \in \mathbb{F}_{\mathcal{T}}$ , which applies to functions in  $\mathbb{S}_{\mathcal{T}}^{n,-1}$ . This is in contrast



with  $f$ , which cannot be applied to functions in  $\mathbb{S}_{\mathcal{T}}^{n,-1}$ . We exploit this property and thereby extend the applicability of dG to load functions in  $H^{-1}(\Omega)$ .

Our intent is to analyze the following algorithm for the approximation of the solution  $u \in H_0^1(\Omega)$  to the coercive problem (2.7).

**Algorithm 9.1 (AFEM-DG-TS)** Given an initial tolerance  $\varepsilon_0 > 0$ , a target tolerance  $\text{tol}$  and initial mesh  $\mathcal{T}_0$ , as well as a safety parameter  $\omega \in (0, 1]$ , AFEM-DG-TS is a two-step algorithm alternating between the resolution of data  $\mathcal{D}$  and the Galerkin solution  $u_{\mathcal{T}}$ :

```

 $[\mathcal{T}, u_{\mathcal{T}}] = \text{AFEM-DG-TS}(\mathcal{T}_0, \varepsilon_0, \omega, \text{tol})$ 
  set  $k = 0$ 
  do
     $[\widehat{\mathcal{T}}_k, \widehat{\mathcal{D}}_k] = \text{DATA}(\mathcal{T}_k, \mathcal{D}, \omega \varepsilon_k)$ 
     $[\mathcal{T}_{k+1}, u_{k+1}] = \text{GALERKIN-DG}(\widehat{\mathcal{T}}_k, \widehat{\mathcal{D}}_k, \varepsilon_k)$ 
     $\varepsilon_{k+1} = \frac{1}{2} \varepsilon_k$ 
     $k \leftarrow k + 1$ 
  while  $\varepsilon_{k-1} > \text{tol}$ 
  return  $\mathcal{T}_k, u_k$ 

```

In AFEM-DG-TS, the module  $\text{DATA}(\mathcal{T}, \mathcal{D}, \tau)$  is the same as described in Section 5.4.2 except that it produces approximate data  $\widehat{\mathcal{D}} \in \mathbb{D}_{\widehat{\mathcal{T}}}$ , defined in (5.2), subordinate with a  $\Lambda$ -admissible refinement  $\widehat{\mathcal{T}}$  of  $\mathcal{T}_0$  for  $\Lambda \geq 0$  rather than  $\Lambda = 0$  (conforming). The discrete data  $\widehat{\mathcal{D}}$  also satisfies the structural assumption (5.51) as discussed in Section 7. It is worth pointing out that the projection  $P_{\mathcal{T}}$  used to approximate the right-hand side  $f \in H^{-1}(\Omega)$  as well as all the results and algorithms presented in Section 7.4 are restricted to conforming subdivisions  $\mathbb{T}$ . We briefly discuss in Section 9.7 the extension of  $P_{\mathcal{T}}$  and DATA to  $\Lambda$ -admissible subdivisions. Algorithm 9.17 describes the module GALERKIN-DG, the counterpart of GALERKIN for dG formulations.

In Section 9.1 we introduce notations and tools relevant for the characterization of discontinuous finite elements. Among them is the operator  $\mathcal{I}_{\mathcal{T}}^{\text{dG}}$  that projects piecewise polynomial functions onto globally continuous piecewise polynomial functions. It is instrumental to derive a Poincaré inequality on the discontinuous spaces and guarantee that the approximation classes  $\mathbb{A}_s^{-1}$  for the solution  $u$  using discontinuous approximation on  $\Lambda$ -admissible subdivisions are equivalent to their conforming counter-parts  $\mathbb{A}_s^0$  introduced in Section 6. We present the discontinuous Galerkin method in Section 9.2. We start with the standard symmetric interior penalty, discuss its drawbacks regarding the unnecessary regularity beyond  $H_0^1(\Omega)$  imposed on the exact solution  $u$ , and describe a reformulation valid in  $H_0^1(\Omega)$ . The latter suffers from lack of consistency that needs to be accounted for. The a posteriori estimates for the perturbed problem (5.5) are derived in Section 9.3. Because the data is polynomial within GALERKIN-DG, the a posteriori estimators

are oscillation free. The GALERKIN-DG module is analyzed in Section 9.4 while the discussion of rate-optimality of AFEM-DG-TS is reserved for Section 9.5.

### 9.1 Discontinuous Galerkin setting

We start with an initial conforming subdivision  $\mathcal{T}_0$  made of simplices or hexahedra satisfying Assumption 6.19 (initial labeling). Given  $\Lambda > 0$ , the refinement procedure REFINE is designed to produce a  $\Lambda$ -admissible sequence of meshes  $\mathbb{T}^\Lambda$  obeying Theorem 3.29 (REFINE for  $\Lambda$ -admissible meshes). From now on, we do not specify the dependency on  $\Lambda$  in the constants.

#### 9.1.1 Basic setting

For  $\mathcal{T} \in \mathbb{T}^\Lambda$ , we denote by  $\mathbb{V}_{\mathcal{T}}^{-1} := \mathbb{S}_{\mathcal{T}}^{n,-1} := \prod_{T \in \mathcal{T}} \mathbb{P}_n(T)$  the space of piecewise polynomials of degree at most  $n \geq 1$  subordinate to a partition  $\mathcal{T}$ . In contrast with the conforming spaces

$$\mathbb{V}_{\mathcal{T}}^0 := \mathbb{S}_{\mathcal{T}}^{n,0} \cap H_0^1(\Omega)$$

considered earlier, the space  $\mathbb{V}_{\mathcal{T}}^{-1}$  consists of (possibly) discontinuous functions across the elements  $T \in \mathcal{T}$  and do not necessarily satisfy the vanishing boundary condition. Continuity across elements and vanishing boundary condition will be weakly imposed in the discontinuous Galerkin formulations.

We recall from Section 3.7 that for a proper (interior) node  $P \in \mathcal{P}$ , the *domain of influence*  $\omega_{\mathcal{T}}(P) = \text{supp}(\psi_P)$  is the support of the Lagrange basis function  $\psi_P \in \mathbb{V}_{\mathcal{T}}^0$  associated with the node  $P$ ; we refer to Fig. 3.11. Since the sequence of meshes is  $\Lambda$ -admissible, Proposition 3.27 (size of the domain of influence) shows that the number of elements  $T \in \mathcal{T}$  such that  $T \subset \omega_{\mathcal{T}}(P)$  is uniformly bounded for  $\mathcal{T} \in \mathbb{T}^\Lambda$ .

The set of faces associated with a subdivision  $\mathcal{T} \in \mathbb{T}^\Lambda$  is denoted  $\mathcal{F}^+ := \mathcal{F}^+(\mathcal{T})$ , and it contains boundary faces as well as interior faces. The set of interior faces is denoted  $\mathcal{F}$ . For a face  $F \in \mathcal{F}^+$ , we denote by  $\{\{v\}\}_F$  and  $[[v]]_F$  the average and jump operators across a face  $F$ . To define them precisely, we associate for each face  $F \in \mathcal{F}^+$  one of the two unit normals  $\mathbf{n}_F$ . The choice of  $\mathbf{n}_F$  is fixed but irrelevant as long as the outward pointing normal to  $\Omega$  is chosen for boundary faces. Let  $T_{\pm} \in \mathcal{T}$  be the elements that share the interior face  $F$ , namely  $F = T_- \cap T_+$ , and  $\mp \mathbf{n}_F$  be their outward pointing normals. Now, given  $v \in \mathbb{V}_{\mathcal{T}}^{-1}$ , let  $v_{\pm} := v|_{T_{\pm}}$  and define for an interior face  $F$

$$\{\{v\}\}_F := \frac{1}{2}(v_- + v_+)|_F, \quad [[v]]_F := (v_- - v_+)|_F, \quad (9.1)$$

By convention, we set  $\{\{v\}\}_F := v_-$  and  $[[v]]_F := v_-$  whenever  $F$  is a boundary face. These definitions extend readily for vector valued functions.

We use the subscript  $\mathcal{T}$  to denote the piecewise version of differential operators. For instance, the *broken gradient*  $\nabla_{\mathcal{T}}$  is the piecewise gradient  $\nabla_{\mathcal{T}} v|_{\hat{T}} = \nabla v|_{\hat{T}}$  for

$T \in \mathcal{T}$  and  $v \in \mathbb{V}_{\mathcal{T}}^{-1}$ . For simplicity, we write

$$\|v\|_{L^2(\tau)}^2 := \sum_{T \in \tau} \|v\|_{L^2(T)}^2$$

for any subset  $\tau \subset \mathcal{T}$  of elements and

$$\|v\|_{L^2(\sigma)}^2 := \sum_{F \in \sigma} \|v\|_{L^2(F)}^2$$

for any subset  $\sigma \subset \mathcal{F}^+$  of faces. We also define a meshsize function  $h := h_{\mathcal{T}} : \overline{\Omega} \rightarrow (0, \infty)$  such that  $h|_{\hat{T}} \approx \text{diam}(T)$  for  $T \in \mathcal{T}$  and  $h|_F \approx \text{diam}(F)$  for  $F \in \mathcal{F}^+$ . With such notations at hand, the broken  $H^1$  space

$$\mathbb{E}_{\mathcal{T}} := H^1(\Omega; \mathcal{T}) = \prod_{T \in \mathcal{T}} H^1(T),$$

is endowed with the mesh dependent seminorm

$$\|v\|_{a, \mathcal{T}}^2 := \|\nabla_{\mathcal{T}} v\|_{L^2(\mathcal{T})}^2 + a \|h^{-1/2} [[v]]\|_{L^2(\mathcal{F}^+)}^2, \quad (9.2)$$

where  $a$  is some positive parameter. We will prove below that this is indeed a norm.

With these notations, we can extend functionals  $f \in \mathbb{F}_{\mathcal{T}}$  in Definition 4.17 to  $\mathbb{V}_{\mathcal{T}}^{-1}$  for  $\mathcal{T} \geq \hat{\mathcal{T}}$ . Before doing so, recall that for  $\hat{f} \in \mathbb{F}_{\hat{\mathcal{T}}}$  and  $v \in H_0^1(\Omega)$  we have

$$\langle \hat{f}, v \rangle = \sum_{\hat{T} \in \hat{\mathcal{T}}} \int_{\hat{T}} \hat{f} v + \sum_{\hat{F} \in \mathcal{F}(\hat{\mathcal{T}})} \int_{\hat{F}} \hat{f} v,$$

where, compared to Definition 4.17, we slightly abused the notation

$$\hat{f}|_{\hat{T}} = \hat{f}_T \in \mathbb{P}_{2n-2}(\hat{T}) \quad \hat{f}|_{\hat{F}} = \hat{f}_{\hat{F}} \in \mathbb{P}_{2n-1}(\hat{F}).$$

In view of this, we can extend the duality pairing to  $\mathbb{V}_{\mathcal{T}}^{-1}$  by setting

$$\langle \hat{f}, v \rangle_{\mathcal{T}} := \sum_{T \in \mathcal{T}} \int_T \hat{f} v + \sum_{F \in \mathcal{F}} \int_F \hat{f} \{v\} \quad (9.3)$$

so that consistency in  $H_0^1(\Omega)$  is preserved

$$\langle \hat{f}, v \rangle_{\mathcal{T}} = \langle \hat{f}, v \rangle \quad \forall v \in H_0^1(\Omega). \quad (9.4)$$

### 9.1.2 Interpolation operator $\mathcal{I}_{\mathcal{T}}^{dG}$

We shall need the interpolation operator  $\mathcal{I}_{\mathcal{T}}^{dG} : \mathbb{E}_{\mathcal{T}} \rightarrow \mathbb{V}_{\mathcal{T}}^0$  from [Bonito and Nochetto 2010]. Its construction is based on an original idea of Clément [Clément 1975], see also [Bernardi and Girault 1998] and other alternatives [Brenner 2003, Bonito, Nochetto and Ntogkas 2021].

Before embarking on the construction of  $\mathcal{I}_{\mathcal{T}}^{dG}$ , we introduce a few notations. For an interior or boundary proper node  $P \in \mathcal{P}$  of the subdivision  $\mathcal{T}$ , we denote by

$$\mathbb{V}_{\omega_{\mathcal{T}}(P)}^0 := H_0^1(\Omega) \cap \prod_{T \subset \omega_{\mathcal{T}}(P)} \mathbb{P}_n(T) \quad (9.5)$$

the space of continuous piecewise polynomial with support on the domain of influence  $\omega_{\mathcal{T}}(P)$  of  $P$  and vanishing on  $\partial\Omega$ . When the underlying grid  $\mathcal{T}$  is clear from the context, we will simplify the notation and write  $\mathbb{V}_P^0 := \mathbb{V}_{\omega_{\mathcal{T}}(P)}^0$  and  $\omega_P := \omega_{\mathcal{T}}(P)$ ; we refer to Figs. 3.9 and 8.6.

We now construct  $\mathcal{I}_{\mathcal{T}}^{\text{dG}}$  in two steps. First, we define  $V_P \in \mathbb{V}_P^0$  locally as satisfying

$$\int_{\omega_P} (v - V_P)w = 0, \quad \forall w \in \mathbb{V}_P^0. \quad (9.6)$$

The value  $V_P(P)$  is then used as the nodal value of  $\mathcal{I}_{\mathcal{T}}^{\text{dG}}v$ , namely

$$\mathcal{I}_{\mathcal{T}}^{\text{dG}}v := \sum_{P \in \mathcal{P}} V_P(P)\psi_P, \quad (9.7)$$

and we recall that  $\{\psi_P\}_{P \in \mathcal{P}}$  is a basis of  $\mathbb{V}_{\mathcal{T}}^0$  (see Section 3); note that  $\mathcal{I}_{\mathcal{T}}^{\text{dG}}v = 0$  on  $\partial\Omega$  for all  $v \in \mathbb{B}_{\mathcal{T}}$ . Moreover, including boundary proper nodes in the definition (9.6)–(9.7) and replacing  $H_0^1(\Omega)$  by  $H^1(\Omega)$  in the definition (9.5),  $\mathcal{I}_{\mathcal{T}}^{\text{dG}}$  easily extends to  $\mathbb{S}_{\mathcal{T}}^{n,0}$  without zero trace; we denote this operator  $\mathcal{I}_{\mathcal{T},+}^{\text{dG}} : \mathbb{B}_{\mathcal{T}} \rightarrow \mathbb{S}_{\mathcal{T}}^{n,0}$ . An immediate property of  $\mathcal{I}_{\mathcal{T}}^{\text{dG}}$  is *local invariance*:

$$v \in \mathbb{V}_{\omega_T}^0 \quad \Rightarrow \quad v = \mathcal{I}_{\mathcal{T}}^{\text{dG}}v \quad \text{in } T, \quad (9.8)$$

where  $\omega_T := \bigcup\{\omega_P : P \in \mathcal{P}, T \subset \omega_P\}$ ; a similar property is valid for  $\mathcal{I}_{\mathcal{T},+}^{\text{dG}}$ . We next gather a few more properties satisfied by  $\mathcal{I}_{\mathcal{T}}^{\text{dG}}$ .

**Lemma 9.2 (interpolation operator)** *Let Assumption 6.19 (initial labeling) hold and let  $\mathcal{T} \in \mathbb{T}^{\Lambda}$ . For  $v \in H_0^1(\Omega)$ , there holds*

$$\|v - \mathcal{I}_{\mathcal{T}}^{\text{dG}}v\|_{L^2(T)} \lesssim \|h\nabla v\|_{L^2(\omega_T)}, \quad \|\nabla \mathcal{I}_{\mathcal{T}}^{\text{dG}}v\|_{L^2(T)} \lesssim \|\nabla v\|_{L^2(\omega_T)}, \quad (9.9)$$

where  $\omega_T$  is defined above. Instead, for  $v \in \mathbb{B}_{\mathcal{T}}$  there holds

$$\begin{aligned} \|v - \mathcal{I}_{\mathcal{T}}^{\text{dG}}v\|_{L^2(T)} + \|h\nabla_{\mathcal{T}}(v - \mathcal{I}_{\mathcal{T}}^{\text{dG}}v)\|_{L^2(T)} \\ \lesssim \|h^{\frac{1}{2}} \llbracket v \rrbracket\|_{L^2(\mathcal{F}^+ \cap \omega_T)} + \|h\nabla_{\mathcal{T}}(v - \Pi_{\mathcal{T}}v)\|_{L^2(\omega_T)}, \end{aligned} \quad (9.10)$$

where  $\Pi_{\mathcal{T}}$  is the  $L^2$  projection operator onto  $\mathbb{V}_{\mathcal{T}}^{-1} = \mathbb{S}_{\mathcal{T}}^{n,-1}$ .

*Proof.* We start with (9.9) and let  $v \in H_0^1(\Omega)$ . The definition (9.6) of the local projection  $V_P \in \mathbb{V}_P^0$  yields for all  $P \in \mathcal{P}$

$$\|V_P\|_{L^2(\omega_P)} \leq \|v\|_{L^2(\omega_P)} \quad \Rightarrow \quad \|V_P\|_{L^\infty(\omega_P)} \lesssim \text{diam}(\omega_P)^{-\frac{d}{2}} \|v\|_{L^2(\omega_P)}.$$

Proposition 3.27 (size of the domain of influence) gives  $\text{diam } \omega_T \leq Ch_T$ , whence the number of  $\omega_P$  containing  $T$  is uniformly bounded. Combining this with the definition (9.7) of  $\mathcal{I}_{\mathcal{T}}^{\text{dG}}$  implies

$$\|\mathcal{I}_{\mathcal{T}}^{\text{dG}}v\|_{L^2(T)} \lesssim \sum_{P \in \mathcal{P}: T \subset \omega_P} |V_P(P)| \|\psi_P\|_{L^2(T)} \lesssim \|v\|_{L^2(\omega_T)} \quad \forall T \in \mathcal{T}. \quad (9.11)$$

Since  $\mathcal{I}_{\mathcal{T}}^{\text{dG}}$  reproduces constants exactly locally, according to (9.8), the first relation in (9.9) follows from invoking the local  $L^2$  stability property (9.11) together with Proposition 6.34 (Bramble-Hilbert for Sobolev spaces). The second relation is proved using the same arguments and an inverse inequality

$$\|\nabla \mathcal{I}_{\mathcal{T}}^{\text{dG}} v\|_{L^2(T)} \lesssim h_T^{-1} \inf_{v_0 \in \mathbb{R}} \|\mathcal{I}_{\mathcal{T}}^{\text{dG}}(v - v_0)\|_{L^2(T)} \lesssim \|\nabla v\|_{L^2(\omega_T)}. \quad (9.12)$$

We consider now  $v \in \mathbb{B}_{\mathcal{T}}$  and let  $\widehat{v} = \Pi_{\mathcal{T}} v \in \mathbb{V}_{\mathcal{T}}^{-1}$ . We intend to prove (9.10) by dealing with  $v - \widehat{v}$  and  $\widehat{v}$  separately and applying the triangle inequality. Since  $v - \widehat{v}$  has zero mean in  $T$  according to (5.66), we apply Lemma 2.3 (second Poincaré inequality) to deduce

$$\|v - \widehat{v}\|_{L^2(T)} \lesssim h_T \|\nabla(v - \widehat{v})\|_{L^2(T)},$$

whence, combining an inverse estimate with (9.11), we further infer that

$$h_T \|\nabla \mathcal{I}_{\mathcal{T}}^{\text{dG}}(v - \widehat{v})\|_{L^2(T)} \lesssim \|\mathcal{I}_{\mathcal{T}}^{\text{dG}}(v - \widehat{v})\|_{L^2(T)} \lesssim \|v - \widehat{v}\|_{L^2(\omega_T)} \lesssim h_T \|\nabla(v - \widehat{v})\|_{L^2(\omega_T)}.$$

This argument yields the inequality (9.10) for  $v - \widehat{v}$ . It remains to deal with  $\widehat{v}$ .

We scale  $\omega_T$  to a reference domain with unit diameter. Estimate (3.49) on the size of the domains of influence guarantees that the number of such reference patches is uniformly finite over  $\mathbb{T}^\Lambda$ . We relabel  $\widehat{v}$  as  $v$  and examine the seminorm  $\|[[v]]\|_{L^2(\mathcal{F}^+ \cap \omega_P)}$  on the space of discontinuous piecewise polynomials

$$\{v \in \Pi_{T \subset \omega_P} \mathbb{P}_n(T) \mid V_P = 0\},$$

where  $V_P$  is defined by (9.6). If this seminorm vanishes then  $v$  is continuous in  $\omega_P$  and thus  $v \in \mathbb{V}_P^0$ , whence the seminorm dominates any norm in this finite dimensional space. Consequently, scaling back gives

$$\|v - V_P\|_{L^2(\omega_P)} + \|h \nabla_{\mathcal{T}}(v - V_P)\|_{L^2(\omega_P)} \lesssim \|h^{\frac{1}{2}} [[v]]\|_{L^2(\mathcal{F}^+ \cap \omega_P)}. \quad (9.13)$$

We now deduce corresponding estimates for  $\mathcal{I}_{\mathcal{T}}^{\text{dG}}$ . For  $T \in \mathcal{T}$  and  $P, Q \in \omega_T \cap \mathcal{P}$ , (9.13) implies that

$$\|V_P - V_Q\|_{L^2(T)} + \|h \nabla_{\mathcal{T}}(V_P - V_Q)\|_{L^2(T)} \lesssim \|h^{\frac{1}{2}} [[v]]\|_{L^2(\mathcal{F}^+ \cap \omega_T)}. \quad (9.14)$$

Consequently, the definition (9.7) of  $\mathcal{I}_{\mathcal{T}}^{\text{dG}}$  yields

$$v - \mathcal{I}_{\mathcal{T}}^{\text{dG}} v = v - \sum_{P \in \omega_T \cap \mathcal{P}} V_P \psi_P = (v - V_Q) - \sum_{P \in \omega_T \cap \mathcal{P}} (V_P - V_Q) \psi_P,$$

which, combined with (9.13) and (9.14), implies

$$\|v - \mathcal{I}_{\mathcal{T}}^{\text{dG}} v\|_{L^2(T)} + \|h \nabla_{\mathcal{T}}(v - \mathcal{I}_{\mathcal{T}}^{\text{dG}} v)\|_{L^2(T)} \lesssim \|h^{1/2} [[v]]\|_{L^2(\mathcal{F}^+ \cap \omega_T)}.$$

This is the desired estimate (9.10) for  $v = \widehat{v} \in \mathbb{V}_{\mathcal{T}}^{-1}$ . To finish the proof we still need to express the right-hand side of the last inequality in terms of  $v \in \mathbb{B}_{\mathcal{T}}$ . Applying the triangle inequality we are left with estimating  $\|[[v - \widehat{v}]]\|_{L^2(F)}$  for

any  $F \in \mathcal{F}^+ \cap \omega_T$ . If  $T_F \in \mathcal{T}$  is an element within  $\omega_T$  that contains  $F$  in its boundary, we employ the scaled trace inequality to arrive at

$$h_T^{\frac{1}{2}} \|v - \widehat{v}\|_{L^2(F)} \lesssim h_T \|\nabla(v - \widehat{v})\|_{L^2(T_F)} + \|v - \widehat{v}\|_{L^2(T_F)} \lesssim h_T \|\nabla(v - \widehat{v})\|_{L^2(T_F)}.$$

Finally, collecting all the estimates completes the proof.  $\square$

We discuss now consequences of Lemma 9.2. The first one is that jumps are solely responsible for controlling the discrepancy between  $v \in \mathbb{V}_{\mathcal{T}}^{-1}$  and  $\mathcal{I}_{\mathcal{T}}^{\text{dG}} v \in \mathbb{V}_{\mathcal{T}}^0$ :

$$\|v - \mathcal{I}_{\mathcal{T}}^{\text{dG}} v\|_{L^2(T)} + \|h \nabla_{\mathcal{T}}(v - \mathcal{I}_{\mathcal{T}}^{\text{dG}} v)\|_{L^2(T)} \lesssim \|h^{\frac{1}{2}} \llbracket v \rrbracket\|_{L^2(\mathcal{F}^+ \cap \omega_T)}, \quad (9.15)$$

because  $\Pi_{\mathcal{T}} v = v$  in  $\omega_T$ . We next observe that (9.10) is also valid for  $\mathcal{I}_{\mathcal{T},+}^{\text{dG}}$  with the same proof. We can thus apply (9.10) for  $\mathcal{I}_{\mathcal{T},+}^{\text{dG}}$  to  $w = v - \mathcal{I}_{\mathcal{T},+}^{\text{dG}} v$ , use the invariance of  $\mathcal{I}_{\mathcal{T},+}^{\text{dG}}$  in  $\mathbb{S}_{\mathcal{T}}^{n,0}$ , and its continuity across internal faces in  $\mathcal{F}$ , to deduce

$$\begin{aligned} \|v - \mathcal{I}_{\mathcal{T},+}^{\text{dG}} v\|_{L^2(T)} + \|h \nabla_{\mathcal{T}}(v - \mathcal{I}_{\mathcal{T},+}^{\text{dG}} v)\|_{L^2(T)} &\lesssim \|h \nabla_{\mathcal{T}}(v - \Pi_{\mathcal{T}} v)\|_{L^2(\omega_T)} \\ &+ \|h^{\frac{1}{2}} \llbracket v \rrbracket\|_{L^2(\mathcal{F} \cap \omega_T)} + \|h^{\frac{1}{2}}(v - \mathcal{I}_{\mathcal{T},+}^{\text{dG}} v)\|_{L^2(\partial\Omega \cap \omega_T)}. \end{aligned} \quad (9.16)$$

A third consequence of (9.10) is the following Poincaré-type inequality on  $\mathbb{E}_{\mathcal{T}}$ .

**Lemma 9.3 (Poincaré-type inequality on  $\mathbb{E}_{\mathcal{T}}$ )** *Let  $\mathcal{T} \in \mathbb{T}^{\Lambda}$  be a  $\Lambda$ -admissible refinement of  $\mathcal{T}_0$  satisfying Assumption 6.19 (initial labeling). There exists  $C_P = C_P(\Omega, \mathcal{T}_0)$ , such that for all  $v \in \mathbb{E}_{\mathcal{T}}$  there holds*

$$\|v\|_{L^2(\Omega)} \leq C_P \left( \|\nabla_{\mathcal{T}} v\|_{L^2(\mathcal{T})} + \|h^{-\frac{1}{2}} \llbracket v \rrbracket\|_{L^2(\mathcal{F})} + \|h^{-\frac{1}{2}} v\|_{L^2(\partial\Omega)} \right). \quad (9.17)$$

In particular, if  $v = 0$  on  $\partial\Omega$  then (9.17) is a dG version of (2.2).

*Proof.* We argue locally with (9.10). First we realize that an argument similar to (9.12) yields  $\|\nabla_{\mathcal{T}}(v - \Pi_{\mathcal{T}} v)\|_{L^2(\omega_T)} \lesssim \|\nabla_{\mathcal{T}} v\|_{L^2(\omega_T)}$ , whence adding over  $T \in \mathcal{T}$

$$\|v - \mathcal{I}_{\mathcal{T}}^{\text{dG}} v\|_{L^2(\Omega)} + \|\nabla \mathcal{I}_{\mathcal{T}}^{\text{dG}} v\|_{L^2(\Omega)} \lesssim \|\nabla_{\mathcal{T}} v\|_{L^2(\Omega)} + \|h^{-\frac{1}{2}} \llbracket v \rrbracket\|_{L^2(\mathcal{F}^+)}.$$

It thus suffices to write

$$\|v\|_{L^2(\Omega)} \leq \|\mathcal{I}_{\mathcal{T}}^{\text{dG}} v\|_{L^2(\Omega)} + \|v - \mathcal{I}_{\mathcal{T}}^{\text{dG}} v\|_{L^2(\Omega)},$$

and invoke (2.2) for  $\mathcal{I}_{\mathcal{T}}^{\text{dG}} v \in H_0^1(\Omega)$  together with the preceding inequality.  $\square$

Another important property obtained using the interpolation operator  $\mathcal{I}_{\mathcal{T}}^{\text{dG}}$  is that the approximation classes  $\mathbb{A}_s^0 := \mathbb{A}_s(H_0^1(\Omega); \mathcal{T}_0)$  defined using globally continuous piecewise polynomial approximations of degree  $\leq n$  on conforming subdivisions are equivalent to those without global continuity on  $\Lambda$ -admissible subdivisions  $\mathcal{T} \in \mathbb{T}^{\Lambda}$ , provided  $\|\cdot\|_{1,\mathcal{T}}$  (defined in (9.2)) is used as norm on  $\mathbb{E}_{\mathcal{T}}$ . We define

$$\sigma_N^{n,-1}(v) := \inf_{\mathcal{T} \in \mathbb{T}_N^{\Lambda}} \inf_{v_{\mathcal{T}} \in \mathbb{S}_{\mathcal{T}}^{n,-1}} \|v - v_{\mathcal{T}}\|_{1,\mathcal{T}} \quad (9.18)$$

and  $\mathbb{A}_s^{-1} := \mathbb{A}_s^{-1}(H_0^1(\Omega); \mathcal{T}_0)$  to be the class of functions  $v \in H_0^1(\Omega)$  such that

$$|v|_{\mathbb{A}_s^{-1}} := \sup_{N \geq \#\mathcal{T}_0} \left( N^s \sigma_N^{n,-1}(v) \right) < \infty \quad \Rightarrow \quad \sigma_N^{n,-1}(v) \leq |v|_{\mathbb{A}_s^{-1}} N^{-s}.$$

Note the scaling parameter  $a$  for jumps in the definition of  $\sigma_N^{n,-1}$  is just  $a = 1$ .

The following result can be traced back to [Bonito and Nochetto 2010].

**Proposition 9.4 (equivalence of classes for  $u$ )** *Let  $\mathcal{T}_0$  be an initial conforming subdivision satisfying Assumption 6.19 (initial labeling). There are two constants  $m \in \mathbb{N}$  and  $C \geq 1$  such that for all  $N \geq \#\mathcal{T}_0$  and all  $v \in H_0^1(\Omega)$*

$$\sigma_N^{n,-1}(v) \leq \sigma_N^{n,0}(v) \quad \text{and} \quad \sigma_{mN}^{n,0}(v) \leq C \sigma_N^{n,-1}(v).$$

*In particular, the approximation classes coincide  $\mathbb{A}_s^0 \equiv \mathbb{A}_s^{-1}$ ,  $s \geq 0$ .*

*Proof.* The start with the first inequality. For  $v \in H_0^1(\Omega)$  and  $N \geq \#\mathcal{T}_0$ , we let  $\mathcal{T} \in \mathbb{T}_N$  be a conforming subdivision of  $\mathcal{T}_0$  and  $v_{\mathcal{T}}^0 \in \mathbb{V}_{\mathcal{T}}^0 \subset \mathbb{V}_{\mathcal{T}}^{-1}$  be such that

$$\sigma_N^{n,0}(v) = |v - v_{\mathcal{T}}^0|_{H_0^1(\Omega)}.$$

Because  $v - v_{\mathcal{T}}^0 \in H_0^1(\Omega)$ , we have  $\|v - v_{\mathcal{T}}^0\|_{1,\mathcal{T}} = |v - v_{\mathcal{T}}^0|_{H_0^1(\Omega)}$  and thus

$$\sigma_N^{n,-1}(v) \leq |v - v_{\mathcal{T}}^0|_{H_0^1(\Omega)} = \sigma_N^{n,0}(v).$$

We now prove the second inequality. For  $v \in H_0^1(\Omega)$  and  $N \geq \#\mathcal{T}_0$ , let  $\mathcal{T} \in \mathbb{T}_N^\Lambda$  be a  $\Lambda$ -admissible mesh with  $N$  elements and  $v_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}^{-1}$  be so that

$$\|v - v_{\mathcal{T}}\|_{1,\mathcal{T}} = \sigma_N^{n,-1}(v).$$

We first show that  $\mathcal{I}_{\mathcal{T}}^{\text{dG}} v_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}^0$  satisfies

$$|v - \mathcal{I}_{\mathcal{T}}^{\text{dG}} v_{\mathcal{T}}|_{H_0^1(\Omega)} \lesssim \sigma_N^{n,-1}(v).$$

Indeed, using the triangle inequality we obtain

$$\|v - \mathcal{I}_{\mathcal{T}}^{\text{dG}} v_{\mathcal{T}}\|_{1,\mathcal{T}} \leq \|v - v_{\mathcal{T}}\|_{1,\mathcal{T}} + \|v_{\mathcal{T}} - \mathcal{I}_{\mathcal{T}}^{\text{dG}} v_{\mathcal{T}}\|_{1,\mathcal{T}}.$$

Interpolation estimate (9.15) yields

$$\|v_{\mathcal{T}} - \mathcal{I}_{\mathcal{T}}^{\text{dG}} v_{\mathcal{T}}\|_{1,\mathcal{T}} \lesssim \|h^{-1/2} [[v_{\mathcal{T}}]]\|_{L^2(\mathcal{F}^+)}, \quad (9.19)$$

because  $v_{\mathcal{T}} - \mathcal{I}_{\mathcal{T}}^{\text{dG}} v_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}^{-1}$ , whence

$$|v - \mathcal{I}_{\mathcal{T}}^{\text{dG}} v_{\mathcal{T}}|_{H_0^1(\Omega)} = \|v - \mathcal{I}_{\mathcal{T}}^{\text{dG}} v_{\mathcal{T}}\|_{1,\mathcal{T}} \leq C \sigma_N^{n,-1}(v)$$

as claimed for a constant  $C \geq 1$  independent of  $v$  and  $N$ . To assert an estimate on  $\sigma_N^{n,0}(v)$ , we now exhibit a conforming refinement  $\overline{\mathcal{T}}$  of  $\mathcal{T}$  with a comparable number of elements. To do this, we note that because  $\mathcal{T} \in \mathbb{T}_N^\Lambda$  is  $\Lambda$ -admissible, it is the product of successive calls  $[\mathcal{T}_j] = \text{REFINE}(\mathcal{T}_{j-1}, T_{j-1})$ ,  $j = 1, \dots, J$ , where  $\mathcal{T}_j$  is

the smallest  $\Lambda$ -admissible refinement of  $\mathcal{T}_{j-1}$  such that the element  $T_{j-1} \in \mathcal{T}_{j-1}$  is bisected once. We now let  $\overline{\mathcal{T}} \in \mathbb{T}$  be the conforming subdivision obtained from the successive calls  $[\overline{\mathcal{T}}_j] = \text{REFINE}(\overline{\mathcal{T}}_{j-1}, \{T_{j-1}\} \cap \overline{\mathcal{T}}_{j-1})$  with  $\overline{\mathcal{T}}_0 = \mathcal{T}_0$  but where this time REFINE produces the smallest *conforming* refinement of  $\overline{\mathcal{T}}_{j-1}$  where the element of  $T_{j-1}$  is bisected once if  $T_{j-1} \in \overline{\mathcal{T}}_{j-1}$  or otherwise  $\overline{\mathcal{T}}_j = \overline{\mathcal{T}}_{j-1}$ . A simple induction argument, exploiting the minimality of the meshes generated by REFINE, reveals that  $\overline{\mathcal{T}}_j \geq \mathcal{T}_j$  for  $0 \leq j \leq J$ . Consequently, Theorem 3.16 (complexity of REFINE) guarantees that

$$\#\overline{\mathcal{T}} - \#\mathcal{T}_0 \leq D \sum_{j=0}^{J-1} \#(\{T_{j-1}\} \cap \overline{\mathcal{T}}_{j-1}) \leq DJ \leq D(\#\mathcal{T} - \#\mathcal{T}_0)$$

whence  $\#\overline{\mathcal{T}} \leq D\#\mathcal{T} \leq mN$  with  $m := \lceil D \rceil$  because  $D \geq 1$ .

Therefore,  $\mathbb{V}_{\overline{\mathcal{T}}}^0 \subset \mathbb{V}_{\mathcal{T}}^0$  because  $\overline{\mathcal{T}}$  is a conforming refinement of  $\mathcal{T}$ . Since  $\#\overline{\mathcal{T}} \leq mN$  and  $\mathcal{I}_{\overline{\mathcal{T}}}^{\text{dG}} v_{\mathcal{T}} \in \mathbb{V}_{\overline{\mathcal{T}}}^0$ , we deduce

$$\sigma_{mN}^{n,0}(v) \leq |v - \mathcal{I}_{\overline{\mathcal{T}}}^{\text{dG}} v_{\mathcal{T}}|_{H_0^1(\Omega)} \leq C\sigma_N^{n,-1}(v),$$

which is the desired inequality. Finally, the equivalence of moduli of approximation yields  $\mathbb{A}_s^0 \equiv \mathbb{A}_s^{-1}$  and completes the proof.  $\square$

**Remark 9.5 (equivalent classes for  $\mathcal{D}$ )** The approximation classes for data  $\mathcal{D} = (A, c, f)$ , namely  $\mathbb{M}_s((L^r(\Omega))^{d \times d}; \mathcal{T}_0, \mathbb{C}_s(L^q(\Omega); \mathcal{T}_0)$  and  $\mathbb{F}_s(H^{-1}(\Omega); \mathcal{T}_0)$ , are defined for conforming subdivisions in Section 6. However, repeating the construction of the smallest conforming refinement  $\overline{\mathcal{T}}$  of any  $\Lambda$ -admissible subdivision  $\mathcal{T}$ , and using the fact that  $\#\overline{\mathcal{T}} \approx \#\mathcal{T}$  proved above, we deduce that these classes are equivalent to their counter-parts on non-conforming meshes. Therefore, we do not repeat the proof here and use from now on the same notation to denote the approximation classes on  $\Lambda$ -admissible subdivisions.

## 9.2 Discontinuous Galerkin formulation

This section discusses the SOLVE routine at the core of the module GALERKIN-DG. Recall that within the two-step method AFEM-DG-TS, data  $\mathcal{D} = (A, c, f)$  is approximated by  $\widehat{\mathcal{D}} = (\widehat{A}, \widehat{c}, \widehat{f}) \in \mathbb{D}_{\widehat{\mathcal{T}}}$  subordinate to a partition  $\widehat{\mathcal{T}} \in \mathbb{T}^\Lambda$ . For a subdivision  $\mathcal{T} \in \mathbb{T}^\Lambda$ ,  $\mathcal{T} \geq \widehat{\mathcal{T}}$ , the Galerkin solution  $[u_{\mathcal{T}}] = \text{SOLVE}(\mathcal{T})$  is constructed to approximate  $\widehat{u} = u(\widehat{\mathcal{D}}) \in H_0^1(\Omega)$ , the exact weak solution of the perturbed problem (5.5) with approximate data  $\widehat{\mathcal{D}} = (\widehat{A}, \widehat{c}, \widehat{f})$  constructed using Algorithm 7.23 (DATA). Corollary 7.24 (performance of DATA) guarantees that the output  $[\widehat{\mathcal{D}}, \widehat{\mathcal{T}}]$  of DATA satisfies the structural assumption

$$\widehat{A} \in M(\widehat{a}_1, \widehat{a}_2), \quad \widehat{c} \in R(\widehat{c}_1, \widehat{c}_2) \quad (9.20)$$



with  $0 < \widehat{\alpha}_1 \leq \widehat{\alpha}_2$  and  $-\frac{\widehat{\alpha}_1}{2C_P^2} \leq \widehat{c}_1 \leq \widehat{c}_2$  upon replacing the Poincaré constant  $C_P$  by the larger constant  $C_P$  appearing in Lemma 9.3 in Algorithm 7.14 (CONSTRAINT-c). We do not specify the dependency on  $\widehat{\alpha}_1$ ,  $\widehat{\alpha}_2$ ,  $\widehat{c}_1$ , and  $\widehat{c}_2$  of the constants appearing in the analysis below. We also emphasize that the constants involved in (9.20) do not depend on  $\widehat{\mathcal{T}}$  and are thus uniform among all the discrete data constructed within AFEM-DG-TS.

Relation (9.20) not only ensures the existence and uniqueness of a solution  $\widehat{u} \in H_0^1(\Omega)$  satisfying the perturbed problem (5.5) but also, as we shall see in Corollary 9.8, the existence and uniqueness of its discontinuous Galerkin approximation. We first present the standard symmetric interior penalty method and point out its consistency requires the exact solution  $u \in H^s(\Omega)$ ,  $s > 3/2$ . To circumvent this rather restrictive assumption, we introduce lifting operators allowing a reformulation valid in  $H^1(\Omega)$ . However, this reformulation is only consistent on the conforming subspace  $\mathbb{V}_{\mathcal{T}}^0 = \mathbb{V}_{\mathcal{T}}^{-1} \cap H_0^1(\Omega)$  and requires our analysis to decompose the discrete space  $\mathbb{V}_{\mathcal{T}}^{-1}$  into  $\mathbb{V}_{\mathcal{T}}^0$  and its complement  $\mathbb{V}_{\mathcal{T}}^{\perp}$  with respect to an appropriate scalar product.

### 9.2.1 The symmetric interior penalty method

The symmetric interior penalty (SIP) formulation is the most standard discontinuous Galerkin method. For  $\mathcal{T} \in \mathbb{T}^\Lambda$ , it consists in finding  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}^{-1}$  satisfying

$$\mathcal{B}_{\mathcal{T}}[u_{\mathcal{T}}, v] = \langle \widehat{f}, v \rangle_{\mathcal{T}}, \quad \forall v \in \mathbb{V}_{\mathcal{T}}^{-1}, \quad (9.21)$$

where  $\mathcal{B}_{\mathcal{T}} : \mathbb{V}_{\mathcal{T}}^{-1} \times \mathbb{V}_{\mathcal{T}}^{-1} \rightarrow \mathbb{R}$  is the bilinear form defined by

$$\begin{aligned} \mathcal{B}_{\mathcal{T}}[w, v] := & \int_{\Omega} (\nabla_{\mathcal{T}} v \cdot \widehat{\mathbf{A}} \nabla_{\mathcal{T}} w + \widehat{c} w v) - \sum_{F \in \mathcal{F}^+} \int_F [[v]] \mathbf{n}_F \cdot \{\{\widehat{\mathbf{A}} \nabla_{\mathcal{T}} w\}\} \\ & - \sum_{F \in \mathcal{F}^+} \int_F [[w]] \mathbf{n}_F \cdot \{\{\widehat{\mathbf{A}} \nabla_{\mathcal{T}} v\}\} + \kappa \sum_{F \in \mathcal{F}^+} \int_F h_F^{-1} [[w]] [[v]]. \end{aligned} \quad (9.22)$$

The parameter  $\kappa > 0$  is responsible for keeping the discontinuity of the Galerkin solution under control and its value is discussed below. Unless specified otherwise, all the constant appearing in the discussion below are independent of  $\kappa$  and the notation  $A \lesssim B$  signifies  $A \leq CB$  with a constant  $C$  independent of the discretization parameters and  $\kappa$ .

A few comments regarding the weak formulation (9.21) are in order. An integration by parts reveals that the method is consistent whenever the exact solution satisfies the additional regularity  $u \in H^s(\Omega)$ ,  $s > 3/2$ . However, we do not make this assumption in the analysis below but rather extend the formulation to the energy space  $\mathbb{E}_{\mathcal{T}} \supset \mathbb{V}_{\mathcal{T}}^{-1}$  using lifting operators. The same integration by parts also indicates that the term  $\sum_{F \in \mathcal{F}^+} \int_F [[w]] \mathbf{n}_F \cdot \{\{\widehat{\mathbf{A}} \nabla v\}\}$  is not necessary but included to achieve a symmetric formulation. Recall that  $\widehat{\mathbf{A}}$  constructed by DATA is symmet-

ric. In addition, the presence of  $\langle \widehat{f}, v \rangle_{\mathcal{T}}$  is not standard but allows for right-hand sides  $\widehat{f} \in \mathbb{F}_{\widehat{\mathcal{T}}}$  and in turn for  $f \in H^{-1}(\Omega)$  within the AFEM-DG-TS algorithm.

### 9.2.2 Lifting operators

The interior penalty bilinear form (9.22) includes inter-element terms

$$\sum_{F \in \mathcal{F}^+} \int_F [[v]] \mathbf{n}_F \cdot \{\{\widehat{\mathbf{A}} \nabla_{\mathcal{T}} w\}\} + \sum_{F \in \mathcal{F}^+} \int_F [[w]] \mathbf{n}_F \cdot \{\{\widehat{\mathbf{A}} \nabla_{\mathcal{T}} v\}\}, \quad (9.23)$$

which are not defined on  $H^1(\Omega)$  but on  $H^s(\Omega)$ ,  $s > 3/2$ . In turn, the method is consistent when  $u \in H^s(\Omega)$ ,  $s > 3/2$ . The key ingredient to extend  $\mathcal{B}_{\mathcal{T}}[w, v]$  to  $\mathbb{B}_{\mathcal{T}} \times \mathbb{B}_{\mathcal{T}}$  without additional regularity is a lifting operator [Brezzi, Manzini, Marini, Pietra and Russo 2000, Arnold, Brezzi, Cockburn and Marini 2002, Perugia and Schötzau 2003, Houston, Schötzau and Wihler 2004, 2007, Bonito and Nochetto 2010] introduced in this section.

For  $n' > 0$ , we define  $\mathcal{L}_{\mathcal{T}}^{n'} : \mathbb{B}_{\mathcal{T}} \rightarrow [\mathbb{S}_{\mathcal{T}}^{n', -1}]^d$  by the relations

$$\int_{\Omega} \mathcal{L}_{\mathcal{T}}^{n'}[v] \cdot \widehat{\mathbf{A}} \mathbf{w} = \sum_{F \in \mathcal{F}^+} \int_F [[v]] \mathbf{n}_F \cdot \{\{\widehat{\mathbf{A}} \mathbf{w}\}\}, \quad \forall \mathbf{w} \in [\mathbb{S}_{\mathcal{T}}^{n', -1}]^d. \quad (9.24)$$

From this definition, we easily deduce a  $L^2$  stability estimate.

**Lemma 9.6 (stability of lift)** *Let  $\mathcal{T} \in \mathbb{T}^{\Lambda}$  be a  $\Lambda$ -admissible subdivision of  $\mathcal{T}_0$  satisfying Assumption 6.19 (initial labeling). Assume  $\widehat{\mathbf{A}} \in M(\widehat{\alpha}_1, \widehat{\alpha}_2)$  with  $0 < \widehat{\alpha}_1 \leq \widehat{\alpha}_2$ . For  $n' \geq 0$  and all  $v \in \mathbb{S}_{\mathcal{T}}^{n', -1}$ , there holds*

$$\|\mathcal{L}_{\mathcal{T}}^{n'}[v]\|_{L^2(\Omega)} \leq C \|h^{-1/2} [[v]]\|_{L^2(\mathcal{F}^+)}, \quad (9.25)$$

where  $C = C(\widehat{\alpha}_2/\widehat{\alpha}_1, n', \mathcal{T}_0)$ .

*Proof.* Let  $v \in \mathbb{S}_{\mathcal{T}}^{n', -1}$  and set  $\mathbf{w} = \mathcal{L}_{\mathcal{T}}^{n'}[v]$  in (9.24) to write

$$\begin{aligned} \|\widehat{\mathbf{A}}^{1/2} \mathcal{L}_{\mathcal{T}}^{n'}[v]\|_{L^2(\Omega)}^2 &= \int_{\Omega} \mathcal{L}_{\mathcal{T}}^{n'}[v] \cdot \widehat{\mathbf{A}} \mathcal{L}_{\mathcal{T}}^{n'}[v] \\ &= \sum_{F \in \mathcal{F}^+} \int_F h^{-1/2} [[v]] \mathbf{n}_F \cdot h^{1/2} \{\{\widehat{\mathbf{A}} \mathcal{L}_{\mathcal{T}}^{n'}[v]\}\} \\ &\leq \|h^{-1/2} [[v]]\|_{L^2(\mathcal{F}^+)} \|h^{1/2} \{\{\widehat{\mathbf{A}} \mathcal{L}_{\mathcal{T}}^{n'}[v]\}\}\|_{L^2(\mathcal{F}^+)}. \end{aligned}$$

A local inverse estimate along with the eigenvalue bounds for  $\widehat{\mathbf{A}} \in M(\widehat{\alpha}_1, \widehat{\alpha}_2)$  yields

$$\|h^{1/2} \{\{\widehat{\mathbf{A}} \mathcal{L}_{\mathcal{T}}^{n'}[v]\}\}\|_{L^2(\mathcal{F}^+)} \leq C \widehat{\alpha}_2 \|\mathcal{L}_{\mathcal{T}}^{n'}[v]\|_{L^2(\Omega)},$$

where  $C$  only depends on  $n'$  and on the shape regularity constant of  $\mathcal{T}_0$ . Combining the above two inequalities and taking advantage again of the assumption  $\widehat{\mathbf{A}} \in M(\widehat{\alpha}_1, \widehat{\alpha}_2)$  implies (9.25).  $\square$

We record two estimates based on (9.25) and used multiple times in the analysis

below. Combining the estimate (9.25) on the lifting operator with assumption (9.20) and a Cauchy-Schwarz inequality we find that

$$\int_{\Omega} \mathcal{L}_{\mathcal{T}}^{n'}[v] \cdot \widehat{\mathbf{A}} \nabla_{\mathcal{T}} w \leq C_{\text{lift}} \|h^{-1/2} [[v]]\|_{L^2(\mathcal{F}^+)} \|\nabla_{\mathcal{T}} w\|_{L^2(\mathcal{T})}, \quad \forall v, w \in \mathbb{E}_{\mathcal{T}}, \quad (9.26)$$

for a constant  $C_{\text{lift}} = C_{\text{lift}}(\widehat{\alpha}_1, \widehat{\alpha}_2, n', \mathcal{T}_0)$  and in particular independent of the discretization parameters and  $\kappa$ . This, together with a Young inequality, yields for any  $\epsilon > 0$  the second estimate for all  $v, w \in \mathbb{E}_{\mathcal{T}}$

$$\int_{\Omega} \mathcal{L}_{\mathcal{T}}^{n'}[v] \cdot \widehat{\mathbf{A}} \nabla_{\mathcal{T}} w \leq \frac{C_{\text{lift}}^2}{2\epsilon} \|h^{-1/2} [[v]]\|_{L^2(\mathcal{F}^+)}^2 + \frac{\epsilon}{2} \|\nabla_{\mathcal{T}} w\|_{L^2(\mathcal{T})}^2. \quad (9.27)$$

We now return to the SIP weak formulation (9.21) and take advantage of the lifting operators to deduce an equivalent expression of the bilinear form  $\mathcal{B}_{\mathcal{T}}$  on  $\mathbb{V}_{\mathcal{T}}^{-1}$ , which is well defined on  $\mathbb{E}_{\mathcal{T}}$ . The problematic inter-elements terms (9.23) are equivalently rewritten as

$$\int_{\Omega} \mathcal{L}_{\mathcal{T}}^{n'}[v] \cdot \widehat{\mathbf{A}} \nabla_{\mathcal{T}} w + \int_{\Omega} \mathcal{L}_{\mathcal{T}}^{n'}[w] \cdot \widehat{\mathbf{A}} \nabla_{\mathcal{T}} v \quad (9.28)$$

provided

$$\nabla_{\mathcal{T}} \mathbb{V}_{\mathcal{T}}^{-1} \subset [\mathbb{S}_{\mathcal{T}}^{n', -1}]^d.$$

The above condition is satisfied when  $n' \geq n - 1$  for subdivisions  $\mathcal{T}$  made of simplices and  $n' \geq n$  for hexahedra. To continue with an analysis incorporating both case, we set  $n' = n$  and write  $\mathcal{L}_{\mathcal{T}} := \mathcal{L}_{\mathcal{T}}^n$ . With this choice, the bilinear form  $\mathcal{B}_{\mathcal{T}}$  in the symmetric interior penalty method (9.21) reads

$$\begin{aligned} \mathcal{B}_{\mathcal{T}}[w, v] &= a_{\mathcal{T}}[w, v] - \int_{\Omega} \mathcal{L}_{\mathcal{T}}[v] \cdot \widehat{\mathbf{A}} \nabla_{\mathcal{T}} w \\ &\quad - \int_{\Omega} \mathcal{L}_{\mathcal{T}}[w] \cdot \widehat{\mathbf{A}} \nabla_{\mathcal{T}} v + \kappa \sum_{F \in \mathcal{F}^+} \int_F h_F^{-1} [[w]] [[v]], \end{aligned} \quad (9.29)$$

for all  $w, v \in \mathbb{V}_{\mathcal{T}}^{-1}$  and where we used

$$a_{\mathcal{T}}[w, v] := \int_{\Omega} \nabla_{\mathcal{T}} v \cdot \widehat{\mathbf{A}} \nabla_{\mathcal{T}} w + \widehat{c} w v \quad (9.30)$$

to denote the bilinear form related to the conforming method.

Expression (9.29) is well defined for  $w, v \in \mathbb{E}_{\mathcal{T}}$  and the weak formulation (9.21) is well-posed. These two claims follow from Corollary 9.8 below, which in turn is a consequence of the next result focusing on the bilinear form  $a_{\mathcal{T}}$ ; we recall (5.52).

**Lemma 9.7 (properties of  $a_{\mathcal{T}}$ )** *Let  $\mathcal{T} \in \mathbb{T}^{\Lambda}$  be a  $\Lambda$ -admissible refinement of  $\mathcal{T}_0$  satisfying Assumption 6.19 (initial labeling). Furthermore, assume that  $\widehat{\mathbf{A}}$  and  $\widehat{c}$  satisfy the structural assumption (9.20). Then, we have*

$$a_{\mathcal{T}}[w, v] \leq (\widehat{\alpha}_2 + |\widehat{c}_2| C_P^2) \|v\|_{1, \mathcal{T}} \|w\|_{1, \mathcal{T}}, \quad \forall v, w \in \mathbb{E}_{\mathcal{T}} \quad (9.31)$$

and

$$a_{\mathcal{T}}[v, v] \geq \frac{\widehat{\alpha}_1}{2} \|\nabla_{\mathcal{T}} v\|_{L^2(\Omega)}^2 + \min(0, \widehat{c}_1) C_P^2 \|h^{-1/2} [[v]]\|_{L^2(\mathcal{F}^+)}^2, \quad \forall v \in \mathbb{E}_{\mathcal{T}}, \quad (9.32)$$

where  $C_P$  is the constant in Lemma 9.3 (Poincaré-type inequality in  $\mathbb{E}_{\mathcal{T}}$ ).

*Proof.* We start with the continuity estimate (9.31). The assumption on the discretized coefficients implies that for  $v, w \in \mathbb{E}_{\mathcal{T}}$  there holds

$$a_{\mathcal{T}}[w, v] \leq \alpha_2 \|\nabla_{\mathcal{T}} w\|_{L^2(\mathcal{T})} \|\nabla_{\mathcal{T}} v\|_{L^2(\mathcal{T})} + |c_2| \|w\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)}.$$

It remains to invoke Lemma 9.3 (Poincaré-type inequality on  $\mathbb{E}_{\mathcal{T}}$ ) to deduce (9.31).

Similarly for the partial coercivity estimate (9.32), we have

$$a_{\mathcal{T}}[v, v] \geq \widehat{\alpha}_1 \|\nabla_{\mathcal{T}} v\|_{L^2(\Omega)}^2 + \widehat{c}_1 \|v\|_{L^2(\Omega)}^2 \geq \widehat{\alpha}_1 \|\nabla_{\mathcal{T}} v\|_{L^2(\Omega)}^2 + \min(0, \widehat{c}_1) C_P^2 \|v\|_{1, \mathcal{T}}^2$$

and the desired estimate follows from the assumption  $-\frac{\widehat{\alpha}_1}{2C_P^2} \leq \widehat{c}_1$ .  $\square$

For the next result, we recall that the discrete norm  $\|\cdot\|_{\kappa, \mathcal{T}}$  is defined in (9.2).

**Corollary 9.8 (properties of  $\mathcal{B}_{\mathcal{T}}$ )** *Let  $\mathcal{T} \in \mathbb{T}^{\Lambda}$  be a  $\Lambda$ -admissible refinement of  $\mathcal{T}_0$  satisfying Assumption 6.19 (initial labeling). Furthermore assume that  $\widehat{\mathbf{A}}$  and  $\widehat{c}$  satisfy the structural assumption (9.20). There exist a constant  $C_{\text{cont}}$  such that*

$$\mathcal{B}_{\mathcal{T}}[w, v] \leq C_{\text{cont}} \|v\|_{\kappa, \mathcal{T}} \|w\|_{\kappa, \mathcal{T}} \quad \forall v, w \in \mathbb{E}_{\mathcal{T}}. \quad (9.33)$$

Moreover, there are constants  $\bar{\kappa}_{\text{stab}}, C_{\text{coer}} > 0$  such that for all  $\kappa > \bar{\kappa}_{\text{stab}}$ , there holds

$$C_{\text{coer}} \|v\|_{\kappa, \mathcal{T}}^2 \leq \mathcal{B}_{\mathcal{T}}[v, v] \quad \forall v \in \mathbb{E}_{\mathcal{T}}. \quad (9.34)$$

In particular, the Galerkin formulation (9.21) has a unique solution  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}^{-1}$ .

*Proof.* The continuity estimate is a direct consequence of the continuity estimate (9.31), estimate (9.26) for the lifting terms, and Cauchy-Schwarz inequality

$$\kappa \sum_{F \in \mathcal{F}^+} \int_F h_F^{-1} [[w]] [[v]] \leq \kappa \|h^{-1/2} [[w]]\|_{L^2(\mathcal{F}^+)} \|h^{-1/2} [[v]]\|_{L^2(\mathcal{F}^+)},$$

which holds for all  $v, w \in \mathbb{E}_{\mathcal{T}}$ .

We now focus on the coercivity estimate (9.34) and start from (9.32), which we write for  $v \in \mathbb{E}_{\mathcal{T}}$  as

$$\frac{\widehat{\alpha}_1}{2} \|\nabla_{\mathcal{T}} v\|_{L^2(\mathcal{T})}^2 - \max(0, -c_1 \widehat{C}_P^2) \|h^{-1/2} [[v]]\|_{L^2(\mathcal{F}^+)}^2 \leq a_{\mathcal{T}}[v, v]. \quad (9.35)$$

Furthermore, the terms involving the lifting operators in the definition (9.29) of the bilinear form  $\mathcal{B}_{\mathcal{T}}$  reduce to  $-2 \int_{\Omega} \mathcal{L}_{\mathcal{T}}[v] \cdot \widehat{\mathbf{A}} \nabla_{\mathcal{T}} v$  when  $w = v$ . Hence, the estimate (9.27) with  $\epsilon = \widehat{\alpha}_1/4$  implies that

$$2 \left| \int_{\Omega} \mathcal{L}_{\mathcal{T}}[v] \cdot \widehat{\mathbf{A}} \nabla_{\mathcal{T}} v \right| \leq \frac{\widehat{\alpha}_1}{4} \|\nabla_{\mathcal{T}} v\|_{L^2(\mathcal{T})}^2 + \frac{4C_{\mathcal{L}}^2}{\widehat{\alpha}_1} \|h^{-1/2} [[v]]\|_{L^2(\mathcal{F}^+)}^2.$$

Gathering the above inequalities and recalling definition (9.29) of  $\mathcal{B}_{\mathcal{T}}$ , we find that

$$\frac{\alpha_1}{4} \|\nabla_{\mathcal{T}} v\|_{L^2(\mathcal{T})}^2 + (\kappa - \bar{\kappa}_{\text{stab}}) \|h^{-1/2} [[v]]\|_{L^2(\mathcal{F}^*)}^2 \leq \mathcal{B}_{\mathcal{T}}[v, v]$$

with

$$\bar{\kappa}_{\text{stab}} := \frac{4C_{\mathcal{L}}^2}{\alpha_1} + \max(0, -\widehat{c}_1 C_P^2).$$

The desired coercivity estimate directly follows provided  $\kappa > \bar{\kappa}_{\text{stab}}$ .  $\square$

### 9.2.3 Partial consistency and role of the conforming Galerkin solution

From now on, we shall use the expression of  $\mathcal{B}_{\mathcal{T}}$  in (9.29) extending  $\mathcal{B}_{\mathcal{T}}$  to  $\mathbb{E}_{\mathcal{T}} \times \mathbb{E}_{\mathcal{T}}$ . This reformulation comes at the price of a partial consistency. Since  $\mathcal{L}_{\mathcal{T}}[v] = 0$  whenever  $v \in H_0^1(\Omega)$  and the duality product  $\langle \cdot, \cdot \rangle_{\mathcal{T}}$  satisfies the consistency (9.4), we have

$$\mathcal{B}_{\mathcal{T}}[\widehat{u}, v] = a_{\mathcal{T}}[\widehat{u}, v] = \langle \widehat{f}, v \rangle_{\mathcal{T}} \quad \forall v \in H_0^1(\Omega), \quad (9.36)$$

which indicates that the reformulation (9.29) using lifts is consistent on  $H_0^1(\Omega)$ . However, (9.36) does not hold for all  $v \in \mathbb{V}_{\mathcal{T}}^{-1}$ .

This suggests splitting  $\mathbb{V}_{\mathcal{T}}^{-1}$  into a conforming space where the consistency holds and its orthogonal complement. We decompose the discontinuous space as

$$\mathbb{V}_{\mathcal{T}}^{-1} = \mathbb{V}_{\mathcal{T}}^0 \oplus \mathbb{V}_{\mathcal{T}}^{\perp}, \quad (9.37)$$

where  $\mathbb{V}_{\mathcal{T}}^0 = \mathbb{V}_{\mathcal{T}}^{-1} \cap H_0^1(\Omega)$  is the finest conforming subspace of  $\mathbb{V}_{\mathcal{T}}^{-1}$  and  $\mathbb{V}_{\mathcal{T}}^{\perp}$  is the orthogonal complement with respect to the  $\mathcal{B}_{\mathcal{T}}[\cdot, \cdot]$  scalar product. Note that the later is well defined provided the assumption on the penalty parameter  $\kappa > \bar{\kappa}_{\text{stab}}$ , required by Corollary 9.8, is satisfied. From now on, we assume this is the case and point out that although the constants appearing in the analysis below do not depend on  $\kappa$ , they may depend on  $\bar{\kappa}_{\text{stab}}$ .

We also emphasize that there might not be a conforming subdivision associated with  $\mathbb{V}_{\mathcal{T}}^0$ . The latter is the span of the basis functions associated with proper nodes; see Figure 3.11 for an illustration and refer to Section 8 for more details. Consequently, the analysis provided below relies on the decomposition (9.37) of the space  $\mathbb{V}_{\mathcal{T}}$  rather than on a subdivision  $\mathcal{T}$ . It is also worth pointing out that the conforming part  $u_{\mathcal{T}}^0 \in \mathbb{V}_{\mathcal{T}}^0$  of the Galerkin solution  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}^{-1}$  satisfies

$$\mathcal{B}_{\mathcal{T}}[u_{\mathcal{T}}^0, v] = \langle \widehat{f}, v \rangle, \quad \forall v \in \mathbb{V}_{\mathcal{T}}^0. \quad (9.38)$$

Hence,  $u_{\mathcal{T}}^0$  is the conforming Galerkin approximation on  $\mathbb{V}_{\mathcal{T}}^0$ . As we shall see this finest coarser conforming Galerkin solution plays a critical role in the convergence of AFEM-DG-TS. This justifies the orthogonal decomposition (9.37) associated with the  $\mathcal{B}_{\mathcal{T}}$  scalar product.

Another advantage of using the  $\mathcal{B}_{\mathcal{T}}$ -orthogonal decomposition (9.37) is that it offers a control on the non-conforming component of  $v \in \mathbb{V}_{\mathcal{T}}^{-1}$  by its scaled jumps. To achieve this, the operator  $\mathcal{I}_{\mathcal{T}}^{\text{dG}}$  defined by (9.7) is instrumental.

**Lemma 9.9 (control of non-conformity)** *Let  $\mathcal{T} \in \mathbb{T}^\Lambda$  be a  $\Lambda$ -admissible refinement of  $\mathcal{T}_0$  satisfying Assumption 6.19 (initial labeling). Assume that  $\widehat{\mathbf{A}}$  and  $\widehat{\mathbf{c}}$  satisfy the structural assumption (9.20). For  $\kappa > \bar{\kappa}_{\text{stab}}$ , if  $v = v^0 + v^\perp \in \mathbb{V}_{\mathcal{T}}^{-1}$  according to (9.37), then*

$$\|v^\perp\|_{\kappa, \mathcal{T}} \lesssim \kappa^{1/2} \|h^{-1/2} \llbracket v^\perp \rrbracket\|_{L^2(\mathcal{F}^+)} = \kappa^{1/2} \|h^{-1/2} \llbracket v \rrbracket\|_{L^2(\mathcal{F}^+)}.$$

*Proof.* Because  $\mathcal{I}_{\mathcal{T}}^{\text{dG}} v \in \mathbb{V}_{\mathcal{T}}^0$ , the orthogonal decomposition (9.37) implies that

$$\mathcal{B}_{\mathcal{T}}[v^\perp, v^\perp] \leq \mathcal{B}_{\mathcal{T}}[v - \mathcal{I}_{\mathcal{T}}^{\text{dG}} v, v - \mathcal{I}_{\mathcal{T}}^{\text{dG}} v].$$

The desired result follows from the coercivity (9.34) and continuity (9.33) of  $\mathcal{B}_{\mathcal{T}}$  along with the interpolation estimate (9.10).  $\square$

### 9.3 A posteriori error estimates

We derive a residual error estimate for the discontinuous Galerkin method. Because the data  $\widehat{\mathcal{D}} \in \mathbb{D}_{\mathcal{T}}$  is discrete, the analysis is free from data oscillation. In the notation introduced in Section 4, this means  $\mathcal{E}_{\mathcal{T}} = \eta_{\mathcal{T}}$  where for  $v \in \mathbb{E}_{\mathcal{T}}$

$$\eta_{\mathcal{T}}^2(v) := \sum_{T \in \mathcal{T}} \eta_{\mathcal{T}}(v, T)^2,$$

and

$$\eta_{\mathcal{T}}(v, T)^2 := h_T \sum_{F \subset \partial T \setminus \partial \Omega} \|j_{\mathcal{T}}(v) - \widehat{f}\|_{L^2(F)}^2 + h_T^2 \|r_{\mathcal{T}}(v)\|_{L^2(T)}^2$$

with  $j_{\mathcal{T}}(v)|_F := \mathbf{n}_F \cdot \llbracket \widehat{\mathbf{A}} \nabla_{\mathcal{T}} v \rrbracket$  and  $r_{\mathcal{T}}(v)|_T := \widehat{f} - \widehat{\mathbf{c}}v + \text{div}_{\mathcal{T}}(\widehat{\mathbf{A}} \nabla_{\mathcal{T}} v)$ .

We start with a result mimicking the conforming argument, discuss its drawback and improvement next.

**Lemma 9.10 (a posteriori error estimates)** *Let  $\mathcal{T} \in \mathbb{T}^\Lambda$  be a  $\Lambda$ -admissible refinement of  $\mathcal{T}_0$  satisfying Assumption 6.19 (initial labeling). Assume that  $\widehat{\mathbf{A}}$  and  $\widehat{\mathbf{c}}$  satisfy the structural assumption (9.20). If  $\kappa > \bar{\kappa}_{\text{stab}}$ , then there holds*

$$\|\widehat{u} - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}}^2 \lesssim \eta_{\mathcal{T}}(u_{\mathcal{T}})^2 + \kappa \|h^{-1/2} \llbracket u_{\mathcal{T}} \rrbracket\|_{L^2(\mathcal{F}^+)}^2 \quad (9.39)$$

and

$$C_L \eta_{\mathcal{T}}(u_{\mathcal{T}}) \leq \|\widehat{u} - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}}, \quad (9.40)$$

for some constant  $C_L$ .

*Proof.* We start with the upper bound (9.39). To exploit the consistency (9.36) in  $\mathbb{V}_{\mathcal{T}}^0$ , we decompose the error  $e := \widehat{u} - u_{\mathcal{T}} \in \mathbb{E}_{\mathcal{T}}$  into a conforming part  $e^0 := \widehat{u} - u_{\mathcal{T}}^0 \in H_0^1(\Omega)$  and a non-conforming part  $e^\perp := -u_{\mathcal{T}}^\perp \in \mathbb{V}_{\mathcal{T}}^{-1}$  according to (9.37). The proof thus relies on techniques used in the conforming theory coupled with Lemma 9.9 (control of non-conformity). We denote by  $C$  a generic constant independent of the discretization and  $\kappa$  but possibly depending on  $\bar{\kappa}_{\text{stab}}$ .

Using the coercivity (9.34) and partial consistency (9.36), we get

$$C_{\text{coer}} \|e\|_{\kappa, \mathcal{T}}^2 \leq \mathcal{B}_{\mathcal{T}}[e, e] = \mathcal{B}_{\mathcal{T}}[e, e^0 - I_{\mathcal{T}}e^0] - \mathcal{B}_{\mathcal{T}}[e, u_{\mathcal{T}}^{\perp}], \quad (9.41)$$

where  $I_{\mathcal{T}}$  is the Scott-Zhang interpolant provided in Proposition 3.5. For the first term, we note that since  $e^0 - I_{\mathcal{T}}e^0 \in H_0^1(\Omega)$  we have

$$\mathcal{B}_{\mathcal{T}}[e, e^0 - I_{\mathcal{T}}e^0] = a_{\mathcal{T}}[e, e^0 - I_{\mathcal{T}}e^0] - \int_{\Omega} \mathcal{L}_{\mathcal{T}}[e] \cdot \mathbf{A} \nabla(e^0 - I_{\mathcal{T}}e^0).$$

For the term involving the bilinear form  $a_{\mathcal{T}}$  we proceed as in the conforming case to arrive at

$$a_{\mathcal{T}}[e, e^0 - I_{\mathcal{T}}e^0] \lesssim \eta_{\mathcal{T}}(u_{\mathcal{T}}) \|\nabla e^0\|_{L^2(\Omega)}.$$

This, combined with estimate (9.26) on the lifting operators and the  $H^1$  stability of the Scott-Zhang interpolant, yield

$$\mathcal{B}_{\mathcal{T}}[e, e^0 - I_{\mathcal{T}}e^0] \lesssim \left( \eta_{\mathcal{T}}(u_{\mathcal{T}})^2 + \|h^{-1/2} \llbracket u_{\mathcal{T}} \rrbracket \|_{L^2(\mathcal{F}^+)}^2 \right)^{1/2} \|\nabla e^0\|_{L^2(\Omega)}.$$

We rewrite  $e^0 = e + u_{\mathcal{T}}^{\perp}$  and use the estimate on the non-conforming component provided by Lemma 9.9 along with a Young inequality to write

$$\mathcal{B}_{\mathcal{T}}[e, e^0 - I_{\mathcal{T}}e^0] \leq C \left( \eta_{\mathcal{T}}(u_{\mathcal{T}})^2 + \kappa \|h^{-1/2} \llbracket u_{\mathcal{T}} \rrbracket \|_{L^2(\mathcal{F}^+)}^2 \right) + \frac{C_{\text{coer}}}{4} \|\nabla e\|_{L^2(\Omega)}^2.$$

For the second term in (9.41), the continuity (9.33) of the bilinear form  $\mathcal{B}_{\mathcal{T}}$ , Lemma 9.9 again, and a Young inequality yield

$$\mathcal{B}_{\mathcal{T}}[e, u_{\mathcal{T}}^{\perp}] \leq \frac{C_{\text{coer}}}{4} \|e\|_{\kappa, \mathcal{T}}^2 + C\kappa \|h^{-1/2} \llbracket u_{\mathcal{T}} \rrbracket \|_{L^2(\mathcal{F}^+)}^2.$$

Returning to (9.41), we find that

$$\|e\|_{\mathcal{T}}^2 \lesssim \eta_{\mathcal{T}}(u_{\mathcal{T}})^2 + \kappa \|h^{-1/2} \llbracket u_{\mathcal{T}} \rrbracket \|_{L^2(\mathcal{F}^+)}^2,$$

which is the desired upperbound.

We finally deal with the lower bound (9.40). For  $T \in \mathcal{T}$  and  $v \in H_0^1(T)$ , we get

$$\int_T (-\text{div}_{\mathcal{T}}(\widehat{\mathbf{A}} \nabla_{\mathcal{T}} u_{\mathcal{T}}) + \widehat{c} u_{\mathcal{T}} - \widehat{f}) v = \int_T \nabla v \cdot \widehat{\mathbf{A}} \nabla_{\mathcal{T}}(\widehat{u} - u_{\mathcal{T}}) v + \widehat{c}(\widehat{u} - u_{\mathcal{T}}) v.$$

For an interior face  $F \in \mathcal{F}$ ,  $v \in H_0^1(\omega_F)$  with  $\omega_F := \{T \in \mathcal{T} : T \cap F \neq \emptyset\}$ , we have

$$\begin{aligned} \int_F (\llbracket \widehat{\mathbf{A}} \nabla_{\mathcal{T}} u_{\mathcal{T}} \rrbracket - \widehat{f}) v &= \int_{\omega_F} (-\text{div}_{\mathcal{T}}(\widehat{\mathbf{A}} \nabla_{\mathcal{T}} u_{\mathcal{T}}) + \widehat{c} u_{\mathcal{T}} - \widehat{f}) v \\ &\quad - \int_{\omega_F} \nabla v \cdot \widehat{\mathbf{A}} \nabla_{\mathcal{T}}(\widehat{u} - u_{\mathcal{T}}) - \widehat{c}(\widehat{u} - u_{\mathcal{T}}) v. \end{aligned}$$

The desired lower bound follows from the same arguments as in the conforming case; we refer to Proposition 4.12 (partial lower bound).  $\square$

Upper bound (9.39) may suggest adding the jump term  $\kappa^{1/2} \|h^{-1/2} [[u_{\mathcal{T}}]] \|_{L^2(\mathcal{F}^+)}$  to the residual estimator  $\eta_{\mathcal{T}}(u_{\mathcal{T}})$ . This would result in a clean upper bound but, because of the presence of negative powers of the meshsize, would be at the expense of destroying the monotonicity property of the estimator; see e.g. Proposition 4.56 (estimator reduction). The later is instrumental in the analysis provided below.

The next result mitigates the effect of the additional jump term by showing that  $\|h^{-1/2} [[u_{\mathcal{T}}]] \|_{L^2(\mathcal{F}^+)}$  can be bounded by  $\eta_{\mathcal{T}}(u_{\mathcal{T}})/\kappa$  and thus can be absorbed by the estimator in the upper bound provided  $\kappa$  is sufficiently large. We follow the proof provided in [Bonito and Nochetto 2010] and refer to [Karakashian and Pascal 2007, eq. (3.20)] for an alternative (original) proof.

**Lemma 9.11 (discontinuity control)** *Let  $\mathcal{T} \in \mathbb{T}^{\Lambda}$  be a  $\Lambda$ -admissible refinement of  $\mathcal{T}_0$  satisfying Assumption 6.19 (initial labeling). Assume that  $\hat{\mathbf{A}}$  and  $\hat{\mathbf{c}}$  satisfy the structural assumption (9.20). There exists a constant  $\bar{\kappa}_{\text{jump}} \geq \bar{\kappa}_{\text{stab}} > 0$  such that if  $\kappa \geq \bar{\kappa}_{\text{jump}}$  then there holds*

$$\|h^{-1/2} [[u_{\mathcal{T}}]] \|_{L^2(\mathcal{F}^+)} \lesssim \kappa^{-1} \eta_{\mathcal{T}}(u_{\mathcal{T}}).$$

*Proof.* For  $v_{\mathcal{T}}^0 \in \mathbb{V}_{\mathcal{T}}^0$  we realize that because  $[[v_{\mathcal{T}}^0]] = 0$ , the coercivity estimate (9.34) implies that

$$C_{\text{coerc}} \kappa \|h^{-1/2} [[u_{\mathcal{T}}]] \|_{L^2(\mathcal{F}^+)}^2 \leq \mathcal{B}_{\mathcal{T}}[u_{\mathcal{T}} - v_{\mathcal{T}}^0, u_{\mathcal{T}} - v_{\mathcal{T}}^0]. \quad (9.42)$$

We now rewrite the right-hand side of (9.42) to produce residual terms. Since  $u_{\mathcal{T}}$  solves (9.21), we have

$$\mathcal{B}_{\mathcal{T}}[u_{\mathcal{T}} - v_{\mathcal{T}}^0, u_{\mathcal{T}} - v_{\mathcal{T}}^0] = \langle \hat{f}, u_{\mathcal{T}} - v_{\mathcal{T}}^0 \rangle_{\mathcal{T}} - \mathcal{B}_{\mathcal{T}}[v_{\mathcal{T}}^0, u_{\mathcal{T}} - v_{\mathcal{T}}^0]. \quad (9.43)$$

We concentrate for the moment on the second term. Since  $[[v_{\mathcal{T}}^0]] = 0$ , the stabilization term vanishes as well:

$$\kappa \sum_{F \in \mathcal{F}^+} \int_F [[v_{\mathcal{T}}^0]] [[u_{\mathcal{T}} - v_{\mathcal{T}}^0]] = 0.$$

Hence, writing  $v_{\mathcal{T}}^0 = u_{\mathcal{T}} + (v_{\mathcal{T}}^0 - u_{\mathcal{T}})$ , we deduce that

$$\begin{aligned} \mathcal{B}_{\mathcal{T}}[v_{\mathcal{T}}^0, u_{\mathcal{T}} - v_{\mathcal{T}}^0] &= a_{\mathcal{T}}[u_{\mathcal{T}}, u_{\mathcal{T}} - v_{\mathcal{T}}^0] \\ &\quad - a_{\mathcal{T}}[u_{\mathcal{T}} - v_{\mathcal{T}}^0, u_{\mathcal{T}} - v_{\mathcal{T}}^0] - \int_{\Omega} \nabla v_{\mathcal{T}}^0 \cdot \hat{\mathbf{A}} \mathcal{L}_{\mathcal{T}}[u_{\mathcal{T}}], \end{aligned}$$

where we invoked again the property  $[[v_{\mathcal{T}}^0]] = 0$  to infer that  $\mathcal{L}_{\mathcal{T}}[v_{\mathcal{T}}^0] = 0$ . Integrating by parts the first term on the right-hand side, adding it to the first term on the right-hand side of (9.43), and using the extended definition (9.3) of the duality



pairing leads to the following expression involving the residuals  $r_{\mathcal{T}}(u_{\mathcal{T}})$ ,  $j_{\mathcal{T}}(u_{\mathcal{T}})$ :

$$\begin{aligned} \mathcal{B}_{\mathcal{T}}[u_{\mathcal{T}} - v_{\mathcal{T}}^0, u_{\mathcal{T}} - v_{\mathcal{T}}^0] &= \int_{\Omega} r_{\mathcal{T}}(u_{\mathcal{T}})(u_{\mathcal{T}} - v_{\mathcal{T}}^0) + \int_{\mathcal{F}^+} (j_{\mathcal{T}}(u_{\mathcal{T}}) - \widehat{f}) \{u_{\mathcal{T}} - v_{\mathcal{T}}^0\} \\ &\quad + a_{\mathcal{T}}[u_{\mathcal{T}} - v_{\mathcal{T}}^0, u_{\mathcal{T}} - v_{\mathcal{T}}^0] - \int_{\Omega} \nabla(u_{\mathcal{T}} - v_{\mathcal{T}}^0) \cdot \widehat{A} \mathcal{L}_{\mathcal{T}}[u_{\mathcal{T}}]. \end{aligned}$$

We point out that we have also employed the definition (9.24) of lift to rewrite the resulting face terms. Inserting this estimate back in (9.42), together with the bound (9.26) for lifts and the continuity estimate (9.31) of  $a_{\mathcal{T}}$ , gives

$$\begin{aligned} \kappa \|h^{-1/2} [[u_{\mathcal{T}}]]\|_{L^2(\mathcal{F}^+)}^2 &\lesssim \|u_{\mathcal{T}} - v_{\mathcal{T}}^0\|_{1,\mathcal{T}}^2 \\ &\quad + \eta_{\mathcal{T}}(u_{\mathcal{T}}) \left( \|h^{-1}(u_{\mathcal{T}} - v_{\mathcal{T}}^0)\|_{L^2(\Omega)} + \|h^{-1/2} \{u_{\mathcal{T}} - v_{\mathcal{T}}^0\}\|_{L^2(\mathcal{F}^+)} \right). \end{aligned}$$

Note that the presence of  $\|\cdot\|_{1,\mathcal{T}}$  rather than  $\|\cdot\|_{\kappa,\mathcal{T}}$  on the right-hand side of the above estimate is critical for the argument below. The former is independent of  $\kappa$  and can thus be absorbed on the left-hand side for sufficiently large  $\kappa$  provided  $v_{\mathcal{T}}^0 = \mathcal{I}_{\mathcal{T}}^{\text{dG}} u_{\mathcal{T}}$ . In fact, the interpolation estimates (9.15) in turn imply

$$\begin{aligned} \|h^{-1}(u_{\mathcal{T}} - \mathcal{I}_{\mathcal{T}}^{\text{dG}} u_{\mathcal{T}})\|_{L^2(\Omega)} &+ \|h^{-1/2} \{u_{\mathcal{T}} - \mathcal{I}_{\mathcal{T}}^{\text{dG}} u_{\mathcal{T}}\}\|_{L^2(\mathcal{F}^+)} \\ &+ \|u_{\mathcal{T}} - \mathcal{I}_{\mathcal{T}}^{\text{dG}} u_{\mathcal{T}}\|_{1,\mathcal{T}} \lesssim \|h^{-1/2} [[u_{\mathcal{T}}]]\|_{L^2(\mathcal{F}^+)}. \end{aligned}$$

Hence, a Young's inequality yields

$$(\kappa - \bar{\kappa}_{\text{stab}}) \|h^{-1/2} [[u_{\mathcal{T}}]]\|_{L^2(\mathcal{F}^+)}^2 \lesssim \kappa^{-1} \eta_{\mathcal{T}}(u_{\mathcal{T}})^2 + \|h^{-1/2} [[u_{\mathcal{T}}]]\|_{L^2(\mathcal{F}^+)}^2,$$

and the desired estimate follows provided  $\kappa$  is sufficiently large.  $\square$

As a direct consequence of the previous Lemma, we obtain a simpler practical upper bound.

**Corollary 9.12 (stabilization-free a posteriori upper bound)** *Under the assumptions of Lemma 9.11, there exists a constant  $C_U$  such that for all  $\kappa \geq \bar{\kappa}_{\text{jump}}$  we have*

$$\|\widehat{u} - u_{\mathcal{T}}\|_{\kappa,\mathcal{T}} \leq C_U \eta_{\mathcal{T}}(u_{\mathcal{T}}). \quad (9.44)$$

*Proof.* Combine the upper bound (9.39) and Lemma 9.11.  $\square$

The partial consistency (9.36) leads to *partial Galerkin orthogonality*

$$\mathcal{B}_{\mathcal{T}}[\widehat{u} - u_{\mathcal{T}}, v] = 0 \quad \forall v \in \mathbb{V}_{\mathcal{T}}^0. \quad (9.45)$$

This would suggest that a quasi-best approximation (Cea's lemma) result in the full space  $\mathbb{V}_{\mathcal{T}}^{-1}$  is questionable. However, the lack of consistency is built into the jump terms which are in turn controlled by the estimator weighted by a negative power of the penalty parameter  $\kappa$ . It thus remains to resort to the lower bound to return to the error and derive a quasi-best approximation estimate for sufficiently large  $\kappa$ . We prove this result next, which expresses the important fact that dG is quasi-optimal with respect to the norm  $\|\cdot\|_{\kappa,\mathcal{T}}$  defined in (9.2). This has two significant

consequences: first it leads to quasi-monotonicity of the error upon refinement (see Corollary 9.14 below), and second it dictates the approximation class for dG already alluded to in Proposition 9.4 (equivalence of classes for  $u$ ).

**Corollary 9.13 (Cea's lemma)** *Under the assumptions of Lemma 9.11, there is  $\bar{\kappa}_{Cea} \geq \bar{\kappa}_{jump}$  such that for  $\kappa \geq \bar{\kappa}_{Cea}$ , there holds*

$$\|\widehat{u} - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}} \leq C_{Cea} \inf_{v_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}^{-1}} \|\widehat{u} - v_{\mathcal{T}}\|_{\kappa, \mathcal{T}}. \quad (9.46)$$

*Proof.* We combine the orthogonal decomposition (9.37) and the partial Galerkin orthogonality (9.45) to write

$$\mathcal{B}_{\mathcal{T}}[\widehat{u} - u_{\mathcal{T}}, \widehat{u} - u_{\mathcal{T}}] = \mathcal{B}_{\mathcal{T}}[\widehat{u} - u_{\mathcal{T}}, \widehat{u} - v_{\mathcal{T}}] - \mathcal{B}_{\mathcal{T}}[\widehat{u} - u_{\mathcal{T}}, u_{\mathcal{T}}^{\perp}] + \mathcal{B}_{\mathcal{T}}[\widehat{u} - u_{\mathcal{T}}, v_{\mathcal{T}}^{\perp}]$$

for all  $v_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}$ . Invoking the coercivity and continuity of  $\mathcal{B}_{\mathcal{T}}$  in Lemma 9.8) (properties of  $\mathcal{B}_{\mathcal{T}}$ ) in conjunction with Lemma 9.9 (control of discontinuity) yields

$$\|\widehat{u} - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}} \lesssim \|\widehat{u} - v_{\mathcal{T}}\|_{\kappa, \mathcal{T}} + \kappa^{1/2} \|h^{-1/2} [[u_{\mathcal{T}}]]\|_{L^2(\mathcal{F}^+)} + \kappa^{1/2} \|h^{-1/2} [[v_{\mathcal{T}}]]\|_{L^2(\mathcal{F}^+)}.$$

Applying now Lemma 9.11 (discontinuity control) results in

$$\|h^{-1/2} [[u_{\mathcal{T}}]]\|_{L^2(\mathcal{F}^+)} \lesssim \kappa^{-1} \eta_{\mathcal{T}}(u_{\mathcal{T}}) \lesssim \kappa^{-1} \|\widehat{u} - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}}$$

because of the lower bound (9.40). We thus end up with

$$\|\widehat{u} - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}} \lesssim \|\widehat{u} - v_{\mathcal{T}}\|_{\kappa, \mathcal{T}} + \kappa^{-1/2} \|\widehat{u} - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}},$$

which for  $\kappa$  sufficiently large gives the desired bound.  $\square$

With this best approximation result, we deduce the following crucial property.

**Corollary 9.14 (quasi-monotonicity)** *Under the assumptions of Lemma 9.11, there is a constant  $C_{Mo}$  independent of the discretization parameters and  $\kappa$  such that for all  $\kappa \geq \bar{\kappa}_{jump}$  and all  $\mathcal{T}_* \geq \mathcal{T}$  we have*

$$\|\widehat{u} - u_{\mathcal{T}_*}\|_{\kappa, \mathcal{T}_*} \leq C_{Mo} \|\widehat{u} - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}}. \quad (9.47)$$

*Proof.* We rely on the orthogonal decomposition (9.37) to write  $u_{\mathcal{T}} = u_{\mathcal{T}}^0 + u_{\mathcal{T}}^{\perp}$  and on Corollary 9.13 (Cea's lemma). Since  $u_{\mathcal{T}}^0 \in \mathbb{V}_{\mathcal{T}}^0 \subset \mathbb{V}_{\mathcal{T}_*}^{-1}$ , we see that

$$C_{Cea}^{-1} \|\widehat{u} - u_{\mathcal{T}_*}\|_{\kappa, \mathcal{T}_*} \leq \|\widehat{u} - u_{\mathcal{T}}^0\|_{\kappa, \mathcal{T}_*} = |\widehat{u} - u_{\mathcal{T}}^0|_{H_0^1(\Omega)} = \|\widehat{u} - u_{\mathcal{T}}^0\|_{\kappa, \mathcal{T}}.$$

Therefore, adding and subtracting  $u_{\mathcal{T}}^{\perp}$  and making use of Lemma 9.9 (control of non-conformity) together with Lemma 9.11 (discontinuity control) implies

$$\|\widehat{u} - u_{\mathcal{T}_*}\|_{\kappa, \mathcal{T}_*} \lesssim \|\widehat{u} - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}} + \kappa^{-1/2} \eta_{\mathcal{T}}(u_{\mathcal{T}}).$$

It remains to invoke the lower bound (9.40) to deduce the desired result.  $\square$

Corollary 9.14 assumes the same data. In estimating the cost of GALERKIN-DG we need a variant of this result that allows for different data. We establish this next.

**Corollary 9.15 (quasi-monotonicity with different data)** *Let  $\mathcal{T}_* \geq \mathcal{T}$  and  $\widehat{\mathcal{D}}_*, \widehat{\mathcal{D}}$  be discrete data on these meshes. Let  $\widehat{u}_* = u(\widehat{\mathcal{D}}_*)$ ,  $\widehat{u} = u(\widehat{\mathcal{D}}) \in H_0^1(\Omega)$  and  $u_{\mathcal{T}_*} \in \mathbb{V}_{\mathcal{T}_*}^{-1}$ ,  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}^{-1}$  be the corresponding exact and Galerkin solutions. Under the assumptions of Lemma 9.11, for  $\kappa \geq \bar{\kappa}_{\text{jump}}$  there holds*

$$\|\widehat{u}_* - u_{\mathcal{T}_*}\|_{\kappa, \mathcal{T}_*} \leq C_{Mo}(\|\widehat{u} - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}} + |\widehat{u}_* - \widehat{u}|_{H_0^1(\Omega)}). \quad (9.48)$$

*Proof.* We proceed as in the proof of Corollary 9.14 with  $\widehat{u}_*$ , but in the last step use the fact  $u_{\mathcal{T}}$  and  $\widehat{u}$  are the functions associated with the same data  $\widehat{\mathcal{D}}$  and thereby satisfy  $C_L \eta_{\mathcal{T}}(u_{\mathcal{T}}) \leq \|\widehat{u} - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}}$  according to (9.40). Applying the triangle inequality and the property  $\|\widehat{u}_* - \widehat{u}\|_{\kappa, \mathcal{T}} = |\widehat{u}_* - \widehat{u}|_{H_0^1(\Omega)}$  concludes the proof.  $\square$

We end this section with the dG counterpart of Theorem 4.48 (upper bound for corrections). One striking difference is that the lack of consistency prevents the discrete lower bound in the dG context from localizing to the refined set  $\mathcal{T} \setminus \mathcal{T}_*$  for  $\mathcal{T}_* \geq \mathcal{T}$ . Rather, it contains a global jump term that expresses the lack of conformity and vanishes as  $\kappa \rightarrow \infty$  in view of Lemma 9.11 (discontinuity control). This is consistent with the upper bound (9.39). We use the notation  $\omega_{\mathcal{T}}(\tau)$  for a set of elements  $\tau \in \mathcal{T}$  to denote  $\tau$  augmented by one layer of elements

$$\omega(\tau) := \omega_{\mathcal{T}}(\tau) = \bigcup_{T \in \tau} \omega_{\mathcal{T}}(T).$$

**Lemma 9.16 (quasi-localized discrete upper bound)** *Let  $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}^\Lambda$ , with  $\mathcal{T}_* \geq \mathcal{T}$ , be two  $\Lambda$ -admissible refinements of  $\mathcal{T}_0$  satisfying Assumption 6.19 (initial labeling). Assume that  $\widehat{\mathbf{A}}$  and  $\widehat{c}$  satisfy the structural assumption (9.20),  $\widehat{f} \in \mathbb{F}_{\mathcal{T}}$ , and denote by  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}^{-1}$ ,  $u_{\mathcal{T}_*} \in \mathbb{V}_{\mathcal{T}_*}^{-1}$  the two Galerkin solutions associated with  $\mathcal{T}$ ,  $\mathcal{T}_*$  respectively. There is a constant  $C_{LU}$  such that for all  $\kappa > \bar{\kappa}_{\text{stab}}$  we have*

$$\|u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}}^2 \leq C_{LU}^2 \left( \eta_{\mathcal{T}}^2(u_{\mathcal{T}}, \omega(\mathcal{T} \setminus \mathcal{T}_*)) + \kappa \|h^{-1/2} \llbracket u_{\mathcal{T}} \rrbracket \|_{L^2(\mathcal{F}^+)}^2 \right),$$

where  $u_{\mathcal{T}_*} = u_{\mathcal{T}_*}^0 + u_{\mathcal{T}_*}^\perp$  is the orthogonal decomposition according to (9.37). Moreover, if  $\kappa \geq \bar{\kappa}_{\text{jump}}$ , there holds

$$\|u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}}^2 \leq C_{LU}^2 \left( \eta_{\mathcal{T}}^2(u_{\mathcal{T}}, \omega(\mathcal{T} \setminus \mathcal{T}_*)) + \kappa^{-1} \eta_{\mathcal{T}}^2(u_{\mathcal{T}}) \right). \quad (9.49)$$

*Proof.* We decompose  $u_{\mathcal{T}_*} = u_{\mathcal{T}_*}^0 + u_{\mathcal{T}_*}^\perp$  according to (9.37), exploit the partial consistency (9.36) for  $u_{\mathcal{T}_*}^0$  with  $v^0 \in \mathbb{V}_{\mathcal{T}}^0$ , and  $\widehat{f} \in \mathbb{F}_{\mathcal{T}} \subset \mathbb{F}_{\mathcal{T}_*}$  to obtain

$$\mathcal{B}_{\mathcal{T}_*}[u_{\mathcal{T}_*}, v^0] = a_{\mathcal{T}_*}[u_{\mathcal{T}_*}, v^0] = a_{\mathcal{T}}[u_{\mathcal{T}_*}, v^0] = \mathcal{B}_{\mathcal{T}}[u_{\mathcal{T}_*}, v^0] = \langle \widehat{f}, v^0 \rangle_{\mathcal{T}}.$$

Since  $\mathcal{B}_{\mathcal{T}}[u_{\mathcal{T}}, v^0] = \langle \widehat{f}, v^0 \rangle_{\mathcal{T}}$ , we readily see that

$$\mathcal{B}_{\mathcal{T}}[u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}, v^0] = 0 \quad \forall v^0 \in \mathbb{V}_{\mathcal{T}}^0.$$

We rely on this reduced form of Galerkin orthogonality to prove the assertions. To this end, we write  $u_{\mathcal{T}_*}^0 - u_{\mathcal{T}} = e_*^0 - u_{\mathcal{T}}^\perp$ , with  $e_*^0 := u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}^0$ . Using the coercivity

estimate (9.34) for  $v = u_{\mathcal{T}_*}^0 - u_{\mathcal{T}} \in \mathbb{B}_{\mathcal{T}}$  yields for  $\kappa \geq \bar{\kappa}_{\text{stab}}$

$$\begin{aligned} \|u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}}^2 &\lesssim \mathcal{B}_{\mathcal{T}}[u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}, u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}] \\ &= \mathcal{B}_{\mathcal{T}}[u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}, e_*^0] - \mathcal{B}_{\mathcal{T}}[u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}, u_{\mathcal{T}}^\perp]. \end{aligned} \quad (9.50)$$

Note that the last term cannot be localized and accounts for the lack of consistency of the dG method. However, it can be made arbitrarily small by increasing the penalty parameter  $\kappa$ . In fact, combining the continuity (9.33) with Lemma 9.9 (control of non-conformity) gives

$$\mathcal{B}_{\mathcal{T}}[u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}, u_{\mathcal{T}}^\perp] \lesssim \kappa^{1/2} \|h^{-1/2} [[u_{\mathcal{T}}]]\|_{L^2(\mathcal{F}^+)} \|u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}}.$$

To localize  $\mathcal{B}_{\mathcal{T}}[u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}, e_*^0]$ , we choose  $v^0 = \mathcal{I}_{\mathcal{T}}^{\text{dG}} e_*^0$ , where the interpolation operator  $\mathcal{I}_{\mathcal{T}}^{\text{dG}}$  is given by (9.7), and exploit the reduced Galerkin orthogonality. Since  $e_*^0 - \mathcal{I}_{\mathcal{T}}^{\text{dG}} e_*^0 \in H_0^1(\Omega)$ , the decomposition (9.29) of the bilinear form  $\mathcal{B}_{\mathcal{T}}$  reads

$$\begin{aligned} \mathcal{B}_{\mathcal{T}}[u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}, e_*^0] &= \mathcal{B}_{\mathcal{T}}[u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}, e_*^0 - \mathcal{I}_{\mathcal{T}}^{\text{dG}} e_*^0] \\ &= a_{\mathcal{T}}[u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}, e_*^0 - \mathcal{I}_{\mathcal{T}}^{\text{dG}} e_*^0] - \int_{\Omega} \mathcal{L}_{\mathcal{T}}[u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}] \cdot \widehat{\mathbf{A}} \nabla_{\mathcal{T}}(e_*^0 - \mathcal{I}_{\mathcal{T}}^{\text{dG}} e_*^0). \end{aligned}$$

We handle the first term as in the conforming case (Theorem 4.48), namely

$$a_{\mathcal{T}}(u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}, e_*^0) \lesssim \eta_{\mathcal{T}}(u_{\mathcal{T}}, \omega(\mathcal{T} \setminus \mathcal{T}_*)) |u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}^0|_{H_0^1(\Omega)}.$$

Note that the interpolation estimate (9.9) for  $\mathcal{I}_{\mathcal{T}}^{\text{dG}}$  is responsible for the appearance of  $\omega(\mathcal{T} \setminus \mathcal{T}_*)$  rather than the smaller set  $\mathcal{T} \setminus \mathcal{T}_*$ .

For the second term, we use the lift estimate (9.26) along with the  $H^1$ -stability of  $\mathcal{I}_{\mathcal{T}}^{\text{dG}}$  and  $[[u_{\mathcal{T}_*}^0]] = 0$  to write

$$\int_{\Omega} \mathcal{L}_{\mathcal{T}}[u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}] \cdot \widehat{\mathbf{A}} \nabla_{\mathcal{T}} e_*^0 \lesssim \|h^{-1/2} [[u_{\mathcal{T}}]]\|_{L^2(\mathcal{F}^+)} |u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}^0|_{H_0^1(\Omega)}.$$

Inserting the estimates into (9.50), and recalling that  $1 \lesssim \bar{\kappa}_{\text{stab}} \leq \kappa$ , we find that

$$\begin{aligned} \|u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}}^2 &\lesssim \kappa^{1/2} \|h^{-1/2} [[u_{\mathcal{T}}]]\|_{L^2(\mathcal{F}^+)} \|u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}} \\ &\quad + \left( \eta_{\mathcal{T}}(u_{\mathcal{T}}, \mathcal{T} \setminus \mathcal{T}_*) + \kappa^{1/2} \|h^{1/2} [[u_{\mathcal{T}}]]\|_{L^2(\mathcal{F}^+)} \right) |u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}^0|_{H_0^1(\Omega)}. \end{aligned}$$

Notice that  $u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}^0 = u_{\mathcal{T}_*}^0 - u_{\mathcal{T}} + u_{\mathcal{T}}^\perp$  so that in view of Lemma 9.9 (control of non-conformity), we have

$$|u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}^0|_{H_0^1(\Omega)} = \|u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}^0\|_{\kappa, \mathcal{T}} \lesssim \|u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}} + \kappa^{1/2} \|h^{-1/2} [[u_{\mathcal{T}}]]\|_{L^2(\mathcal{F}^+)}.$$

The first desired inequality follows from the last two estimates. For the second inequality, it suffices to further invoke Lemma 9.11 (discontinuity control).  $\square$

#### 9.4 Module GALERKIN-DG

The main ingredients for the a posteriori estimation have been derived in the previous section and we can now turn our attention to the adaptive method. In

essence, it is the same as in the conforming case (Algorithm 5.4) but accounting for the perturbation arising from the non-conforming setting. Compared to Algorithm 5.4, SOLVE ( $\mathcal{T}$ ) determines the discontinuous Galerkin solution to (9.21) and REFINE ( $\mathcal{T}, \mathcal{M}$ ) produce the smallest  $\Lambda$ -admissible refinement of  $\mathcal{T}$  where all the marked elements  $\mathcal{M}$  are refined at least  $b \geq 1$  times.

**Algorithm 9.17 (GALERKIN-DG)** Let  $\widehat{\mathcal{T}} \geq \mathcal{T}_0$  be a  $\Lambda$ -admissible refinement,  $\Lambda \geq 0$ , of a suitable initial mesh  $\mathcal{T}_0$ . Let data  $\widehat{\mathcal{D}} = (\widehat{\mathbf{A}}, \widehat{c}, \widehat{f}) \in \mathbb{D}_{\widehat{\mathcal{T}}}$  be discrete on  $\widehat{\mathcal{T}}$  and  $\varepsilon > 0$  be a stopping tolerance. The following routine creates a  $\Lambda$ -admissible refinement  $\mathcal{T} \geq \widehat{\mathcal{T}}$  and discontinuous Galerkin solution  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}^{-1}$  for data  $\widehat{\mathcal{D}}$  such that  $\eta_{\mathcal{T}}(u_{\mathcal{T}}) \leq \varepsilon$ .

```

 $[\mathcal{T}, u_{\mathcal{T}}] = \text{GALERKIN-DG}(\widehat{\mathcal{T}}, \widehat{\mathcal{D}}, \varepsilon)$ 
  set  $j = 0$ ,  $\mathcal{T}_0 = \widehat{\mathcal{T}}$  and do
     $[u_j] = \text{SOLVE}(\mathcal{T}_j)$ ;
     $[\{\eta_j(u_j, T)\}_{T \in \mathcal{T}_j}] = \text{ESTIMATE}(u_j, \mathcal{T}_j, \widehat{\mathcal{D}})$ ;
    if  $\eta_j(u_j) \leq \varepsilon$ ;
      return  $(\mathcal{T}_j, u_j)$ 
     $[\mathcal{M}_j] = \text{MARK}(\{\eta_j(u_j, T)\}_{T \in \mathcal{T}_j}, \mathcal{T}_j, \theta)$ ;
     $[\mathcal{T}_{j+1}] = \text{REFINE}(\mathcal{T}_j, \mathcal{M}_j)$ ;
     $j \leftarrow j + 1$ ;
  while true

```

We start the analysis of GALERKIN-DG by investigating how the energy norm  $\mathcal{B}_{\mathcal{T}}[v, v]^{1/2}$  changes upon refining  $\mathcal{T}$ . Note that in the conforming case, Lemma 5.2 (Pythagoras) directly provides the relation  $\|\widehat{u} - u_{\mathcal{T}_*}\|_{\Omega} \leq \|\widehat{u} - u_{\mathcal{T}}\|_{\Omega}$ . In the non-conforming setting, the constant on the right-hand side is no longer 1 and jumps terms are present in the estimate. Regardless, it is possible to assess the effect of refinement in the energy norm and, in turn, compare two consecutive Galerkin solutions  $u_{\mathcal{T}}$  and  $u_{\mathcal{T}_*}$  where  $\mathcal{T} \in \mathbb{T}^{\Lambda}$  and  $\mathcal{T}_* = \text{REFINE}(\mathcal{T}, \mathcal{M})$  for some  $\mathcal{M} \subset \mathcal{T}$ . This is the subject of the next three results but before embarking on this path, we mention a key ingredient for this comparison to hold: *the routine REFINE does not refine elements in  $\mathcal{T}$  more than  $d$  times for  $b = 1$* ; see Corollary 3.31. This implies for any  $F \in \mathcal{F}^+$  and  $F_* \in \mathcal{F}_*^+$  with  $F_* \subset F$ , one has

$$h_F \lesssim h_{F_*}. \quad (9.51)$$

**Lemma 9.18 (mesh perturbation)** Let  $\mathcal{T} \in \mathbb{T}^{\Lambda}$  be a  $\Lambda$ -admissible refinement of  $\mathcal{T}_0$  satisfying Assumption 6.19 (initial labeling),  $\mathcal{M} \subset \mathcal{T}$ , and  $\mathcal{T}_* = \text{REFINE}(\mathcal{T}, \mathcal{M})$ . Assume that  $\widehat{\mathbf{A}}$  and  $\widehat{c}$  satisfy the structural assumption (9.20). There is a constant  $C$  such that for  $0 < \varepsilon < 1$  and all  $v \in \mathbb{E}_{\mathcal{T}}$  there holds

$$\mathcal{B}_{\mathcal{T}_*}[v, v] \leq (1 + \varepsilon)\mathcal{B}_{\mathcal{T}}[v, v] + C\varepsilon^{-1}\kappa\|h^{-1/2}[[v]]\|_{L^2(\mathcal{F}^+)}^2. \quad (9.52)$$

*Proof.* Because  $\nabla_{\mathcal{T}_*} v = \nabla_{\mathcal{T}} v$  when  $v \in \mathbb{E}_{\mathcal{T}}$ , we directly deduce that

$$\begin{aligned} \mathcal{B}_{\mathcal{T}_*}[v, v] &= \mathcal{B}_{\mathcal{T}}[v, v] + 2 \int_{\Omega} (\mathcal{L}_{\mathcal{T}}[v] - \mathcal{L}_{\mathcal{T}_*}[v]) \cdot \widehat{\mathbf{A}} \nabla_{\mathcal{T}} v \\ &\quad + \kappa \|h_*^{-1/2} [[v]]\|_{L^2(\mathcal{F}_*^+)}^2 - \kappa \|h_*^{-1/2} [[v]]\|_{L^2(\mathcal{F}^+)}^2, \end{aligned} \quad (9.53)$$

where  $h_* := h_{\mathcal{T}_*}$  denotes the mesh size of  $\mathcal{T}_*$ .

Unlike the broken gradients, the lifting operators are affected by refinements. However, this effect is controlled by the scaled jumps as we show now. Using estimate (9.27) twice with  $\epsilon = \frac{\varepsilon C_{\text{coer}}}{2}$  yields

$$\begin{aligned} &2 \int_{\Omega} (\mathcal{L}_{\mathcal{T}}[v] - \mathcal{L}_{\mathcal{T}_*}[v]) \cdot \widehat{\mathbf{A}} \nabla_{\mathcal{T}} v \\ &\leq \varepsilon C_{\text{coer}} \|\nabla_{\mathcal{T}} v\|_{L^2(\mathcal{T})}^2 + \frac{2C_{\text{lift}}^2}{\varepsilon C_{\text{coer}}} \left( \|h_*^{-1/2} [[v]]\|_{L^2(\mathcal{F}_*^+)}^2 + \|h_*^{-1/2} [[v]]\|_{L^2(\mathcal{F}^+)}^2 \right). \end{aligned}$$

Hence, the coercivity estimate (9.34) gives

$$\begin{aligned} &2 \int_{\Omega} (\mathcal{L}_{\mathcal{T}}[v] - \mathcal{L}_{\mathcal{T}_*}[v]) \cdot \widehat{\mathbf{A}} \nabla_{\mathcal{T}} v \\ &\leq \epsilon \mathcal{B}_{\mathcal{T}}[v, v] + \frac{2C_{\text{lift}}^2}{\varepsilon C_{\text{coer}}} \left( \|h_*^{-1/2} [[v]]\|_{L^2(\mathcal{F}_*^+)}^2 + \|h_*^{-1/2} [[v]]\|_{L^2(\mathcal{F}_*^+)}^2 \right). \end{aligned}$$

Inserting this back into (9.53), and using the fact that the jumps of  $v$  occur only on  $\mathcal{F}^+ \subset \mathcal{F}_*^+$ , the meshsize relation (9.51) proves the desired estimate.  $\square$

**Lemma 9.19 (comparison of solutions)** *Let  $\mathcal{T} \in \mathbb{T}^{\Lambda}$  be a  $\Lambda$ -admissible refinement of  $\mathcal{T}_0$  satisfying Assumption 6.19 (initial labeling),  $\mathcal{M} \subset \mathcal{T}$ , and  $\mathcal{T}_* = \text{REFINE}(\mathcal{T}, \mathcal{M})$ . Assume that  $\widehat{\mathbf{A}}$  and  $\widehat{c}$  satisfy the structural assumption (9.20). Let  $\widehat{u} = u(\widehat{\mathcal{D}}) \in H_0^1(\Omega)$  be the solution of the perturbed problem (5.5) with discrete data  $\widehat{\mathcal{D}}$  and denote by  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}^{-1}$ ,  $u_{\mathcal{T}_*} \in \mathbb{V}_{\mathcal{T}_*}^{-1}$  the Galerkin solutions associated to  $\mathcal{T}$ ,  $\mathcal{T}_*$  respectively with data  $\widehat{\mathcal{D}}$ . Let  $\bar{\kappa}_{\text{jump}}$  be as in Lemma 9.11. There exists a constant  $C_{\text{comp}}$  such that for all  $\kappa \geq \bar{\kappa}_{\text{jump}}$  and all  $0 < \epsilon < 1$  we have*

$$\begin{aligned} \mathcal{B}_{\mathcal{T}_*}[\widehat{u} - u_{\mathcal{T}_*}, \widehat{u} - u_{\mathcal{T}_*}] &\leq (1 + \epsilon) \mathcal{B}_{\mathcal{T}}[\widehat{u} - u_{\mathcal{T}}, \widehat{u} - u_{\mathcal{T}}] \\ &\quad - \frac{C_{\text{coer}}}{2} \|\nabla_{\mathcal{T}_*}(u_{\mathcal{T}_*} - u_{\mathcal{T}})\|_{L^2(\mathcal{T}_*)}^2 + \frac{C_{\text{comp}}}{\epsilon \kappa} (\eta_{\mathcal{T}}(u_{\mathcal{T}})^2 + \eta_{\mathcal{T}_*}(u_{\mathcal{T}_*})^2). \end{aligned}$$

*Proof.* We invoke the partial Galerkin orthogonality (9.45) of  $\widehat{u} - u_{\mathcal{T}_*}$  upon testing with  $v^0 := u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}^0 \in \mathbb{V}_{\mathcal{T}_*}^0$ :

$$\mathcal{B}_{\mathcal{T}_*}[\widehat{u} - u_{\mathcal{T}_*}, \widehat{u} - u_{\mathcal{T}_*}] = \mathcal{B}_{\mathcal{T}_*}[\widehat{u} - u_{\mathcal{T}_*} + v^0, \widehat{u} - u_{\mathcal{T}_*} + v^0] - \mathcal{B}_{\mathcal{T}_*}[v^0, v^0].$$

Note that

$$\widehat{u} - u_{\mathcal{T}_*} + v^0 = \widehat{u} - u_{\mathcal{T}} + u_{\mathcal{T}}^{\perp} - u_{\mathcal{T}_*}^{\perp}$$

and  $\|v^0\|_{\kappa, \mathcal{T}_*} = \|\nabla_{\mathcal{T}_*} v^0\|_{L^2(\mathcal{T}_*)}$ , which is critical for the argument below. Hence,

from the coercivity and continuity of  $\mathcal{B}_{\mathcal{T}_*}$  (Corollary 9.8), we deduce that

$$\begin{aligned} \mathcal{B}_{\mathcal{T}_*}[\widehat{u} - u_{\mathcal{T}_*}, \widehat{u} - u_{\mathcal{T}_*}] &\leq \mathcal{B}_{\mathcal{T}_*}[\widehat{u} - u_{\mathcal{T}}, \widehat{u} - u_{\mathcal{T}}] + 2C_{\text{cont}}^{1/2} \mathcal{B}_{\mathcal{T}_*}[\widehat{u} - u_{\mathcal{T}}, \widehat{u} - u_{\mathcal{T}}]^{1/2} \|u_{\mathcal{T}}^\perp - u_{\mathcal{T}_*}^\perp\|_{\kappa, \mathcal{T}_*} \\ &\quad + C_{\text{cont}} \|u_{\mathcal{T}}^\perp - u_{\mathcal{T}_*}^\perp\|_{\kappa, \mathcal{T}_*}^2 - C_{\text{coer}} \|\nabla_{\mathcal{T}_*}(u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}^0)\|_{L^2(\mathcal{T}_*)}^2. \end{aligned}$$

We now apply the reverse triangle inequality and Young's inequality

$$\|\nabla_{\mathcal{T}_*}(u_{\mathcal{T}_*}^0 - u_{\mathcal{T}}^0)\|_{L^2(\mathcal{T}_*)}^2 \geq \frac{1}{2} \|\nabla_{\mathcal{T}_*}(u_{\mathcal{T}_*} - u_{\mathcal{T}})\|_{L^2(\mathcal{T}_*)}^2 - \|\nabla_{\mathcal{T}_*}(u_{\mathcal{T}_*}^\perp - u_{\mathcal{T}}^\perp)\|_{L^2(\mathcal{T}_*)}^2$$

to deduce that for any  $0 < \varepsilon < 1$

$$\begin{aligned} \mathcal{B}_{\mathcal{T}_*}[\widehat{u} - u_{\mathcal{T}_*}, \widehat{u} - u_{\mathcal{T}_*}] &\leq (1 + \varepsilon) \mathcal{B}_{\mathcal{T}_*}[\widehat{u} - u_{\mathcal{T}}, \widehat{u} - u_{\mathcal{T}}] \\ &\quad - \frac{C_{\text{coer}}}{2} \|\nabla_{\mathcal{T}_*}(u_{\mathcal{T}_*} - u_{\mathcal{T}})\|_{L^2(\mathcal{T}_*)}^2 + \frac{C}{\varepsilon} \|u_{\mathcal{T}_*}^\perp - u_{\mathcal{T}}^\perp\|_{\kappa, \mathcal{T}_*}^2, \end{aligned}$$

where  $C$  is for the remainder of this proof a constant independent of the discretization parameters and  $\kappa$ .

To bound the last term we recall Lemma 9.9 (control of non-conformity)

$$\|u_{\mathcal{T}_*}^\perp - u_{\mathcal{T}}^\perp\|_{\kappa, \mathcal{T}_*} \leq \kappa^{1/2} \|h_*^{-1/2} [[u_{\mathcal{T}_*}]]\|_{L^2(\mathcal{F}_*^+)} + \kappa^{1/2} \|h_*^{-1/2} [[u_{\mathcal{T}}]]\|_{L^2(\mathcal{F}_*^+)},$$

and notice that the last integral over  $\mathcal{F}_*^+$  has weights relative to the local meshsize of  $\mathcal{T}_* \geq \mathcal{T}$ . Since for consecutive meshes the local meshsizes are comparable, according to (9.51), we can write  $\|h_*^{-1/2} [[u_{\mathcal{T}}]]\|_{L^2(\mathcal{F}_*^+)}$  instead. Inserting these expressions in the preceding estimate, and using Lemma 9.18 (mesh perturbation) to replace  $\mathcal{B}_{\mathcal{T}_*}$  with  $\mathcal{B}_{\mathcal{T}}$  on the right-hand side, yields

$$\begin{aligned} \mathcal{B}_{\mathcal{T}_*}[\widehat{u} - u_{\mathcal{T}_*}, \widehat{u} - u_{\mathcal{T}_*}] &\leq (1 + \varepsilon) \mathcal{B}_{\mathcal{T}}[\widehat{u} - u_{\mathcal{T}}, \widehat{u} - u_{\mathcal{T}}] - \frac{C_{\text{coer}}}{2} \|\nabla_{\mathcal{T}_*}(u_{\mathcal{T}_*} - u_{\mathcal{T}})\|_{L^2(\mathcal{T}_*)}^2 \\ &\quad + C \frac{\kappa}{\varepsilon} \left( \|h_*^{-1/2} [[u_{\mathcal{T}_*}]]\|_{L^2(\mathcal{F}_*^+)}^2 + \|h_*^{-1/2} [[u_{\mathcal{T}}]]\|_{L^2(\mathcal{F}_*^+)}^2 \right), \end{aligned}$$

where  $2\varepsilon$  has been relabeled  $\varepsilon$ . Finally, to derive the desired estimate, it remains to invoke Lemma 9.11 (discontinuity control).  $\square$

Combining Lemma 9.19 (comparison of solutions) with Lemma 9.10 (a posteriori error estimate), we derive the following dG version of Lemma 5.2 (Pythagoras).

**Corollary 9.20 (quasi-orthogonality of dG errors)** *If the assumptions of Lemma 9.19 hold, then for all  $\kappa \geq \kappa_{QO} := \frac{C_{\text{comp}}}{\varepsilon^2 C_L}$  and  $0 < \varepsilon \leq \frac{1}{4}$  there holds*

$$\|\widehat{u} - u_{\mathcal{T}_*}\|_{\kappa, \mathcal{T}_*}^2 \leq (1 + 4\varepsilon) \|\widehat{u} - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}}^2 - \frac{C_{\text{coer}}}{2} \|\nabla_{\mathcal{T}_*}(u_{\mathcal{T}_*} - u_{\mathcal{T}})\|_{L^2(\mathcal{T}_*)}^2.$$

*Proof.* We make use of the lower bound (9.40), and set  $D := \frac{C_{\text{comp}}}{\varepsilon C_L}$ , to rewrite the estimate of Lemma 9.19 as follows:

$$\left(1 - \frac{D}{\kappa}\right) \|\widehat{u} - u_{\mathcal{T}_*}\|_{\kappa, \mathcal{T}_*}^2 \leq \left(1 + \varepsilon + \frac{D}{\kappa}\right) \|\widehat{u} - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}}^2 - \frac{C_{\text{coer}}}{2} \|\nabla_{\mathcal{T}_*}(u_{\mathcal{T}_*} - u_{\mathcal{T}})\|_{L^2(\mathcal{T}_*)}^2.$$

For  $\kappa \geq \kappa_{\text{QO}} := \frac{D}{\varepsilon}$  this inequality implies

$$\|\widehat{u} - u_{\mathcal{T}_*}\|_{\kappa, \mathcal{T}_*}^2 \leq \frac{1+2\varepsilon}{1-\varepsilon} \|\widehat{u} - u_{\mathcal{T}}\|_{\kappa, \mathcal{T}}^2 - \frac{C_{\text{coer}}}{2(1-\varepsilon)} \|\nabla_{\mathcal{T}_*}(u_{\mathcal{T}_*} - u_{\mathcal{T}})\|_{L^2(\mathcal{T}_*)}^2.$$

It remains to realize that  $\frac{1+2\varepsilon}{1-\varepsilon} \leq 1 + 4\varepsilon$  provided  $\varepsilon \leq \frac{1}{4}$ .  $\square$

The last ingredient to prove convergence of GALERKIN-DG is a dG version of Proposition 4.56 (estimator reduction) with  $f = f_* \in \mathbb{F}_{\mathcal{T}}$ . It turns out that the same estimate and proof are valid for dG except that the  $H_0^1$ -seminorm is to be replaced by the broken  $H_0^1$ -seminorm. We thus state the result without proof.

**Proposition 9.21 (estimator reduction)** *Given  $\mathcal{T} \in \mathbb{T}^\Lambda$  and a subset  $\mathcal{M} \subset \mathcal{T}$  of elements marked for refinement, let  $\mathcal{T}_* = \text{REFINE}(\mathcal{T}, \mathcal{M})$ . If  $f = P_{\mathcal{T}}f \in \mathbb{F}_{\mathcal{T}}$ , then there is a constant  $C_{\text{Lip}}$  such that for all  $v \in \mathbb{V}_{\mathcal{T}}$ ,  $v_* \in \mathbb{V}_{\mathcal{T}_*}$  and any  $\delta > 0$*

$$\eta_{\mathcal{T}_*}(v_*, \mathcal{T}_*)^2 \leq (1+\delta)(\eta_{\mathcal{T}}(v, \mathcal{T})^2 - \lambda \eta_{\mathcal{T}}(v, \mathcal{M})^2) + (1+\delta^{-1}) C_{\text{Lip}}^2 \|\nabla_{\mathcal{T}_*}(v_* - v)\|_{L^2(\mathcal{T}_*)}^2.$$

We are now in a position to prove a contraction property between two consecutive iterations of the adaptive loop GALERKIN-DG.

**Theorem 9.22 (contraction property)** *Let  $\widehat{\mathcal{T}}$  be a  $\Lambda$ -admissible refinement of  $\mathcal{T}_0$  satisfying Assumption 6.19 (initial labeling). Let  $\widehat{\mathcal{D}} \in \mathbb{D}_{\widehat{\mathcal{T}}}$  such that  $\widehat{\mathbf{A}}$  and  $\widehat{c}$  satisfy the structural assumption (9.20). Let  $\theta \in (0, 1]$  be the Dörfler marking parameter used in the MARK module and  $\{\mathcal{T}_j, \mathbb{V}_j, u_j\}_{j=0}^J$  be a sequence of conforming meshes, finite element spaces and discrete solutions  $u_j = u_j(\widehat{\mathcal{D}}) \in \mathbb{V}_j$  created within GALERKIN-DG. If  $\widehat{u} \in H_0^1(\Omega)$  is the exact solution of (5.5) with discrete data  $\widehat{\mathcal{D}}$ , then there exist constants  $\bar{\kappa}_{\text{conv}} \geq 0$ ,  $\gamma > 0$ , and  $0 < \alpha < 1$  independent of the discretization parameters and  $\kappa$ , such that for all  $\kappa \geq \bar{\kappa}_{\text{conv}}$  and  $0 \leq j < J$*

$$B_{j+1}^2 + \gamma \eta_{\mathcal{T}_{j+1}}^2(u_{j+1}) \leq \alpha^2 (B_j^2 + \gamma \eta_{\mathcal{T}_j}^2(u_j)), \quad (9.54)$$

where  $B_j := (\mathcal{B}_{\mathcal{T}_j}[\widehat{u} - u_j, \widehat{u} - u_j])^{1/2}$  is the dG norm of  $\widehat{u} - u_j$ .

*Proof.* In essence, we proceed as in Theorem 5.8 (general contraction property) for the conforming case but with minor changes that account for non-conformity. We only explain the differences below. For  $j \geq 0$ , we shorten the notations and write  $\eta_j := \eta_{\mathcal{T}_j}(u_j)$  and  $E_j := \|\nabla_{\mathcal{T}_{j+1}}(u_{j+1} - u_j)\|_{L^2(\mathcal{T}_{j+1})}$ .

Corollary 9.20 (quasi-orthogonality of dG errors) gives for any  $0 < \varepsilon \leq \frac{1}{4}$

$$B_{j+1}^2 \leq (1 + 4\varepsilon) B_j^2 - \frac{C_{\text{coer}}}{2} E_j^2.$$

Combining Proposition 9.21 (estimator reduction), written in terms of  $\mathcal{T} = \mathcal{T}_j$ ,  $\mathcal{T}_* = \mathcal{T}_{j+1}$ ,  $v = u_j$  and  $v_* = u_{j+1}$ , with Dörfler marking  $\eta_j(u_j, \mathcal{M}_j) \geq \theta \eta_j$  yields

$$\eta_{j+1}^2 \leq (1 + \delta)(1 - \lambda \theta^2) \eta_j^2 + (1 + \delta^{-1}) C_{\text{Lip}}^2 E_j^2$$



for any  $\delta > 0$ . We now multiply this inequality by  $\gamma > 0$  and add it to the previous one with the following choice of parameters:

$$\delta = \frac{1 - \frac{\lambda\theta^2}{2}}{1 - \lambda\theta^2} - 1, \quad \gamma = \frac{C_{\text{coer}}}{2C_{\text{Lip}}^2(1 + \delta^{-1})}.$$

Consequently, the terms involving  $E_j^2$  cancel out and we end up with

$$\begin{aligned} B_{j+1}^2 + \gamma\eta_{j+1}^2 &\leq (1 + 4\varepsilon)B_j^2 + \gamma(1 + \delta)(1 - \lambda\theta^2)\eta_j^2 \\ &= \left(1 + 4\varepsilon - \frac{\gamma\lambda\theta^2}{4}\right)B_j^2 + \gamma\left(1 - \frac{\lambda\theta^2}{4}\right)\eta_j^2. \end{aligned}$$

We finally choose  $\varepsilon := \frac{\gamma\lambda\theta^2}{32}$  to obtain (9.54) with  $\alpha^2 = \max\left\{1 - \frac{\gamma\lambda\theta^2}{8}, 1 - \frac{\lambda\theta^2}{4}\right\}$ , and conclude the proof.  $\square$

**Corollary 9.23 (linear convergence)** *Under the assumptions of Theorem 9.22, and if  $0 < \alpha < 1$ ,  $\gamma > 0$ ,  $\bar{\kappa}_{\text{conv}} > 0$  are the constants in (9.54), then for all  $\kappa \geq \bar{\kappa}_{\text{conv}}$  there holds*

$$\|\widehat{u} - u_k\|_{\kappa, \mathcal{T}_k} \leq C_* \alpha^{k-j} \|\widehat{u} - u_j\|_{\kappa, \mathcal{T}_j},$$

for some constant  $C_*$  independent of the discretization parameters and  $\kappa$ .

*Proof.* Let  $e_k^2 = \|\widehat{u} - u_k\|_{\kappa, \mathcal{T}_k}^2$  and  $B_k^2 = \mathcal{B}_{\mathcal{T}_k}[\widehat{u} - u_k, \widehat{u} - u_k]^2$ , and use the coercivity estimate (9.34), the contraction property (9.54), the continuity estimate (9.33) and the lower bound (9.40) to arrive at

$$C_{\text{coer}}e_k^2 \leq B_k^2 \leq \alpha^{2(k-j)}(B_j^2 + \gamma\eta_j^2) \leq \alpha^{2(k-j)}\left(C_{\text{cont}} + \frac{\gamma}{C_L^2}\right)e_j^2.$$

This is the desired estimate in disguised with  $C_* := \frac{1}{C_{\text{coer}}}\left(C_{\text{cont}} + \frac{\gamma}{C_L^2}\right)^{1/2}$ .  $\square$

We end the discussion on GALERKIN-DG by deriving the optimality property of the Dörfler marking strategy. We mimic the proof of Lemma 6.16 (Dörfler marking) but directly use the optimal parameter  $\mu = \frac{1}{2}$  to simplify the argument. We refer to the discussion after Lemma 6.16 for the role of  $\mu$  and its influence on  $\theta_0$ . Notice that  $\theta_0$  depends on  $\kappa^{-1}$  because of its appearance in the perturbed localized upper bound (9.49). It plays a similar role to  $\sigma$  in Assumption 6.15 (restriction on  $\omega$ ) in the presence of oscillations (one step method with switch).

**Lemma 9.24 (Dörfler marking)** *Let  $\mathcal{T}_* \geq \mathcal{T}$  be two  $\Lambda$ -admissible refinements of  $\mathcal{T}_0$  satisfying Assumption 6.19 (initial labeling). Let  $\widehat{\mathcal{D}} \in \mathbb{D}_{\widehat{\mathcal{T}}}$  such that  $\widehat{\mathbf{A}}$  and  $\widehat{\mathbf{c}}$  satisfy the structural assumption (9.20). Let  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}^{-1}$ ,  $u_{\mathcal{T}_*} \in \mathbb{V}_{\mathcal{T}_*}^{-1}$  denote the Galerkin solutions associated with  $\mathcal{T}$ ,  $\mathcal{T}_*$  respectively and  $\widehat{u} \in H_0^1(\Omega)$  denotes the solution to (5.5) with discrete data  $\widehat{\mathcal{D}}$ . Assume  $\kappa > \bar{\kappa}_D := \max(\bar{\kappa}_{\text{stab}}, 4C_{\text{Lip}}^2 C_{LU}^2)$ . If*

$$\eta_{\mathcal{T}_*}(u_{\mathcal{T}_*}) \leq \frac{1}{2} \eta_{\mathcal{T}}(u_{\mathcal{T}}) \quad (9.55)$$

then the refined set  $\mathcal{T} \setminus \mathcal{T}_*$  satisfies the Dörfler property

$$\eta_{\mathcal{T}}(u_{\mathcal{T}}, \omega(\mathcal{T} \setminus \mathcal{T}_*)) \geq \theta_0 \eta_{\mathcal{T}}(u_{\mathcal{T}}), \quad (9.56)$$

$$\text{with } 0 < \theta_0^2 := \theta_0^2(\kappa) := \frac{1 - 4C_{\text{Lip}}^2 C_{\text{LU}}^2 \kappa^{-1}}{4C_{\text{Lip}}^2 C_{\text{LU}}^2} < \frac{1}{4C_{\text{Lip}}^2 C_{\text{LU}}^2}.$$

*Proof.* To relate  $\eta_{\mathcal{T}}$  with  $\eta_{\mathcal{T}_*}$ , we invoke Proposition 9.21 (estimator reduction) with  $\delta = 1$ , along with the localized upper bound (9.49), to write

$$\eta_{\mathcal{T}}(u_{\mathcal{T}})^2 \leq 2\eta_{\mathcal{T}_*}(u_{\mathcal{T}_*})^2 + 2C_{\text{Lip}}^2 C_{\text{LU}}^2 \left( \eta_{\mathcal{T}}(u_{\mathcal{T}}, \omega(\mathcal{T} \setminus \mathcal{T}_*))^2 + \kappa^{-1} \eta_{\mathcal{T}}(u_{\mathcal{T}})^2 \right).$$

This, combined with (9.55) yields

$$\left( \frac{1}{2} - 2C_{\text{Lip}}^2 C_{\text{LU}}^2 \kappa^{-1} \right) \eta_{\mathcal{T}}(u_{\mathcal{T}})^2 \leq 2C_{\text{Lip}}^2 C_{\text{LU}}^2 \eta_{\mathcal{T}}(u_{\mathcal{T}}, \omega(\mathcal{T} \setminus \mathcal{T}_*))^2$$

for  $\kappa \geq \bar{\kappa}_{\text{D}}$ . This is the desired results in disguised.  $\square$

### 9.5 Convergence of AFEM-DG-TS

Algorithm 9.1 (AFEM-DG-TS) relies on two modules: GALERKIN-DG and DATA. We have analyzed the performance of GALERKIN-DG in the previous section and showed in Section 7 that the output  $[\widehat{\mathcal{T}}_k, \widehat{\mathcal{D}}_k] = \text{DATA}(\mathcal{T}_k, \mathcal{D}, \omega \varepsilon_k)$  satisfies

$$\|\mathcal{D} - \widehat{\mathcal{D}}_k\|_{D(\Omega)} \leq \omega \varepsilon_k \quad \Rightarrow \quad |u - \widehat{u}_k|_{H_0^1(\Omega)} \leq C_D \omega \varepsilon_k. \quad (9.57)$$

Recall that  $u = u(\mathcal{D})$ ,  $\widehat{u}_k = u(\widehat{\mathcal{D}}_k) \in H_0^1(\Omega)$  are the exact solutions to (2.7) with exact data  $\mathcal{D}$  and discrete data  $\widehat{\mathcal{D}}_k$ , respectively. We also recall that  $\widehat{\mathcal{D}}_k$  satisfies the structural assumption (9.20) uniformly in  $k$  and thus  $C_D$  does not depend on  $k$ .

We start with a result guaranteeing that the cost of GALERKIN-DG does not depend on the iteration counter  $k$  within AFEM-DG-TS.

**Lemma 9.25 (computational cost of GALERKIN-DG)** *For any  $\kappa \geq \bar{\kappa}_{\text{conv}}$  and any  $k \in \mathbb{N}$ , the number of sub-iterations  $J_k$  inside a call of GALERKIN-DG at iteration  $k$  of Algorithm 9.1 (AFEM-DG-TS) is bounded independently of  $k$ .*

*Proof.* We proceed as in the proof of Proposition 5.27 (computational cost of GALERKIN) for the conforming case, and focus on the essential differences. We fix the iteration counter  $k \geq 1$ , recall that the output of the  $(k-1)$ -th loop of AFEM-DG-TS is  $[\mathcal{T}_k, u_k] = \text{GALERKIN-DG}(\widehat{\mathcal{T}}_{k-1}, \widehat{\mathcal{D}}_{k-1}, \varepsilon_{k-1})$ , and denote by  $\mathcal{T}_{k,j}$  and  $\widehat{u}_{k,j} \in \mathbb{V}_{\mathcal{T}_{k,j}}^{-1}$  the  $j$ -th mesh and Galerkin solution to (9.21) with data  $\widehat{\mathcal{D}}_k$  in the  $k$ -th loop of AFEM-DG-TS. The exact solution to the perturbed problem (5.5) with discrete coefficient  $\widehat{\mathcal{D}}_k$  is  $\widehat{u}_k = u(\widehat{\mathcal{D}}_k)$ .

We recall that  $\mathcal{T}_{k,0} = \widehat{\mathcal{T}}_k$  is the mesh produced by DATA, and assume that  $u_{k,0} \in \mathbb{V}_{k,0}$  satisfies  $\eta_{\mathcal{T}_{k,0}}(u_{k,0}) > \varepsilon_k$  because otherwise  $J_k = 0$  and there is nothing to prove. In view of Corollary 9.23 (linear convergence), all we need to prove is that the error  $\|u_{k,0} - \widehat{u}_k\|_{\kappa, \mathcal{T}_{k,0}}$  entering GALERKIN-DG is bounded by  $\varepsilon_k$ . We resort to Corollary 9.15 (quasi-monotonicity with different data) to write

$$\|u_{k,0} - \widehat{u}_k\|_{\kappa, \mathcal{T}_{k,0}} \leq C_{\text{Mo}} \left( \|u_k - \widehat{u}_{k-1}\|_{\kappa, \mathcal{T}_k} + |\widehat{u}_k - \widehat{u}_{k-1}|_{H_0^1(\Omega)} \right).$$

The appearance of the last term is the only difference with respect to Proposition 5.27. However, in view of property (9.57) of DATA, we infer that

$$|\widehat{u}_k - \widehat{u}_{k-1}|_{H_0^1(\Omega)} \leq |\widehat{u}_k - u|_{H_0^1(\Omega)} + |\widehat{u}_{k-1} - u|_{H_0^1(\Omega)} \leq C_D \omega (\varepsilon_k + \varepsilon_{k-1}) = 3C_D \omega \varepsilon_k.$$

Moreover, the stabilization-free upper bound (9.44) implies

$$\|u_k - \widehat{u}_{k-1}\|_{\kappa, \mathcal{T}_k} \leq C_U \eta_{\mathcal{T}_k}(u_k) \leq C_U \varepsilon_{k-1} = 2C_U \varepsilon_k,$$

which, combined with the lower bound (9.40), further yields

$$\|u_{k,0} - \widehat{u}_k\|_{\kappa, \mathcal{T}_{k,0}} \leq (3C_D \omega + 2C_U) \varepsilon_k = \Lambda \varepsilon_k \quad \Rightarrow \quad \eta_{\mathcal{T}_{k,0}} \leq C_L^{-1} \Lambda \varepsilon_k.$$

This is the requisite estimate. In fact, recalling Corollary 9.23, we see that

$$\eta_{\mathcal{T}_{k,j}}(u_{k,j}) \leq C_L^{-1} \|u_{k,j} - \widehat{u}_k\|_{\kappa, \mathcal{T}_{k,j}} \leq C_L^{-1} C_* \alpha^j \|u_{k,0} - \widehat{u}_k\|_{\kappa, \mathcal{T}_{k,0}} \leq C_L^{-1} C_* \Lambda \varepsilon_k \alpha^j.$$

Since GALERKIN-DG stops when  $\eta_{\mathcal{T}_{k,k}}(u_{J_k,k}) \leq \varepsilon_k$ , we finally conclude as in the proof of Proposition 5.27 that  $J_k$  is independent of  $k$ .  $\square$

The proof of convergence of AFEM-DG-TS is identical to the proof of Proposition 5.29 upon replacing the semi-norm  $|\cdot|_{H_0^1(\Omega)}$  by the appropriate dG norm. It is therefore not repeated here.

**Proposition 9.26 (convergence of AFEM-DG-TS)** *For any  $\kappa \geq \bar{\kappa}_{\text{conv}}$  and  $k \geq 0$ , the  $(k+1)$ -th iteration of AFEM-DG-TS terminates and requires a finite number of inner iterations of GALERKIN-DG independent of  $k$ . Moreover, if  $u \in H_0^1(\Omega)$  denotes the solution to (2.7), there exists a constant  $C_*$  such that the output of  $[\mathcal{T}_{k+1}, u_{k+1}] = \text{GALERKIN-DG}(\widehat{\mathcal{T}}_k, \widehat{\mathcal{D}}_k, \varepsilon_k)$  satisfies*

$$\|u - u_{k+1}\|_{\kappa, \mathcal{T}_{k+1}} \leq C_* \varepsilon_k, \quad \forall k \geq 0.$$

Therefore, AFEM-DG-TS stops after  $K < 2 + \frac{\log \frac{\varepsilon_0}{\text{tol}}}{\log 2}$  iterations and delivers

$$\|u - u_{K+1}\|_{\kappa, \mathcal{T}_K} \leq C_* \text{tol}.$$

## 9.6 Rate-optimality of AFEM-DG-TS

To derive rates of convergence for the discontinuous Galerkin method, we proceed similarly to Section 6 for the conforming case. Recall that in the  $k$ th-step of Algorithm 9.1 (AFEM-DG-TS), the output of  $[\widehat{\mathcal{T}}_k, \widehat{\mathcal{D}}_k] = \text{DATA}(\mathcal{T}_k, \mathcal{D}, \omega \varepsilon_k)$  is fed to  $[\mathcal{T}_{k+1}, u_{k+1}] = \text{GALERKIN-DG}(\widehat{\mathcal{T}}_k, \widehat{\mathcal{D}}_k, \varepsilon_k)$ , which in turn iterates  $J_k$  times.

Lemma 9.25 shows that  $J_k$  is uniformly bounded in  $k$ , and we assume that  $J_k \geq 1$  for otherwise the module GALERKIN-DG is skipped altogether. We denote by  $(\mathcal{T}_{k,j}, \mathcal{M}_{k,j}, u_{\mathcal{T}_{k,j}})$  the triplets of grids, marked sets and discrete solutions computed within GALERKIN-DG  $(\widehat{\mathcal{T}}_k, \widehat{\mathcal{D}}_k, \varepsilon_k)$  for  $0 \leq j < J_k$ . Note that

$$\widehat{\varepsilon}_{k,j} := \eta_{\mathcal{T}_{k,j}}(u_{\mathcal{T}_{k,j}}, \widehat{\mathcal{D}}_k) > \varepsilon_k, \quad 0 \leq j < J_k$$

so that together with the lower a posteriori error estimate (9.40), we infer that

$$\|\widehat{u}_k - u_{\mathcal{T}_{k,j}}\|_{\kappa, \mathcal{T}_{k,j}} \geq C_L \widehat{\varepsilon}_{k,j} > C_L \varepsilon_k,$$

where  $\widehat{u}_k = u(\widehat{\mathcal{D}}_k) \in H_0^1(\Omega)$  is the exact solution with approximate data  $\widehat{\mathcal{D}}_k$ .

The module DATA guarantees (9.57), and the parameter  $\omega$  modulates the discrepancy between  $u$  and  $\widehat{u}_k$  relative to  $\varepsilon_k$ . The error due to data approximation can be made small relative to the finite element approximation by choosing  $\omega$  much smaller than 1. In addition, we have established Lemma 9.24 (Dörfler marking) for  $\theta_0 < 1$  which implies a Dörfler property for any  $0 < \theta \leq \theta_0$ . The restrictions on the parameters  $\kappa$ ,  $\omega$ , and  $\theta$  are gathered in the following assumption.

**Assumption 9.27 (restrictions on  $\kappa$ ,  $\omega$  and  $\theta$ )** Assume that  $\kappa > \max(\bar{\kappa}_D, \bar{\kappa}_{\text{conv}})$ , that  $0 < \omega \leq \frac{1}{4} C_{\text{Mo}}^{-1} C_L C_D^{-1}$  and that  $0 < \theta \leq \theta_0(\kappa)$ , where  $\kappa_D$  and  $\theta_0$  are defined in Lemma 9.24 (Dörfler marking).

Note that if Assumption 9.27 is valid then

$$|u - \widehat{u}_k|_{H_0^1(\Omega)} \leq \frac{1}{4} C_{\text{Mo}}^{-1} C_L \varepsilon_k. \quad (9.58)$$

The next results rely on Assumption 6.3 (approximability of  $u$ ) and Assumption 6.10 (approximability of data). They are stated and proved for conforming meshes and continuous approximations of  $u$ . However, Proposition 9.4 (equivalence of classes for  $u$ ) and Remark 9.5 (equivalence of classes for  $\mathcal{D}$ ) show that these classes coincide with the conforming case.

**Proposition 9.28 (cardinality of marked sets)** *Let Assumptions 6.3 (approximability of  $u$ ), 6.17 (cardinality of  $\mathcal{M}$ ), and 9.27 (restrictions on  $\kappa$ ,  $\omega$  and  $\theta$ ) hold. If  $\widehat{\varepsilon}_{k,0} > \varepsilon_k$ , then GALERKIN-DG at iteration  $k$  of AFEM-DG-TS is called and the cardinality  $N_{k,j}(u)$  of the marked set  $\mathcal{M}_{k,j}$  satisfies*

$$N_{k,j}(u) \lesssim |u|_s^{1/s} \|u - u_{\mathcal{T}_{k,j}}\|_{\kappa, \mathcal{T}_{k,j}}^{-1/s} \quad \forall 0 \leq j < J_k. \quad (9.59)$$

*Proof.* Fix  $0 \leq j < J_k$  and set

$$\delta := \frac{1}{2} C_{\text{Mo}}^{-1} C_L \eta_{\mathcal{T}_{k,j}}(u_{\mathcal{T}_{k,j}}) \geq \frac{1}{2} C_{\text{Mo}}^{-1} C_L \varepsilon_k,$$

because  $\eta_{\mathcal{T}_{k,j}}(u_{\mathcal{T}_{k,j}}) > \varepsilon_k$  for  $j < J_k$ . Thanks to (9.58),  $\widehat{u}_k$  is an  $(\frac{1}{2} C_{\text{Mo}}^{-1} C_L \varepsilon_k)$ -approximation of order  $s$  to  $u$  according to Lemma 6.13 ( $\varepsilon$ -approximation of order

s). Therefore, there exists a conforming mesh  $\mathcal{T}_\delta \in \mathbb{T}^\Lambda$  and  $u_{\mathcal{T}_\delta}^0 \in \mathbb{V}_{\mathcal{T}_\delta}$  such that

$$\|\widehat{u}_k - u_{\mathcal{T}_\delta}^0\|_{\kappa, \mathcal{T}_\delta} = |\widehat{u}_k - u_{\mathcal{T}_\delta}^0|_{H_0^1(\Omega)} \leq \delta, \quad \#\mathcal{T}_\delta \lesssim |u|_{\frac{1}{s}}^{\frac{1}{s}} \delta^{-\frac{1}{s}}.$$

To compare  $\mathcal{T}_\delta$  with  $\mathcal{T}_{k,j}$  we consider the overlay  $\mathcal{T}_* = \mathcal{T}_{k,j} \oplus \mathcal{T}_\delta$ , which satisfies

$$\#\mathcal{T}_* \leq \#\mathcal{T}_{k,j} + \#\mathcal{T}_\delta - \#\mathcal{T}_0;$$

see Proposition 8.15 (mesh overlay is  $\Lambda$ -admissible). Let  $u_{\mathcal{T}_*} \in \mathbb{V}_{\mathcal{T}_*}^{-1}$  be the Galerkin solution on the subspace  $\mathbb{V}_{\mathcal{T}_*}^{-1}$  and invoke Corollary 9.13 (Cea's lemma) to write

$$\eta_{\mathcal{T}_*}(u_{\mathcal{T}_*}) \leq C_L^{-1} \|\widehat{u}_k - u_{\mathcal{T}_*}\|_{\kappa, \mathcal{T}_*} \leq C_L^{-1} C_{\text{Mo}} \|\widehat{u}_k - u_{\mathcal{T}_\delta}^0\|_{\kappa, \mathcal{T}_*} = C_L^{-1} C_{\text{Mo}} |\widehat{u}_k - u_{\mathcal{T}_\delta}^0|_{H_0^1(\Omega)},$$

whence  $\eta_{\mathcal{T}_*}(u_{\mathcal{T}_*}) \leq C_L^{-1} C_{\text{Mo}} \delta$  and

$$\eta_{\mathcal{T}_*}(u_{\mathcal{T}_*}) \leq \frac{1}{2} \eta_{\mathcal{T}_{k,j}}(u_{\mathcal{T}_{k,j}}).$$

Applying Lemma 9.24 (Dörfler marking) to  $\mathcal{T}_*$  and  $\mathcal{T}_{k,j}$  we infer that the enlarged refined set  $\omega(\mathcal{T}_{k,j} \setminus \mathcal{T}_*)$  satisfies the Dörfler marking property

$$\eta_{\mathcal{T}_{k,j}}(u_{\mathcal{T}_{k,j}}, \omega(\mathcal{T}_{k,j} \setminus \mathcal{T}_*)) \geq \theta \eta_{\mathcal{T}_*}(u_{\mathcal{T}_*})$$

since  $0 < \theta \leq \theta_0$  by Assumption 9.27. The Dörfler marking involves a minimal set  $\mathcal{M}_{k,j}$  according to Assumption 6.17, which thus implies

$$N_{k,j}(u) \leq \#\omega(\mathcal{T}_{k,j} \setminus \mathcal{T}_*) \lesssim \#(\mathcal{T}_{k,j} \setminus \mathcal{T}_*) \leq \#\mathcal{T}_\delta - \#\mathcal{T}_0 \lesssim |u|_{\frac{1}{s}}^{\frac{1}{s}} \delta^{-\frac{1}{s}} \lesssim |u|_{\frac{1}{s}}^{\frac{1}{s}} \varepsilon_k^{-\frac{1}{s}}.$$

because  $\#(\mathcal{T}_{k,j} \setminus \mathcal{T}_*) \leq \#\mathcal{T}_* - \#\mathcal{T}_{k,j}$ . This concludes the proof.  $\square$

**Corollary 9.29 (quasi-optimality of GALERKIN-DG)** *Let Assumptions 6.3 (approximability of  $u$ ), 6.17 (cardinality of  $\mathcal{M}$ ), and 9.27 (restrictions on  $\kappa$ ,  $\omega$  and  $\theta$ ) hold. Assume  $\kappa \geq \max(\bar{\kappa}_{\text{conv}}, \bar{\kappa}_D)$ . Then, the total number of marked elements  $N_k(u)$  within a call to GALERKIN-DG satisfies*

$$N_k(u) \leq J C_0 |u|_{\frac{1}{s}}^{\frac{1}{s}} \varepsilon_k^{-\frac{1}{s}},$$

where  $J \geq J_k$  is a uniform upper bound for the number of iterations of GALERKIN-DG according to Lemma 9.25 (computational cost of GALERKIN-DG).

*Proof.* Use that  $N_k(u) = \sum_{j=0}^{J_k-1} N_{k,j}(u)$  and combine Propositions 9.28 (cardinality of marked sets) and 9.25 (computational cost of GALERKIN-DG).  $\square$

We finally address the rate-optimality of the two-step algorithm AFEM-DG-TS.

**Theorem 9.30 (rate-optimality of AFEM-DG-TS)** *Let Assumptions 6.3 (approximability of  $u$ ), 6.10 (approximability of data), 6.11 (quasi-optimality of DATA), 6.17 (cardinality of  $\mathcal{M}$ ), 6.19 (initial labeling), and 9.27 (restrictions on  $\kappa$ ,  $\omega$  and  $\theta$ ) hold. Then, AFEM-DG-TS gives rise to a sequence  $(\mathcal{T}_k, \mathbb{V}_{\mathcal{T}_k}^{-1}, u_{\mathcal{T}_k})_{k=0}^{K+1}$  such that*

$$\|u - u_{\mathcal{T}_k}\|_{\kappa, \mathcal{T}_k} \leq C(u, \mathcal{D}) (\#\mathcal{T}_k)^{-s} \quad 1 \leq k \leq K+1,$$

where  $0 < s = \min\{s_u, s_{\mathcal{D}}\} = \min\{s_u, s_A, s_C, s_f\} \leq \frac{n}{d}$  and

$$C(u, \mathcal{D}) = C_* \left( |u|_{\mathbb{A}_{s_u}}^{\frac{1}{s_u}} + |A|_{\mathbb{M}_{s_A}}^{\frac{1}{s_A}} + |c|_{\mathbb{C}_{s_C}}^{\frac{1}{s_C}} + |f|_{\mathbb{F}_{s_f}}^{\frac{1}{s_f}} \right)^s$$

with constant  $C_* > 0$  independent of  $u$  and  $\mathcal{D}$ .

*Proof.* Assumptions 6.3, 6.17, and 9.27 combined with Corollary 9.29 for  $u$ , and Assumptions 6.10 and 6.11 for  $\mathcal{D}$ , imply the existence of a constant  $C_{\#}$  such that the total number of marked elements within one loop of AFEM-DG-TS is

$$N_k(u) + N_k(\mathcal{D}) \leq C_{\#} \left( |u|_{\mathbb{A}_{s_u}}^{\frac{1}{s_u}} + |\mathcal{D}|_{\mathbb{A}_{s_{\mathcal{D}}}}^{\frac{1}{s_{\mathcal{D}}}} \right) \varepsilon_k^{-\frac{1}{s}},$$

with  $s_u, s_{\mathcal{D}} \leq \frac{n}{d}$ . Moreover, upon termination DATA and GALERKIN-DG give

$$|u - \widehat{u}_k|_{H_0^1(\Omega)} \leq \frac{1}{4} C_{\text{Mo}}^{-1} C_L \varepsilon_k,$$

$$\|\widehat{u}_k - u_{\mathcal{T}_{k+1}}\|_{\kappa, \mathcal{T}_{k+1}} \leq C_U \eta_{\mathcal{T}_{k+1}}(u_{\mathcal{T}_{k+1}}) \leq C_U \varepsilon_k,$$

because of (9.58) and (9.44). This implies by triangle inequality

$$\|u - u_{\mathcal{T}_{k+1}}\|_{\kappa, \mathcal{T}_{k+1}} \leq \left( \frac{1}{4} C_{\text{Mo}}^{-1} C_L + C_U \right) \varepsilon_k.$$

We finally conclude as in Theorem 6.24 (rate-optimality of AFEM-TS).  $\square$

Remark 6.25 about the role of  $\omega$ ,  $\theta^*$  and Remark 6.26 about the optimality of the result, written after Theorem 6.24 for the conforming case, remain valid for the non-conforming case and are not repeated here.

### 9.7 Operator $P_{\mathcal{T}}$ and routine DATA on $\Lambda$ -admissible partitions

In this section we have used extensively the notion of  $\Lambda$ -admissible meshes for the design and study of dG methods, including forcing  $f \in H^{-1}(\Omega)$ . To this end, as well as for the design of the two-step AFEM for dG, namely AFEM-DG-TS, the construction of the local projection  $P_{\mathcal{T}}f \in \mathbb{F}_{\mathcal{T}}$  is critical. We discuss this now.

Recall that for a conforming partition  $\mathcal{T} \in \mathbb{T}$ ,  $P_{\mathcal{T}}f$  is defined as a projection to  $\mathbb{F}_{\mathcal{T}}$ ; see Definition 4.24 (projection onto discrete functionals). The definition and subsequent properties of  $P_{\mathcal{T}}$  hinge on extensions  $E_F$  for  $F \in \mathcal{F}$ , studied in Lemma 4.20 (extending from faces), as well as on bubble functions  $\phi_T$ ,  $T \in \mathcal{T}$ , and  $\phi_F$ ,  $F \in \mathcal{F}$  satisfying Assumption 4.21 (abstract cut-off).

The definition of the element bubble functions  $\phi_T$  in (4.14) is local to  $T$  and is thus unchanged on non-conforming subdivisions. The situation is different for faces. If  $F$  is a conforming face, we have the conforming definitions of  $E_F$  and  $\phi_F$ . Instead, if  $F$  is a non-conforming face, i.e.  $F = T \cap T_*$  with  $g(T_*) > g(T)$ , and use a virtual conforming refinement of  $\omega_F$  to define  $E_F$  and  $\phi_F$  as in (4.17). Recall that  $g(T)$  is the generation of  $T \in \mathcal{T}$ , and  $\mathcal{T} \in \mathbb{T}$  is a uniform refinement of  $\mathcal{T}_0$  if and only

if  $g$  is constant on  $\mathcal{T}$ . Let  $\overline{\mathcal{T}}$  be the uniform refinement of  $\mathcal{T}_0$  containing  $T_*$ , whence  $g(T) = g(T_*)$  for all  $T \in \overline{\mathcal{T}}$ ;  $\overline{\mathcal{T}}$  is conforming thanks to Assumption 6.19 (initial labeling) on  $\mathcal{T}_0$ . Let  $\overline{T} \in \overline{\mathcal{T}}$  be the element sharing  $F$  with  $T_*$  (and thus contained in  $T$ ) and let  $\overline{\omega}_F := T_* \cup \overline{T}$  be the virtual conforming patch around  $F$ . We now proceed by defining  $E_F$  via (4.4) with  $\omega_F$  replaced by  $\overline{\omega}_F$  and  $\phi_F$  as in (4.17) using the basis functions  $\phi_z, z \in \mathcal{V} \cap F$ , associated with  $\prod_{T \subset \overline{\omega}_F} \mathbb{P}_n(T) \cap H_0^1(\overline{\omega}_F)$ . Note that because  $\mathcal{T}$  is  $\Lambda$ -admissible, Proposition 3.27 guarantees that the diameters of  $\overline{T}, T, T_*, \omega_F$  and  $\overline{\omega}_F$  are all comparable with constants depending on the initial mesh  $\mathcal{T}_0$  and  $\Lambda$ .

Assumption 4.21 is an important ingredient in the analysis of  $P_{\mathcal{T}}$  and it holds true with  $\omega_F$  replaced by  $\overline{\omega}_F$  when  $F$  is a non-conforming face. Therefore, Remark 4.26 (local computation) and Corollary 4.31 (local near-best approximation) are valid for  $\Lambda$ -admissible partitions as well. Consequently, all the algorithms and results presented in Section 7 (data approximation) readily extend to  $\Lambda$ -admissible subdivisions as well. We do not dwell on this matter any longer.

## 10 AFEMs for Inf-Sup Stable Problems

We go back to the functional framework introduced in Section 2.4. Precisely, let the bilinear form  $\mathcal{B} : \mathbb{V} \times \mathbb{W} \rightarrow \mathbb{R}$  be continuous and *inf-sup stable* (i.e., it satisfies one of the equivalent conditions stated in Theorem 2.8 (Nečas)). Given  $f \in \mathbb{W}^*$ , let  $u \in \mathbb{V}$  be the unique solution of the variational problem

$$u \in \mathbb{V} : \quad \mathcal{B}[u, w] = \langle f, w \rangle \quad \forall w \in \mathbb{W}. \quad (10.1)$$

Let  $\mathbb{V}_j \subset \mathbb{V}, \mathbb{W}_j \subset \mathbb{W}$  be finite dimensional subspaces depending on an integer parameter  $j \geq 0$ , such that

$$\dim \mathbb{V}_j = \dim \mathbb{W}_j = n_j, \quad \mathbb{V}_j \subset \mathbb{V}_{j+1}, \quad \mathbb{W}_j \subset \mathbb{W}_{j+1}.$$

(Note that the notation has changed with respect to Section 3.1, where  $\mathbb{V}_N$  was a subspace of dimension  $N$ . Here  $\mathbb{V}_j$  may stand for  $\mathbb{V}_{\mathcal{T}_j}$ , where  $\mathcal{T}_j$  is the  $j$ -th mesh generated by an adaptive algorithm.)

We assume  $\mathcal{B}$  to satisfy a *uniform* discrete inf-sup condition on any product of subspaces  $\mathbb{V}_j \times \mathbb{W}_j$ , i.e., there exists a constant  $\beta > 0$  such that for all  $j$

$$\inf_{v \in \mathbb{V}_j} \sup_{w \in \mathbb{W}_j} \frac{\mathcal{B}[v, w]}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}} \geq \beta. \quad (10.2)$$

Let  $u_j \in \mathbb{V}_j$  be the solution of the (Petrov-)Galerkin problem

$$u_j \in \mathbb{V}_j : \quad \mathcal{B}[u_j, w] = \langle f, w \rangle \quad \forall w \in \mathbb{W}_j. \quad (10.3)$$

The first part of this section, which is mostly based on the recent work by M. Feischl [Feischl 2022], is devoted to studying the convergence of this approximation. Convergence and rate-optimality of different AFEMs will be discussed next in Section 10.3. Applications will be given to the Stokes problem (see Section 10.4) and the mixed formulation of a scalar diffusion problem (see Section 10.5).

### 10.1 Linear convergence of inf-sup stable methods

We make the following key assumptions that guarantee the convergence of the sequence  $u_j$  to  $u$  in the  $\mathbb{V}$ -norm, and comment about them afterwards. The first assumption is a relaxed form of the *general quasi-orthogonality* property introduced in [Carstensen *et al.* 2014] as part of an abstract set of axioms of adaptivity.

**Assumption 10.1 (relaxed quasi-orthogonality)** For each  $N \in \mathbb{N}$  there exists a nondecreasing constant  $C = C(N)$  such that

$$\sum_{k=j}^{j+N} \|u_{k+1} - u_k\|_{\mathbb{V}}^2 \leq C(N) \|u - u_j\|_{\mathbb{V}}^2 \quad j \geq 0, \quad (10.4)$$

and

$$C(N) = o(N), \quad \text{as } N \rightarrow \infty.$$

**Assumption 10.2 (equivalence of error and estimator)** There exist constants  $C_U \geq C_L > 0$  and, for each  $j \geq 0$ , an error estimator  $\eta_j = \eta_j(u_j)$ , such that

$$C_L \eta_j \leq \|u - u_j\|_{\mathbb{V}} \leq C_U \eta_j \quad j \geq 0. \quad (10.5)$$

**Assumption 10.3 (estimator reduction)** There exist constants  $0 < \rho_1 < 1$  and  $C_1 > 0$  such that

$$\eta_{j+1}^2 \leq \rho_1 \eta_j^2 + C_1 \|u_{j+1} - u_j\|_{\mathbb{V}}^2 \quad j \geq 0. \quad (10.6)$$

**Remark 10.4** Assumptions 10.2 and 10.3 are abstract and allow for a general convergence theory. In the context of our model problems of Section 2.3, they are valid for discrete data, i.e., if the coefficients of the linear operator corresponding to the bilinear form  $\mathcal{B}$  are piecewise polynomials on the adopted meshes, and if  $f \in \mathbb{F}_{\mathcal{T}}$  (see Section 4.3). We make this concrete in Sections 10.4 and 10.5 below.

**Remark 10.5** We comment on the significance of Assumption 10.1 upon considering two extreme cases.

1. Assumption 10.1 with  $C(N) = O(1)$  is precisely the general quasi-orthogonality property of [Carstensen *et al.* 2014]. It is valid with  $C(N) = 1$  for  $\mathbb{V} = \mathbb{W}$  and  $\mathcal{B}$  symmetric and coercive. Indeed,

$$\mathcal{B}[u_{k+1} - u_k, u - u_{k+1}] = 0 \quad (\text{Galerkin orthogonality}),$$

whence

$$\|u_{k+1} - u_k\|_{\Omega}^2 + \|u - u_{k+1}\|_{\Omega}^2 = \|u - u_k\|_{\Omega}^2,$$

where  $\|\cdot\|_{\Omega}$  is the energy norm induced by  $\mathcal{B}$ . Adding upon  $k$  and using telescopic



cancellation yields

$$\begin{aligned} \sum_{k=j}^{j+N} \|u_{k+1} - u_k\|_{\Omega}^2 &= \sum_{k=j}^{j+N} \|u - u_k\|_{\Omega}^2 - \|u - u_{k+1}\|_{\Omega}^2 \\ &= \|u - u_j\|_{\Omega}^2 - \|u - u_{j+N+1}\|_{\Omega}^2 \leq \|u - u_j\|_{\Omega}^2. \end{aligned}$$

Finally, the equivalence (2.30) of the norms  $\|\cdot\|_{\mathbb{V}}$  and  $\|\cdot\|_{\Omega}$  yields the result.

2. Assumption 10.1 trivially holds with  $C(N) = O(N)$  for  $\mathcal{B}$  continuous and inf-sup stable. Indeed, choosing in Corollary 3.3 (quasi-monotonicity)  $\mathbb{V}_N = \mathbb{V}_k$  or  $\mathbb{V}_{k+1}$  and  $\mathbb{V}_M = \mathbb{V}_j$  for  $j \leq k$ , and using the triangle inequality gives

$$\|u_{k+1} - u_k\|_{\mathbb{V}}^2 \lesssim \|u_{k+1} - u\|_{\mathbb{V}}^2 + \|u_k - u\|_{\mathbb{V}}^2 \lesssim \|u - u_j\|_{\mathbb{V}}^2.$$

Adding, we get

$$\sum_{k=j}^{j+N} \|u_{k+1} - u_k\|_{\mathbb{V}}^2 \leq C \sum_{k=j}^{j+N} \|u - u_j\|_{\mathbb{V}}^2 = CN \|u - u_j\|_{\mathbb{V}}^2.$$

However, the relation  $C(N) = O(N)$  is not enough for the subsequent analysis. In fact, we need  $C(N) = o(N)$ .

We now prove that the stated assumptions guarantee the linear convergence of the sequence of Petrov-Galerkin solutions (10.3). This result is similar to the convergence result for the estimators given in [Feischl 2022], and exploits the equivalence (10.5) between errors and estimators.

**Theorem 10.6 (linear convergence)** *Under Assumptions 10.1, 10.2 and 10.3, the discretization (10.3) is convergent; precisely, there exist constants  $0 < \rho < 1$  and  $c > 0$  such that*

$$e_{j+i} \leq c\rho^i e_j \quad \forall i, j \in \mathbb{N}, \quad (10.7)$$

where  $e_j := \|u - u_j\|_{\mathbb{V}}$ .

*Proof.* The proof is divided into several steps. Firstly, we set

$$E_k := \|u_k - u_{k-1}\|_{\mathbb{V}}.$$

□ We start by iterating (10.6)  $1 \leq n \leq k$  times to obtain

$$\begin{aligned} \eta_k^2 &\leq \rho_1 \eta_{k-1}^2 + C_1 E_k^2 \\ &\leq \rho_1 (\rho_1 \eta_{k-2}^2 + C_1 E_{k-1}^2) + C_1 E_k^2 \leq \rho_1^2 \eta_{k-2}^2 + C_1 (E_k^2 + E_{k-1}^2) \\ &\leq \rho_1^n \eta_{k-n}^2 + C_1 \sum_{\ell=k-n+1}^k E_{\ell}^2. \end{aligned}$$

We now invoke Assumption 10.2 to state the upper bound

$$e_k^2 \leq c_1 \eta_k^2$$

and the lower bound

$$\eta_k^2 \leq c_2 e_k^2$$

(with  $c_1 = C_U^2$  and  $c_2 = C_L^{-2}$ ). This yields

$$e_k^2 \leq c_1 \eta_k^2 \leq c_1 c_2 \rho_1^n e_{k-n}^2 + c_1 C_1 \sum_{\ell=k-n+1}^k E_\ell^2. \quad (10.8)$$

Let  $n \in \mathbb{N}$  be sufficiently large such that

$$\rho_2 = c_1 c_2 \rho_1^n < 1,$$

and let us relabel  $c_1 C_1$  as  $C_1$  to get

$$e_k^2 \leq \rho_2 e_{k-n}^2 + C_1 \sum_{\ell=k-n+1}^k E_\ell^2. \quad (10.9)$$

This shows that the reduction property (10.6) of the estimator is valid for the error after  $n$  iterations. We cannot expect (10.9) to hold for  $e_k$  with  $n = 1$ , not even in the coercive case: see Example 5.7 (lack of strict monotonicity) for  $A = I$  and  $f = 1$ . It is thus convenient to rewrite (10.9) as follows:

$$e_{kn}^2 \leq \rho_2 e_{(k-1)n}^2 + C_1 \sum_{\ell=(k-1)n+1}^{kn} E_\ell^2. \quad (10.10)$$

□ Sum up (10.10) from  $k = j + 1$  to  $k = j + N$  to get

$$\sum_{k=j+1}^{j+N} e_{kn}^2 \leq \rho_2 \sum_{k=j+1}^{j+N} e_{(k-1)n}^2 + C_1 \sum_{\ell=jn+1}^{(j+N)n} E_\ell^2.$$

Using (10.4) we see that

$$\sum_{\ell=jn+1}^{(j+N)n} E_\ell^2 = \sum_{\ell=jn}^{(j+N)n-1} E_{\ell+1}^2 \leq C(Nn-1) e_{jn}^2 \leq C(Nn) e_{jn}^2,$$

whence

$$\begin{aligned} \sum_{k=j+1}^{j+N} e_{kn}^2 &\leq \rho_2 \sum_{k=j+1}^{j+N} e_{(k-1)n}^2 + C_1 C(Nn) e_{jn}^2 \\ &\leq \rho_2 \left( \sum_{k=j+1}^{j+N} e_{kn}^2 + e_{jn}^2 \right) + C_1 C(Nn) e_{jn}^2. \end{aligned}$$

This implies

$$(1 - \rho_2) \sum_{k=j+1}^{j+N} e_{kn}^2 \leq (\rho_2 + C_1 C(Nn)) e_{jn}^2,$$

or equivalently

$$\frac{1 - \rho_2}{\rho_2 + C_1 C(Nn)} \sum_{k=j+1}^{j+N} e_{kn}^2 \leq e_{jn}^2. \quad (10.11)$$

Let us add the quantity  $\sum_{k=j+1}^{j+N} e_{kn}^2$  to both sides to arrive at

$$\left(1 + \frac{1 - \rho_2}{\rho_2 + C_1 C(Nn)}\right) \sum_{k=j+1}^{j+N} e_{kn}^2 \leq \sum_{k=j}^{j+N} e_{kn}^2.$$

We can rewrite this inequality as follows:

$$\sum_{k=j+1}^{j+N} e_{kn}^2 \leq \rho(N) \sum_{k=j}^{j+N} e_{kn}^2, \quad (10.12)$$

where

$$\begin{aligned} \rho(N) &= \frac{1}{1 + \frac{1 - \rho_2}{\rho_2 + C_1 C(Nn)}} = \frac{\rho_2 + C_1 C(Nn)}{1 + C_1 C(Nn)} \\ &= 1 - \frac{1 - \rho_2}{1 + C_1 C(Nn)} = 1 - \frac{1}{D(N)} \end{aligned}$$

with

$$D(N) = \frac{1 + C_1 C(Nn)}{1 - \rho_2} \rightarrow \infty, \quad N \rightarrow \infty,$$

whenever  $C(N)$  diverges. Therefore, (10.12) is a contraction for the quantity  $\sum_{k=j+1}^{j+N} e_{kn}^2$  with a constant  $\rho(N)$  uniform in  $j$  that may degenerate to 1 as  $N \rightarrow \infty$ .

3 We iterate (10.12) and exploit that the left-hand side has one fewer term than the right-hand side. Take

$$j \rightarrow j + 1 \quad N \rightarrow N - 1$$

to get

$$\sum_{k=j+2}^{j+N} e_{kn}^2 \leq \rho(N-1) \sum_{k=j+1}^{j+N} e_{kn}^2,$$

whence

$$\sum_{k=j+2}^{j+N} e_{kn}^2 \leq \rho(N-1)\rho(N) \sum_{k=j}^{j+N} e_{kn}^2.$$

Iterating, we get

$$e_{(j+N)n}^2 = \sum_{k=j+N}^{j+N} e_{kn}^2 \leq \rho(1) \dots \rho(N-1)\rho(N) \sum_{k=j}^{j+N} e_{kn}^2.$$

We now need to bound the sum on the right-hand side by a single term. To this end, we resort to (10.11)

$$\sum_{k=j+1}^{j+N} e_{kn}^2 \leq \frac{\rho_2 + C_1 C(Nn)}{1 - \rho_2} e_{jn}^2$$

and add  $e_{jn}^2$  on both sides

$$\sum_{k=j}^{j+N} e_{kn}^2 \leq \left(1 + \frac{\rho_2 + C_1 C(Nn)}{1 - \rho_2}\right) e_{jn}^2 = D(N) e_{jn}^2.$$

Altogether, we arrive at

$$e_{(j+N)n}^2 \leq \rho(1) \dots \rho(N) D(N) e_{jn}^2 = D(N) \prod_{k=1}^N \left(1 - \frac{1}{D(k)}\right) e_{jn}^2.$$

4 We estimate the factor on the right-hand side. For  $N_0 > 0$  to be chosen later, set

$$\rho_0 := D(N_0) \prod_{k=1}^{N_0} \left(1 - \frac{1}{D(k)}\right)$$

and compute the logarithm of  $\rho_0$

$$\log(\rho_0) = \log(D(N_0)) + \sum_{k=1}^{N_0} \log\left(1 - \frac{1}{D(k)}\right) \leq \log(D(N_0)) - \sum_{k=1}^{N_0} \frac{1}{D(k)},$$

because the log is concave and  $\log(1+x) \leq x$ . Since we assume in (10.4) that  $D(k) \simeq C(kN_0) = o(k)$ , the series diverges and we see that

$$\log(\rho_0) < 0$$

for  $N_0$  sufficiently large. Summarizing, there exist  $N_0 > 0$  and  $0 < \rho_0 < 1$  such that

$$e_{(j+N_0)n}^2 \leq \rho_0 e_{jn}^2 \quad \forall j \in \mathbb{N}. \quad (10.13)$$

5 For any  $j, i \in \mathbb{N}$ , we now find  $c > 0$  and  $0 < \rho < 1$  such that the inequality

$$e_{j+i} \leq c \rho^i e_j$$

holds. We decompose  $j$  and  $j+i$  in terms of integers  $k, m$

$$\begin{aligned} j &= (k-1)n + \widehat{j}, & k \geq 1, & \quad 0 \leq \widehat{j} < n, \\ j+i &= (k+m)n + \widehat{i}, & m \geq -1, & \quad 0 \leq \widehat{i} < n, \end{aligned}$$

and examine first the case  $m \geq 0$ . We further decompose

$$m = aN_0 + b, \quad a, b \in \mathbb{N}, \quad 0 \leq b < N_0 \quad \Rightarrow \quad a = \frac{m}{N_0} - \frac{b}{N_0}.$$

Note that

$$kn = j - \hat{j} + n > j, \quad m = \frac{i}{n} - \left( \frac{\hat{i} - \hat{j}}{n} + 1 \right), \quad a > \frac{i}{nN_0} - \frac{2 + N_0}{N_0}.$$

Therefore, invoking Corollary 3.3 (quasi-monotonicity)

$$e_{j_2} \leq \frac{\|\mathcal{B}\|}{\beta} e_{j_1} = C_* e_{j_1} \quad j_2 \geq j_1 \geq 0,$$

in conjunction with (10.13), yields

$$e_{j+i} \leq C_* e_{(k+aN_0)n} \leq C_* \rho_0^{\frac{a}{2}} e_{kn} \leq C_*^2 \rho_0^{\frac{a}{2}} e_j < C_*^2 \rho_0^{-\frac{2+N_0}{2N_0}} (\rho_0^{\frac{1}{2nN_0}})^i e_j.$$

This is the desired estimate with  $c = C_*^2 \rho_0^{-\frac{2+N_0}{2N_0}}$  and  $\rho = \rho_0^{\frac{1}{2nN_0}}$  for  $m \geq 0$ . We finally consider  $m = -1$  and use again the error quasi-monotonicity to write

$$e_{j+i} \leq C_* e_j = \frac{C_*}{\rho^i} \rho^i e_j < \frac{C_*}{\rho^n} \rho^i e_j.$$

This concludes the proof with  $c = \max \left\{ C_*^2 \rho_0^{-\frac{2+N_0}{2N_0}}, \frac{C_*}{\rho^n} \right\}$ .  $\square$

**Remark 10.7 (improving on Assumption 10.1)** It is worth underlining that, while linear convergence (10.7) is established in Theorem 10.6 using Assumption 10.1, the same property combined with uniform inf-sup stability tell us a posteriori that the constant  $C(N)$  in (10.4) can be made independent of  $N$ , i.e.  $C(N) = O(1)$ . To see this, we apply the linear convergence bound

$$\|u - u_k\|_{\mathbb{V}} \leq c \rho^{k-j} \|u - u_j\|_{\mathbb{V}}$$

in conjunction with the triangle inequality

$$\|u_{k+1} - u_k\|_{\mathbb{V}} \leq \|u - u_{k+1}\|_{\mathbb{V}} + \|u - u_k\|_{\mathbb{V}} \leq 2c \rho^{k-j} \|u - u_j\|_{\mathbb{V}};$$

summation of a geometric series gives

$$\sum_{k=j}^{j+N} \|u_{k+1} - u_k\|_{\mathbb{V}}^2 \leq C \|u - u_j\|_{\mathbb{V}}^2$$

with  $C = 4c^2 \sum_{\ell=0}^{\infty} \rho^{2\ell} < +\infty$ . This suggests that Assumption 10.1 might be too pessimistic.

## 10.2 Inf-sup stability implies quasi-orthogonality

We aim at proving the following key result in this section.

**Theorem 10.8 (sufficient condition for Assumption 10.1)** *The Assumption 10.1 (relaxed quasi-orthogonality) is valid if the bilinear form  $\mathcal{B} : \mathbb{V} \times \mathbb{W} \rightarrow \mathbb{R}$  is continuous and uniformly inf-sup stable on the sequence of subspaces  $\mathbb{V}_j \times \mathbb{W}_j$ ,  $j \geq 0$ .*

To accomplish this task, we proceed in two steps. Using variational techniques, we first establish an intermediate result formally similar to (10.4) (see Corollary 10.14), but involving the norm of a matrix  $\mathbf{U}$  related to the form  $\mathcal{B}$ . Next, we rely on algebraic techniques to estimate such norm (see Theorem 10.15) and close the proof of the desired result.

In order to perform the first step, we introduce orthonormal bases of the finite element spaces  $\mathbb{V}_j, \mathbb{W}_j, 0 \leq j \leq N$ , and next we biorthogonalize them. This procedure turns out to be crucial.

We start with some notation. Let

$$n_j = \dim \mathbb{V}_j = \dim \mathbb{W}_j.$$

Let  $\mathbb{V}_{j-1}^\perp$  and  $\mathbb{W}_{j-1}^\perp$  denote the orthogonal complements of  $\mathbb{V}_{j-1}$  and  $\mathbb{W}_{j-1}$  within  $\mathbb{V}_j$  and  $\mathbb{W}_j$ , respectively. Let

$$d_j = n_j - n_{j-1} = \dim \mathbb{V}_{j-1}^\perp = \dim \mathbb{W}_{j-1}^\perp$$

be the dimension of the Galerkin update to augment the space  $\mathbb{V}_{j-1}$  into the next space  $\mathbb{V}_j$ , and likewise with the space  $\mathbb{W}_{j-1}$  and  $\mathbb{W}_j$ .

We consider orthonormal bases

$$\mathbf{v} = \{\mathbf{v}(j)\}_{j=0}^N \subset \mathbb{V}_N, \quad \mathbf{w} = \{\mathbf{w}(i)\}_{i=0}^N \subset \mathbb{W}_N \quad (10.14)$$

partitioned into blocks for  $1 \leq j \leq N$

$$\mathbf{v}(j) = (v_k)_{k=n_{j-1}+1}^{n_j} \subset \mathbb{V}_{j-1}^\perp, \quad \mathbf{w}(i) = (w_k)_{k=n_{i-1}+1}^{n_i} \subset \mathbb{W}_{i-1}^\perp, \quad (10.15)$$

and  $\mathbf{v}(0) \subset \mathbb{V}_0, \mathbf{w}(0) \subset \mathbb{W}_0$ . In other words,  $(\mathbf{v}(j), \mathbf{w}(j))$  represent the  $d_j$  new directions added by Galerkin to the current spaces  $(\mathbb{V}_{j-1}, \mathbb{W}_{j-1})$  for  $1 \leq j \leq N$ .

We recall that the bilinear form  $\mathcal{B} : \mathbb{V}_N \times \mathbb{W}_N \rightarrow \mathbb{R}$  satisfies the following uniform properties for all  $0 \leq j \leq N$

(P1) *continuity*:

$$|\mathcal{B}[\mathbf{v}, \mathbf{w}]| \leq \|\mathcal{B}\| \|\mathbf{v}\|_{\mathbb{V}} \|\mathbf{w}\|_{\mathbb{W}} \quad \forall \mathbf{v} \in \mathbb{V}_j, \mathbf{w} \in \mathbb{W}_j; \quad (10.16)$$

(P2) *inf-sup condition*:

$$\beta \|\mathbf{v}\|_{\mathbb{V}} \leq \sup_{\mathbf{w} \in \mathbb{W}_j} \frac{\mathcal{B}[\mathbf{v}, \mathbf{w}]}{\|\mathbf{w}\|_{\mathbb{W}}} \quad \forall \mathbf{v} \in \mathbb{V}_j. \quad (10.17)$$

The block bases  $\mathbf{v}$  and  $\mathbf{w}$  given in (10.14) induce a block matrix

$$\mathbf{B} := (\mathbf{B}(i, j))_{i,j=0}^N \in \mathbb{R}^{n_N \times n_N}$$

defined by

$$\mathbf{B}(i, j) = \mathcal{B}[\mathbf{v}(j), \mathbf{w}(i)]. \quad (10.18)$$

Note that the actual size of  $\mathbf{B}$  is  $n_N = \dim \mathbb{V}_N \gg N$ , and that the following analysis entails expressing important quantities in terms of the number of blocks  $N$  rather than the dimension  $n_N$ .

We will use this block decomposition for a generic matrix

$$\mathbf{M} = (\mathbf{M}(i, j))_{i,j=0}^N \in \mathbb{R}^{n_N \times n_N}$$

and we denote by  $\mathbf{M}[k] = (\mathbf{M}(i, j))_{i,j=0}^k$  the *principal  $k$ -th block* of  $\mathbf{M}$ . Fig. 10.1 shows schematically what this means.

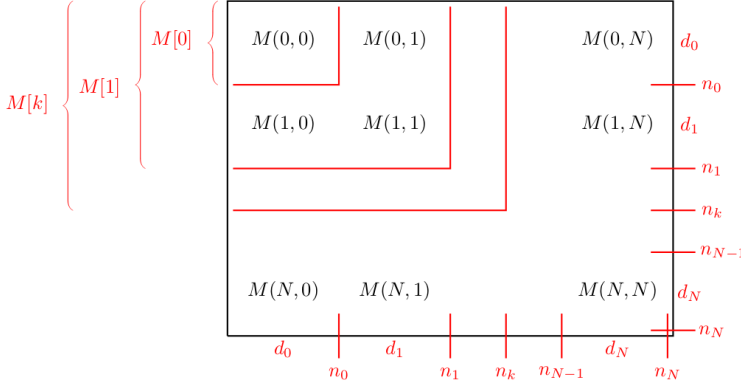


Figure 10.1. Block partition of a matrix  $\mathbf{M} \in \mathbb{R}^{n_N \times n_N}$  with  $(N+1) \times (N+1)$  blocks  $\mathbf{M}(i, j) \in \mathbb{R}^{d_i \times d_j}$  and principal  $k$ -th block  $\mathbf{M}[k] \in \mathbb{R}^{n_k \times n_k}$  with  $0 \leq i, j, k \leq N$ .

We stress that (P2) implies that  $\mathbf{B}[k]$  is uniformly invertible with

$$\|\mathbf{B}[k]^{-1}\|_2 \leq \frac{1}{\beta} \quad \forall 0 \leq k \leq N. \quad (10.19)$$

In fact, (P2) with  $j$  replaced by  $k$  can be rephrased as follows in terms of the coordinates  $\mathbf{v} \in \mathbb{R}^{n_k}$  relative to the orthonormal basis  $\{\mathbf{v}(j)\}_{j=0}^k$  of  $\mathbb{V}_k$  of a generic vector in  $\mathbb{V}_k$

$$\beta \|\mathbf{v}\|_2 \leq \|\mathbf{B}[k]\mathbf{v}\|_2 \quad \forall \mathbf{v} \in \mathbb{R}^{n_k},$$

i.e., setting  $\mathbf{z} = \mathbf{B}[k]\mathbf{v}$ ,

$$\beta \|\mathbf{B}[k]^{-1}\mathbf{z}\|_2 \leq \|\mathbf{z}\|_2 \quad \forall \mathbf{z} \in \mathbb{R}^{n_k},$$

which is precisely (10.19).

A fundamental linear algebra theorem of Gaussian elimination guarantees the existence of a unique normalized block  $\mathbf{LU}$  decomposition of  $\mathbf{B}$  without pivoting due to (10.19):

$$\mathbf{B} = \mathbf{LU}, \quad (10.20)$$

with block partitioning

$$\mathbf{L}(i, j) \in \mathbb{R}^{d_i \times d_j}, \quad \mathbf{L}(i, j) = \mathbf{0} \text{ for } j > i, \quad \mathbf{L}(i, i) = \mathbf{I}(i, i); \quad (10.21)$$

$$\mathbf{U}(i, j) \in \mathbb{R}^{d_i \times d_j}, \quad \mathbf{U}(i, j) = \mathbf{0} \text{ for } i > j. \quad (10.22)$$

### 10.2.1 Matrix representation

The  $k$ -th Galerkin solution  $u_k$  satisfies:

$$u_k \in \mathbb{V}_k : \quad \mathcal{B}[u_k, w] = \langle f, w \rangle, \quad \forall w \in \mathbb{W}_k.$$

Equivalently, if  $\{\gamma(j)\}_{j=0}^k \in \mathbb{R}^{n_k}$  are the coordinates of  $u_k$  with respect to the orthonormal basis  $\{v(j)\}_{j=0}^k$

$$u_k = \sum_{j=0}^k \gamma(j) \cdot v(j),$$

then

$$\sum_{j=0}^k \gamma(j) \cdot \mathcal{B}[v(j), w(i)] = \langle f, w(i) \rangle, \quad \forall 0 \leq i \leq k,$$

or using matrix notation

$$\sum_{j=0}^k \mathbf{B}(i, j) \gamma(j) = \mathbf{f}(i) = \langle f, w(i) \rangle, \quad \forall 0 \leq i \leq k. \quad (10.23)$$

If we further write

$$\gamma_k = (\gamma(j))_{j=0}^k \in \mathbb{R}^{n_k}, \quad \mathbf{f}_k = (\mathbf{f}(i))_{i=0}^k \in \mathbb{R}^{n_k},$$

then (10.23) reduces to

$$\mathbf{B}[k] \gamma_k = \mathbf{f}_k. \quad (10.24)$$

In view of the definition of  $\mathbf{f}_k$  we realize that the  $k$ -th section  $\mathbf{f}_N[k] \in \mathbb{R}^{n_k}$  of  $\mathbf{f}_N$  coincides with  $\mathbf{f}_k$ :

$$\mathbf{f}_N[k] = (\mathbf{f}(i))_{i=0}^k = \mathbf{f}_k.$$

However, this statement is *not* true for the solution  $\gamma_k$  of (10.24), namely

$$\gamma_N[k] \neq \gamma_k.$$

### 10.2.2 Block biorthogonal bases

We define biorthogonal bases  $\tilde{\mathbf{v}} \subset \mathbb{V}_N$  and  $\tilde{\mathbf{w}} \subset \mathbb{W}_N$  as follows:

$$\tilde{\mathbf{v}} := \mathbf{U}^{-T} \mathbf{v} \implies \tilde{\mathbf{v}}(j) = \sum_{m=0}^j \mathbf{U}^{-T}(j, m) \mathbf{v}(m) \quad 0 \leq j \leq N, \quad (10.25)$$

$$\tilde{\mathbf{w}} := \mathbf{L}^{-1} \mathbf{w} \implies \tilde{\mathbf{w}}(i) = \sum_{m=0}^i \mathbf{L}^{-1}(i, m) \mathbf{w}(m) \quad 0 \leq i \leq N. \quad (10.26)$$



We will see below that these bases are convenient to represent the Galerkin solution  $u_k \in \mathbb{V}_k$ . We start with a list of properties.

**Lemma 10.9 (span of new bases)** *The vectors  $\tilde{\mathbf{v}}$  and  $\tilde{\mathbf{w}}$  are bases of  $\mathbb{V}_N$  and  $\mathbb{W}_N$ , respectively, and satisfy*

$$\text{span}\{\tilde{\mathbf{v}}(j)\}_{j=0}^k = \text{span}\{\mathbf{v}(j)\}_{j=0}^k,$$

$$\text{span}\{\tilde{\mathbf{w}}(i)\}_{i=0}^k = \text{span}\{\mathbf{w}(i)\}_{i=0}^k.$$

*Proof.* This relies on the fact that  $\mathbf{U}^{-T}$  and  $\mathbf{L}^{-1}$  are lower triangular and the diagonal blocks are non-singular (i.e., both  $\mathbf{L}$  and  $\mathbf{U}$  are invertible).  $\square$

Consider now the matrix  $\tilde{\mathbf{B}}$  induced by  $(\tilde{\mathbf{v}}, \tilde{\mathbf{w}})$ , namely

$$\tilde{\mathbf{B}} := \mathcal{B}[\tilde{\mathbf{v}}, \tilde{\mathbf{w}}] \in \mathbb{R}^{n_N \times n_N}. \quad (10.27)$$

**Lemma 10.10 (biorthogonality)** *The block matrix  $\tilde{\mathbf{B}}$  is equal to the identity, namely*

$$\tilde{\mathbf{B}}(i, j) = \mathbf{I}(i, j) \quad \forall 0 \leq i, j \leq N.$$

*Proof.* We simply combine the definition (10.27) with (10.25) and (10.26) to deduce, for all  $0 \leq i, j \leq N$ , that

$$\begin{aligned} \tilde{\mathbf{B}}(i, j) &= \mathcal{B}[\tilde{\mathbf{v}}(j), \tilde{\mathbf{w}}(i)] \\ &= \mathcal{B}\left[\sum_{m=0}^j \mathbf{U}^{-T}(j, m) \mathbf{v}(m), \sum_{k=0}^i \mathbf{L}^{-1}(i, k) \mathbf{w}(k)\right] \\ &= \sum_{m=0}^j \sum_{k=0}^i \mathbf{L}^{-1}(i, k) \mathcal{B}[\mathbf{v}(m), \mathbf{w}(k)] \mathbf{U}^{-T}(j, m) \\ &= \sum_{m=0}^j \sum_{k=0}^i \mathbf{L}^{-1}(i, k) \mathbf{B}(k, m) \mathbf{U}^{-1}(m, j) \\ &= (\mathbf{L}^{-1} \mathbf{B} \mathbf{U}^{-T})(i, j) \\ &= (\mathbf{L}^{-1} (\mathbf{L} \mathbf{U}) \mathbf{U}^{-1})(i, j) \\ &= \mathbf{I}(i, j), \end{aligned}$$

as asserted.  $\square$

Generic functions  $v \in \mathbb{V}_N$  and  $w \in \mathbb{W}_N$  can be represented as follows in terms of the old and new bases:

$$v = \sum_{j=0}^N \gamma(j) \cdot \mathbf{v}(j) = \sum_{j=0}^N \tilde{\gamma}(j) \cdot \tilde{\mathbf{v}}(j), \quad (10.28)$$

$$w = \sum_{i=0}^N \alpha(i) \cdot w(i) = \sum_{i=0}^N \tilde{\alpha}(i) \cdot \tilde{w}(i). \quad (10.29)$$

The following lemma relates the coordinates in the two systems.

**Lemma 10.11 (change of basis)** *The coordinates  $\gamma = (\gamma(j))_{j=0}^N$  and  $\alpha = (\alpha(i))_{i=0}^N$  satisfy*

$$\gamma = U^{-1} \tilde{\gamma}, \quad \alpha = L^{-T} \tilde{\alpha}. \quad (10.30)$$

*Proof.* Write (10.28) in vector form and use (10.25) to obtain

$$v = \tilde{\gamma}^T \tilde{v} = \tilde{\gamma}^T (U^{-T} v) = (U^{-1} \tilde{\gamma})^T v = \gamma^T v,$$

whence

$$\gamma = U^{-1} \tilde{\gamma}.$$

Similarly, combining (10.29) and (10.26) yields

$$w = \tilde{\alpha}^T \tilde{w} = \tilde{\alpha}^T (L^{-1} w) = (L^{-T} \tilde{\alpha})^T w = \alpha^T w$$

and

$$\alpha = L^{-T} \tilde{\alpha}.$$

This concludes the proof.  $\square$

### 10.2.3 An intermediate inequality

We intend to prove the following crucial estimate. This result is in [Feischl 2022], but we give a different proof based on variational arguments.

**Proposition 10.12 (quasi-orthogonality - I)** *If  $u_k \in \mathbb{V}_k$  denotes the  $k$ -th Galerkin solution of (10.3), then there holds*

$$\sum_{k=0}^{N-1} \|u_{k+1} - u_k\|_{\mathbb{V}}^2 \leq \frac{\|U\|_2^2}{\beta^2} \|u_N - u_0\|_{\mathbb{V}}^2. \quad (10.31)$$

*Proof.* We proceed in several steps.

$\square$  *Estimate of  $\|u_{k+1} - u_k\|_{\mathbb{V}}$ .* Galerkin orthogonality yields

$$\mathcal{B}[u_{k+1} - u_k, w] = 0 \quad \forall w \in \mathbb{W}_k. \quad (10.32)$$

The uniform discrete inf-sup property (P2) implies the existence of  $\bar{w} \in \mathbb{W}_{k+1}$  with  $\|\bar{w}\|_{\mathbb{W}} = 1$  such that

$$\beta \|u_{k+1} - u_k\|_{\mathbb{V}} \leq \mathcal{B}[u_{k+1} - u_k, \bar{w}]. \quad (10.33)$$

We decompose  $\bar{w}$  orthogonally as follows:

$$\bar{w} = \bar{w}_k + \bar{w}_k^\perp, \quad \bar{w}_k \in \mathbb{W}_k, \quad \bar{w}_k^\perp \in \mathbb{W}_k^\perp,$$

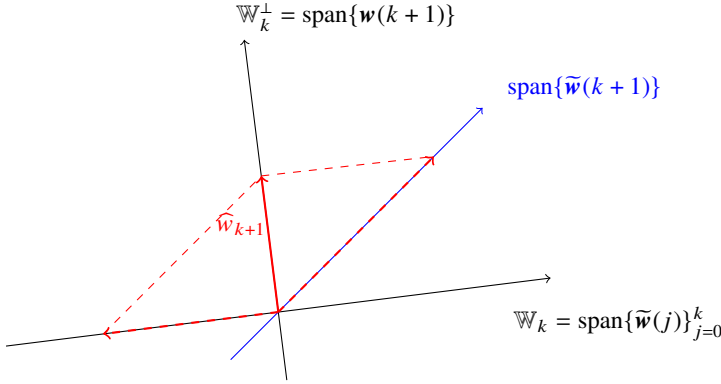


Figure 10.2. Oblique decomposition of the space  $\mathbb{W}_{k+1}$  into the subspaces  $\mathbb{W}_k = \text{span}\{\tilde{\mathbf{w}}(j)\}_{j=0}^k$  and  $\text{span}\{\tilde{\mathbf{w}}(k+1)\}$ .

with  $\|\bar{\mathbf{w}}_k^\perp\|_{\mathbb{W}} \leq 1$ . In view of (10.32), (10.33) also reads

$$\beta \|u_{k+1} - u_k\|_{\mathbb{V}} \leq \mathcal{B}[u_{k+1} - u_k, \bar{\mathbf{w}}_k^\perp] \leq \mathcal{B}[u_{k+1} - u_k, \frac{\bar{\mathbf{w}}_k^\perp}{\|\bar{\mathbf{w}}_k^\perp\|_{\mathbb{W}}}] .$$

We now let  $\hat{\mathbf{w}}_{k+1} := \frac{\bar{\mathbf{w}}_k^\perp}{\|\bar{\mathbf{w}}_k^\perp\|_{\mathbb{W}}} \in \mathbb{W}_k^\perp \subset \mathbb{W}_{k+1}$  and decompose it along the oblique subspaces  $\mathbb{W}_k = \text{span}\{\tilde{\mathbf{w}}(j)\}_{j=0}^k$  and  $\text{span}\{\tilde{\mathbf{w}}(k+1)\}$ , as illustrated in Fig. 10.2. Since  $\mathbb{W}_k^\perp = \text{span}\{\mathbf{w}(k+1)\}$  and  $\mathbf{w}(k+1) = (\mathbf{w}_j)_{j=n_{k+1}}^{n_{k+1}+1}$  is an orthonormal basis, the function  $\hat{\mathbf{w}}_{k+1} \in \mathbb{W}_k^\perp$  can be written uniquely as

$$\hat{\mathbf{w}}_{k+1} = \alpha(k+1) \cdot \mathbf{w}(k+1)$$

with  $\alpha(k+1) \in \mathbb{R}^{d_{k+1}}$  satisfying

$$\|\alpha(k+1)\|_2 = 1 = \|\hat{\mathbf{w}}_{k+1}\|_{\mathbb{W}} .$$

Invoking (10.26), we can express  $\mathbf{w}(k+1)$  in terms of  $\{\tilde{\mathbf{w}}(j)\}_{j=0}^k$  as follows:

$$\begin{aligned} \mathbf{w}(k+1) &= \sum_{j=0}^{k+1} \mathbf{L}(k+1, j) \tilde{\mathbf{w}}(j) \\ &= \tilde{\mathbf{w}}(k+1) + \sum_{j=0}^k \mathbf{L}(k+1, j) \tilde{\mathbf{w}}(j) \end{aligned}$$

because  $\mathbf{L}(k+1, k+1) = \mathbf{I}(k+1, k+1) \in \mathbb{R}^{d_{k+1} \times d_{k+1}}$ . Consequently

$$\mathcal{B}[u_{k+1} - u_k, \hat{\mathbf{w}}_{k+1}] = \mathcal{B}[u_{k+1} - u_k, \alpha(k+1) \cdot \tilde{\mathbf{w}}(k+1)]$$

because (10.32) implies

$$\mathcal{B}[u_{k+1} - u_k, \tilde{\mathbf{w}}(j)] = 0 \quad \forall 0 \leq j \leq k .$$

In addition, the biorthogonality of  $\tilde{\mathbf{w}}(k+1)$  with respect to  $\tilde{\mathbf{v}}(j)$  for  $0 \leq j \leq k$  translates into

$$\mathcal{B}[u_k, \tilde{\mathbf{w}}(k+1)] = 0 = \mathcal{B}[u_0, \tilde{\mathbf{w}}(k+1)].$$

Moreover, Galerkin orthogonality yields

$$\mathcal{B}[u_{k+1}, \tilde{\mathbf{w}}_{k+1}] = \langle f, \tilde{\mathbf{w}}_{k+1} \rangle = \mathcal{B}[u_N, \tilde{\mathbf{w}}_{k+1}],$$

and collecting the preceding expressions we obtain

$$\|u_{k+1} - u_k\| \leq \frac{1}{\beta} \alpha(k+1) \cdot \mathcal{B}[u_N - u_0, \tilde{\mathbf{w}}(k+1)]. \quad (10.34)$$

2 *Estimate of  $\mathcal{B}[u_N - u_0, \tilde{\mathbf{w}}(k+1)]$ .* We exploit the biorthogonality between  $\{\tilde{\mathbf{v}}(j)\}_{j=0}^N$  and  $\{\tilde{\mathbf{w}}(j)\}_{j=0}^N$ . In fact, we write

$$u_N - u_0 = \sum_{j=0}^N \tilde{\gamma}(j) \cdot \tilde{\mathbf{v}}(j)$$

and substitute into the right-hand side of (10.34) to arrive at

$$\mathcal{B}[u_N - u_0, \tilde{\mathbf{w}}(k+1)] = \sum_{j=0}^N \tilde{\gamma}(j) \cdot \mathcal{B}[\tilde{\mathbf{v}}(j), \tilde{\mathbf{w}}(k+1)] = \tilde{\gamma}(k+1).$$

Therefore, (10.34) gives

$$\|u_{k+1} - u_k\|_{\mathbb{V}} \leq \frac{1}{\beta} \alpha(k+1) \cdot \tilde{\gamma}(k+1),$$

whence

$$\|u_{k+1} - u_k\|_{\mathbb{V}} \leq \frac{1}{\beta} \|\tilde{\gamma}(k+1)\|_2$$

because  $\|\alpha(k+1)\|_2 = 1$ .

3 *Final estimate.* Compute

$$\sum_{k=0}^{N-1} \|u_{k+1} - u_k\|_{\mathbb{V}}^2 \leq \frac{1}{\beta^2} \sum_{k=0}^{N-1} \|\tilde{\gamma}(k+1)\|_2^2 \leq \frac{1}{\beta^2} \|\tilde{\gamma}\|_2^2 \leq \frac{\|\mathbf{U}\|_2^2}{\beta^2} \|\gamma\|_2^2,$$

according to (10.30). Since  $\{\mathbf{v}(j)\}_{j=0}^N$  are orthonormal, we get

$$u_N - u_0 = \sum_{j=0}^N \gamma(j) \cdot \mathbf{v}(j) \quad \Rightarrow \quad \|u_N - u_0\|_{\mathbb{V}} = \|\gamma\|_2$$

and

$$\sum_{k=0}^{N-1} \|u_{k+1} - u_k\|_{\mathbb{V}}^2 \leq \frac{\|\mathbf{U}\|_2^2}{\beta^2} \|u_N - u_0\|_{\mathbb{V}}^2,$$

as asserted. This concludes the proof.  $\square$

In order to get the quasi-orthogonality estimate, we still need to compare the errors  $\|u_N - u_0\|_{\mathbb{V}}$  and  $\|u - u_0\|_{\mathbb{V}}$ . The following is a variant of (3.3).

**Lemma 10.13 (stability)** *There holds*

$$\|u_N - u_0\|_{\mathbb{V}} \leq \frac{\|\mathcal{B}\|}{\beta} \|u - u_0\|_{\mathbb{V}}.$$

*Proof.* We use (10.17) and (10.16), in this order, together with Galerkin orthogonality to deduce

$$\begin{aligned} \beta \|u_N - u_0\|_{\mathbb{V}} &\leq \sup_{w \in \mathbb{W}_N} \frac{\mathcal{B}[u_N - u_0, w]}{\|w\|_{\mathbb{W}}} \\ &= \sup_{w \in \mathbb{W}_N} \frac{\mathcal{B}[u - u_0, w]}{\|w\|_{\mathbb{W}}} \leq \|\mathcal{B}\| \|u - u_0\|_{\mathbb{V}}. \end{aligned}$$

This completes the proof.  $\square$

**Corollary 10.14 (quasi-orthogonality - II)** *Let  $u_k \in \mathbb{V}_k$  be the  $k$ -th Galerkin solution of (10.3). Then, for all  $0 \leq j \leq N$ , we have*

$$\sum_{k=j}^{j+N-1} \|u_{k+1} - u_k\|_{\mathbb{V}}^2 \leq \frac{\|\mathcal{B}\|^2}{\beta^4} \|\mathbf{U}\|_2^2 \|u - u_j\|_{\mathbb{V}}^2.$$

*Proof.* Combining Proposition 10.12 with Lemma 10.13 yields

$$\sum_{k=0}^{N-1} \|u_{k+1} - u_k\|_{\mathbb{V}}^2 \leq \frac{\|\mathcal{B}\|^2}{\beta^4} \|\mathbf{U}\|_2^2 \|u - u_0\|_{\mathbb{V}}^2.$$

Finally, replacing  $u_0 \in \mathbb{V}_0$  by the  $j$ -th Galerkin solution  $u_j \in \mathbb{V}_j$ , we obtain the desired estimate.  $\square$

This corollary says that in order to prove Theorem 10.8, i.e., to check the validity of Assumption 10.1, it is enough to investigate the growth of the block triangular factor  $\mathbf{U}$  introduced in (10.20), and precisely to prove that

$$\|\mathbf{U}\|_2 = o(N^{1/2}).$$

This is the second step of our analysis. Actually, we prove something more, which is expressed by the following result.

**Theorem 10.15 (bound of block matrices  $\mathbf{L}$  and  $\mathbf{U}$ )** *There exist constants  $C_{LU} > 0$  and  $p > 2$  such that the block  $\mathbf{LU}$  factors of  $\mathbf{B}$  satisfy*

$$\|\mathbf{U}\|_2 + \|\mathbf{L}\|_2 + \|\mathbf{U}^{-1}\|_2 + \|\mathbf{L}^{-1}\|_2 \leq C_{LU} N^{1/p}. \quad (10.35)$$

The proof of this theorem is lengthy and very technical; it involves subtle linear algebra arguments, which may not be familiar to many readers. For such reasons, we prefer to postpone it to the end of this section (see Section 10.6).

### 10.3 Convergence rates of AFEMs for inf-sup stable methods

In this section, we discuss AFEMs to solve a boundary-value problem admitting a variational formulation of the form

$$u \in \mathbb{V} : \quad \mathcal{B}[u, w] = \langle f, w \rangle \quad \forall w \in \mathbb{W}, \quad (10.36)$$

in which the bilinear form  $\mathcal{B}$  on  $\mathbb{V} \times \mathbb{W}$  is continuous and inf-sup stable, with inf-sup constant  $\beta > 0$ , and  $f \in \mathbb{W}^*$ . We will consider the one-step AFEM given by Algorithm 5.4 (GALERKIN) when all data are discrete, the one-step AFEM with switch given by Algorithm 5.16 (AFEM-SW) when the operator coefficients are discrete but the forcing term is not (as in the Stokes problem), and the general two-step AFEM given by Algorithm 5.1 (AFEM-TS).

#### 10.3.1 Algorithm 5.4 (GALERKIN)

For  $j \geq 0$ , let us denote by  $(\mathcal{T}_j, \mathbb{V}_j, u_j)$  with  $u_j \in \mathbb{V}_j = \mathbb{V}_{\mathcal{T}_j}$ , the sequence of meshes, subspaces and Galerkin approximations to (10.36) generated by GALERKIN. Let

$$\eta_{\mathcal{T}_j}(v) = \eta_{\mathcal{T}_j}(v, f) = \left( \sum_{T \in \mathcal{T}_j} \eta_T(v, T)^2 \right)^{1/2} \quad (10.37)$$

be the PDE error estimators used in the loop. If such estimators fulfil Assumptions 10.2 (equivalence of error and estimator) and 10.3 (estimator reduction), then Theorem 10.6 (linear convergence) applies and the following result holds.

**Proposition 10.16 (convergence and termination of GALERKIN)** *The module GALERKIN produces a sequence  $\{u_j\}$  converging linearly to  $u \in \mathbb{V}$*

$$\|u - u_{j+i}\|_{\mathbb{V}} \leq C \rho^i \|u - u_j\|_{\mathbb{V}} \quad \forall j, i \geq 0, \quad 0 < \rho < 1,$$

*thereby reaching any prescribed accuracy  $\|u - u_j\|_{\mathbb{V}} \leq \varepsilon$  in a finite number of iterations.*

#### 10.3.2 Algorithm 5.16 (AFEM-SW).

This algorithm applies to the situation in which the operator coefficients are discrete, whereas the forcing  $f \in \mathbb{W}^*$  is not. Then, the PDE estimator  $\eta_{\mathcal{T}}(v, f)$  depends on  $f$  via a projection  $P_{\mathcal{T}}f$  upon a finite dimensional subspace of  $\mathbb{W}^*$ . Inspired by Lemma 4.5 (localization of  $H^{-1}$ -norm) we denote by  $\mathbb{W}_{\mathcal{T}}^*$  a suitable decomposition of  $\mathbb{W}^*$  subordinate to  $\mathcal{T}$  with norm  $\|f\|_{\mathbb{W}_{\mathcal{T}}^*}$ . In this part of the discussion, we prefer to make the dependence of  $\eta_{\mathcal{T}}$  upon  $P_{\mathcal{T}}f$  explicit to avoid confusion, so we will write  $\eta_{\mathcal{T}}(v, P_{\mathcal{T}}f)$  rather than  $\eta_{\mathcal{T}}(v, f)$  as usual.

Let us begin by stating two assumptions on the estimator (10.37) to be used in the sequel.

**Assumption 10.17 (Lipschitz continuity of estimator)** There exists a constant  $C_{\text{Lip}} > 0$  such that for any  $\mathcal{T} \in \mathbb{T}$ , any  $v, w \in \mathbb{V}_{\mathcal{T}}$  and any  $f, g \in \mathbb{W}^*$ , we have

$$|\eta_{\mathcal{T}}(v, P_{\mathcal{T}}f) - \eta_{\mathcal{T}}(w, P_{\mathcal{T}}g)| \leq C_{\text{Lip}} (\|v - w\|_{\mathbb{V}} + \|P_{\mathcal{T}}f - P_{\mathcal{T}}g\|_{\mathbb{W}_{\mathcal{T}}^*}).$$

**Assumption 10.18 (monotonicity of estimator)** If  $\mathcal{T} \in \mathbb{T}$  and  $\mathcal{T}_*$  is a refinement of  $\mathcal{T}$ , then the projection operator satisfies  $P_{\mathcal{T}_*}P_{\mathcal{T}} = P_{\mathcal{T}}$  and

$$\eta_{\mathcal{T}_*}(v, P_{\mathcal{T}}f, T) \leq \eta_{\mathcal{T}}(v, P_{\mathcal{T}}f, T) \quad \forall T \in \mathcal{T} \quad \text{and} \quad \forall v \in \mathbb{V}_{\mathcal{T}}, \forall f \in \mathbb{W}^*.$$

It is useful for the subsequent applications to have explicit criteria which guarantee the fulfilment of Assumption 10.3. This is the purpose of the following result.

**Proposition 10.19 (estimator reduction under Dörfler marking)** *Let the estimator  $\eta_{\mathcal{T}}(v, P_{\mathcal{T}}f)$  in (10.37) be used in GALERKIN. Let Assumptions 10.17 and 10.18 be valid. Let  $\mathcal{T}_*$  be a refinement of  $\mathcal{T}$ , with estimator  $\eta_{\mathcal{T}_*}(v, P_{\mathcal{T}_*}f)$ , obtained by bisecting the elements  $T \in \mathcal{M}$  marked in MARK, using a Dörfler condition on the estimator  $\eta_{\mathcal{T}}(u_{\mathcal{T}}, P_{\mathcal{T}}f)$  for the Galerkin solution  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}$ . Suppose that there exists  $\lambda \in (0, 1)$  such that*

$$\eta_{\mathcal{T}_*}(u_{\mathcal{T}}, P_{\mathcal{T}}f, T)^2 \leq \lambda \eta_{\mathcal{T}}(u_{\mathcal{T}}, P_{\mathcal{T}}f, T)^2 \quad \forall T \in \mathcal{M}. \quad (10.38)$$

Then, there exists  $0 < \rho < 1$  and  $C > 0$  such that for all  $v_{\mathcal{T}_*} \in \mathbb{V}_{\mathcal{T}_*}$

$$\eta_{\mathcal{T}_*}(v_{\mathcal{T}_*}, P_{\mathcal{T}_*}f)^2 \leq \rho \eta_{\mathcal{T}}(u_{\mathcal{T}}, P_{\mathcal{T}}f)^2 + C \left( \|v_{\mathcal{T}_*} - u_{\mathcal{T}}\|_{\mathbb{V}}^2 + \|P_{\mathcal{T}_*}f - P_{\mathcal{T}}f\|_{\mathbb{W}_{\mathcal{T}_*}^*}^2 \right).$$

*Proof.* By Assumption 10.17 applied to  $\mathcal{T}_*$ , we have

$$\eta_{\mathcal{T}_*}(v_{\mathcal{T}_*}, P_{\mathcal{T}_*}f) \leq \eta_{\mathcal{T}_*}(u_{\mathcal{T}}, P_{\mathcal{T}}f) + C_{\text{Lip}} \left( \|v_{\mathcal{T}_*} - u_{\mathcal{T}}\|_{\mathbb{V}} + \|P_{\mathcal{T}_*}f - P_{\mathcal{T}}f\|_{\mathbb{W}_{\mathcal{T}_*}^*} \right).$$

Using Assumption 10.18 while extending Proposition 4.56 to the current abstract setting, we have for any  $\delta > 0$

$$\begin{aligned} \eta_{\mathcal{T}_*}(v_{\mathcal{T}_*}, P_{\mathcal{T}_*}f)^2 &\leq (1 + \delta) (\eta_{\mathcal{T}}(u_{\mathcal{T}}, P_{\mathcal{T}}f)^2 - (1 - \lambda) \eta_{\mathcal{T}}(u_{\mathcal{T}}, P_{\mathcal{T}}f, \mathcal{M})^2) \\ &\quad + 2(1 + \delta^{-1}) C_{\text{Lip}}^2 \left( \|v_{\mathcal{T}_*} - u_{\mathcal{T}}\|_{\mathbb{V}}^2 + \|P_{\mathcal{T}_*}f - P_{\mathcal{T}}f\|_{\mathbb{W}_{\mathcal{T}_*}^*}^2 \right). \end{aligned}$$

We conclude using Dörfler condition  $\eta_{\mathcal{T}}(u_{\mathcal{T}}, \mathcal{M}) \geq \theta \eta_{\mathcal{T}}(u_{\mathcal{T}})$  and choosing  $\delta$  small enough.  $\square$

Before proceeding further, let us introduce the quantity

$$\text{osc}_{\mathcal{T}}(f) := \|f - P_{\mathcal{T}}f\|_{\mathbb{W}_{\mathcal{T}}^*},$$

which is a measure of the oscillation of the data  $f$ . If  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}$  is the solution of AFEM-SW, we let  $\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f)$  indicate the full estimator defined by

$$\mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f)^2 := \eta_{\mathcal{T}}(u_{\mathcal{T}}, P_{\mathcal{T}}f)^2 + \text{osc}_{\mathcal{T}}(f)^2. \quad (10.39)$$

We formulate the following assumption on the data oscillation.

**Assumption 10.20 (quasi-monotonicity of oscillation)** There exists a constant  $C_{\text{osc}} > 0$  such that for any  $\mathcal{T} \in \mathbb{T}$  and any admissible refinement  $\mathcal{T}_* \geq \mathcal{T}$ , we have

$$\text{osc}_{\mathcal{T}_*}(f) \leq C_{\text{osc}} \text{osc}_{\mathcal{T}}(f).$$

A consequence of this assumption is the bound

$$\|P_{\mathcal{T}_*}f - P_{\mathcal{T}}f\|_{\mathbb{W}_{\mathcal{T}_*}^*} \leq \|f - P_{\mathcal{T}_*}f\|_{\mathbb{W}_{\mathcal{T}_*}^*} + \|f - P_{\mathcal{T}}f\|_{\mathbb{W}_{\mathcal{T}_*}^*} \leq (1 + C_{\text{osc}}) \text{osc}_{\mathcal{T}}(f),$$

which, inserted into the reduction estimate of Proposition 10.19, gives the existence of  $0 < \rho_0 < 1$  and  $C_0 > 0$  independent of  $\mathcal{T}$  such that

$$\eta_{\mathcal{T}_*}(u_{\mathcal{T}_*}, P_{\mathcal{T}_*}f)^2 \leq \rho_0 \eta_{\mathcal{T}}(u_{\mathcal{T}}, P_{\mathcal{T}}f)^2 + C_0 (\|u_{\mathcal{T}_*} - u_{\mathcal{T}}\|_{\mathbb{V}}^2 + \text{osc}_{\mathcal{T}}(f)^2). \quad (10.40)$$

We aim at establishing a linear convergence result similar to Theorem 10.6 for the sequence  $\{u_j\}_{j=0}^{\infty}$  generated by AFEM-SW. To this end, we introduce as usual the short-hand notation  $e_j = \|u - u_j\|_{\mathbb{V}}$ ,  $E_{j+1} = \|u_{j+1} - u_j\|_{\mathbb{V}}$ ,  $\eta_j = \eta_{\mathcal{T}_j}(u_j, P_{\mathcal{T}_j}f)$ ,  $\text{osc}_j = \text{osc}_{\mathcal{T}_j}(f)$ ,  $\mathcal{E}_j = \mathcal{E}_{\mathcal{T}_j}(u_j, f)$ , and we also introduce the scaled estimator

$$\zeta_j^2 := \eta_j^2 + \gamma \text{osc}_j^2, \quad (10.41)$$

where the parameter  $\gamma > 0$  is to be found. Note that at this point we have three parameters,  $\omega \in (0, 1)$ ,  $\xi \in (0, 1)$ , and  $\gamma > 0$  to play with, and the idea is to find conditions on them such that an inequality similar to (10.9) in the proof of Theorem 10.6 holds true. The following result is an intermediate step.

**Lemma 10.21 (linear estimator reduction)** *Let Assumptions 10.3 (estimator reduction), 10.17 (Lipschitz continuity of estimator), 10.18 (monotonicity of estimator) and 10.20 (quasi-monotonicity of oscillation) be valid. There exists  $\omega_0 > 0$  such that, for any choice of parameters  $0 < \omega \leq \omega_0$  and  $0 < \xi \leq 1/\sqrt{2}$  in AFEM-SW, there exist constants  $0 < \rho < 1$ ,  $\Lambda > 0$ ,  $\gamma \geq 1$  for which*

$$\zeta_k^2 \leq \rho^{k-j} \zeta_j^2 + \Lambda \sum_{\ell=j+1}^k E_{\ell}^2, \quad k \geq j \geq 0. \quad (10.42)$$

*Proof.* We discuss separately the two alternatives in Algorithm 5.16 (AFEM-SW).

□ *Case  $\text{osc}_j \leq \omega \mathcal{E}_j$ .* We use (10.40) to get

$$\eta_{j+1}^2 \leq \rho_0 \eta_j^2 + C_0 E_{j+1}^2 + C_0 \text{osc}_j^2$$

and Assumption 10.20 to write

$$\text{osc}_{j+1}^2 \leq C_{\text{osc}}^2 \text{osc}_j^2.$$

From

$$\text{osc}_j^2 \leq \omega^2 \mathcal{E}_j^2 = \omega^2 (\eta_j^2 + \text{osc}_j^2) \leq \omega^2 (\eta_j^2 + \gamma \text{osc}_j^2) = \omega^2 \zeta_j^2$$



provided  $\gamma \geq 1$ , we deduce

$$\begin{aligned}
 \zeta_{j+1}^2 &= \eta_{j+1}^2 + \gamma \operatorname{osc}_{j+1}^2 \leq \rho_0 \eta_j^2 + C_0 E_{j+1}^2 + (C_0 + \gamma C_{\operatorname{osc}}^2) \operatorname{osc}_j^2 \\
 &\leq \rho_0 \eta_j^2 + C_0 E_{j+1}^2 + (C_0 + \gamma C_{\operatorname{osc}}^2) \omega^2 (\eta_j^2 + \gamma \operatorname{osc}_j^2) \\
 &= [\rho_0 + (C_0 + \gamma C_{\operatorname{osc}}^2) \omega^2] \eta_j^2 + [(C_0 + \gamma C_{\operatorname{osc}}^2) \omega^2] \gamma \operatorname{osc}_j^2 + C_0 E_{j+1}^2 \\
 &\leq [\rho_0 + (C_0 + \gamma C_{\operatorname{osc}}^2) \omega^2] \zeta_j^2 + C_0 E_{j+1}^2.
 \end{aligned} \tag{10.43}$$

Below we will impose

$$\rho_1 := \rho_0 + (C_0 + \gamma C_{\operatorname{osc}}^2) \omega^2 < 1, \tag{10.44}$$

which will yield the desired bound

$$\zeta_{j+1}^2 \leq \rho_1 \zeta_j^2 + C_0 E_{j+1}^2. \tag{10.45}$$

□ *Case  $\operatorname{osc}_j > \omega \mathcal{E}_j$ .* We use  $\mathcal{E}_j^2 = \eta_j^2 + \operatorname{osc}_j^2 > \eta_j^2$  to get

$$\eta_j^2 < \frac{1}{\omega^2} \operatorname{osc}_j^2.$$

Proceeding as in the proofs of Proposition 10.19 (now with  $\mathcal{M} = \emptyset$ ) and of (10.40), we obtain for any  $\delta > 0$

$$\eta_{j+1}^2 \leq (1 + \delta) \eta_j^2 + C_\delta (E_{j+1}^2 + \operatorname{osc}_j^2) \leq (1 - \delta) \eta_j^2 + C_\delta E_{j+1}^2 + \left( \frac{2\delta}{\omega^2} + C_\delta \right) \operatorname{osc}_j^2$$

with  $C_\delta = C_0^2(1 + \delta^{-1})$ . On the other hand, since  $\operatorname{osc}_{j+1}$  is computed after a call to DATA, it satisfies

$$\operatorname{osc}_{j+1}^2 \leq \xi^2 \sigma_j^2 = \xi^2 \omega^2 \mathcal{E}_j^2 < \xi^2 \operatorname{osc}_j^2 = \frac{1 + \xi^2}{2} \operatorname{osc}_j^2 - \frac{1 - \xi^2}{2} \operatorname{osc}_j^2.$$

Combining the two last equations, we obtain

$$\begin{aligned}
 \zeta_{j+1}^2 &= \eta_{j+1}^2 + \gamma \operatorname{osc}_{j+1}^2 \leq (1 - \delta) \eta_j^2 + \gamma \frac{1 + \xi^2}{2} \operatorname{osc}_j^2 + C_\delta E_{j+1}^2 \\
 &\quad + \left( \frac{2\delta}{\omega^2} + C_\delta - \gamma \frac{1 - \xi^2}{2} \right) \operatorname{osc}_j^2.
 \end{aligned}$$

Below we will enforce

$$\Gamma := \frac{2\delta}{\omega^2} + C_\delta - \gamma \frac{1 - \xi^2}{2} \leq 0, \tag{10.46}$$

which will guarantee

$$\zeta_{j+1}^2 \leq \rho_2 \zeta_j^2 + C_\delta E_{j+1}^2, \tag{10.47}$$

with  $\rho_2 := \max(1 - \delta, \frac{1 + \xi^2}{2}) < 1$ .

**[3] Choice of parameters:** Summarizing, in both cases **[1]** and **[2]** we have obtained

$$\zeta_{j+1}^2 \leq \rho \zeta_j^2 + \Lambda E_{j+1}^2 \quad (10.48)$$

with  $\rho := \max(\rho_1, \rho_2) < 1$  and  $\Lambda := \max(C_0, C_\delta)$ , which holds under the conditions (10.44) and (10.46). Iterating (10.48), we obtain the desired bound (10.42).

To fulfil (10.44), we write  $\omega^2$  in the form  $\omega^2 = \frac{\sigma_0}{\gamma}$ , which gives

$$\rho_1 = \rho_0 + (C_0 + \gamma C_{\text{osc}}^2) \frac{\sigma_0}{\gamma} = \rho_0 + \left( \frac{C_0}{\gamma} + C_{\text{osc}}^2 \right) \sigma_0 \leq \rho_0 + (C_0 + C_{\text{osc}}^2) \sigma_0$$

since  $\gamma \geq 1$ , and we pick a  $\sigma_0 > 0$  small enough to make  $\rho_0 + (C_0 + C_{\text{osc}}^2) \sigma_0 < 1$ .

To fulfil (10.46), we use  $\xi \leq 1/\sqrt{2}$  and again  $\omega^2 = \frac{\sigma_0}{\gamma}$  to write

$$\Gamma = C_\delta + \left( \frac{2\delta}{\sigma_0} - \frac{1 - \xi^2}{2} \right) \gamma \leq C_\delta + \left( \frac{2\delta}{\sigma_0} - \frac{1}{4} \right) \gamma.$$

Choosing  $\delta = \delta_0 = \sigma_0/16$ , yields

$$\Gamma \leq C_{\delta_0} - \frac{1}{8} \gamma \leq 0 \quad \text{provided} \quad \gamma \geq 8C_{\delta_0}.$$

In conclusion, setting  $\gamma_0 = \max(1, 8C_{\delta_0})$  and  $\omega_0 = \sqrt{\sigma_0/\gamma_0}$ , we fulfil both conditions (10.44) and (10.46) for any  $0 < \omega \leq \omega_0$ , by choosing the scaling parameter  $\gamma = \sigma_0/\omega^2 \geq \gamma_0$ . This completes the proof.  $\square$

Before establishing the linear convergence result for Algorithm 5.16 (one-step AFEM with switch), we need to extend Assumption 10.2 (equivalence of error and estimator) to the present situation, in which the estimator  $\eta_{\mathcal{T}}$  is replaced by the full estimator  $\mathcal{E}_{\mathcal{T}}$  defined in (10.39); see Theorem 4.45 (modified residual estimator).

**Assumption 10.22 (equivalence of error and full estimator)** There are constants  $C_U \geq C_L > 0$  such that

$$C_L \mathcal{E}_j \leq \|u - u_j\|_{\mathbb{V}} \leq C_U \mathcal{E}_j \quad j \geq 0, \quad (10.49)$$

where  $\mathcal{E}_j = \mathcal{E}_{\mathcal{T}_j}(u_{\mathcal{T}_j}, f)$ .

**Theorem 10.23 (linear convergence for AFEM-SW)** *Let Assumptions 10.22 (equivalence of error and full estimator), 10.3 (estimator reduction), 10.17 (Lipschitz continuity of estimator) and 10.20 (quasi-monotonicity of oscillation) be valid. There exists  $\omega_0 \in (0, 1]$  such that, for any choice of parameters  $0 < \omega < \omega_0$  and  $0 < \xi, \theta < 1$  in AFEM-SW, constants  $0 < \rho < 1$  and  $c > 0$  exist for which*

$$e_{j+1} \leq c \rho^i e_j \quad \forall i, j \in \mathbb{N}, \quad (10.50)$$

where  $e_j := \|u - u_j\|_{\mathbb{V}}$ .

*Proof.* By Assumption 10.22 and  $\gamma \geq 1$  in (10.41), we get the equivalence of

error and scaled estimator

$$\frac{C_L^2}{\gamma} \zeta_j^2 = \frac{C_L^2}{\gamma} (\eta_j^2 + \gamma \operatorname{osc}_j^2) \leq C_L^2 \mathcal{E}_j^2 \leq e_j^2 \leq C_U^2 \mathcal{E}_j^2 \leq C_U^2 (\eta_j^2 + \gamma \operatorname{osc}_j^2) = C_U^2 \zeta_j^2.$$

Invoking (10.42) yields

$$\begin{aligned} e_k^2 &\leq C_U^2 \zeta_k^2 \leq C_U^2 \rho^{k-j} \zeta_j^2 + C_U^2 \Lambda \sum_{\ell=j+1}^k E_\ell^2, \\ &\leq \gamma \frac{C_U^2}{C_L^2} \rho^{k-j} e_j^2 + C_U^2 \Lambda \sum_{\ell=j+1}^k E_\ell^2. \end{aligned}$$

This inequality is similar to the expression (10.8) obtained in Step  $\square$  in the proof of Theorem 10.6. Therefore, we can finally proceed as in that proof and obtain the desired result.  $\square$

### 10.3.3 Algorithm 5.28 (AFEM-TS)

As usual, DATA produces discrete data  $\widehat{\mathcal{D}}_k$  on a mesh  $\widehat{\mathcal{T}}_k \geq \mathcal{T}_k$ , whereas GALERKIN produces an approximation  $u_{k+1}$  on a mesh  $\mathcal{T}_{k+1} \geq \widehat{\mathcal{T}}_k$  to the exact solution  $\widehat{u}_k$  of the boundary-value problem of interest with data  $\widehat{\mathcal{D}}_k$ . Its kernel is given in Algorithm 5.4 (GALERKIN).

In order to proceed, we need some notation and some assumptions. Let us denote by  $D(\Omega)$  the space of admissible data  $\mathcal{D}$  for the boundary-value problem at hand; let  $\|\mathcal{D}\|_{D(\Omega)}$  be a (quasi-)norm on  $D(\Omega)$ . If  $\mathcal{D}$  collects all data of problem (10.36), we write  $\mathcal{B} = \mathcal{B}(\mathcal{D})$  and  $\mathcal{F} = \mathcal{F}(\mathcal{D}) = \langle f(\mathcal{D}), \cdot \rangle$  to highlight the dependence of the bilinear and linear forms on the chosen data; similarly, we write  $u = u(\mathcal{D})$  for the corresponding solution. A perturbation  $\widehat{\mathcal{D}}$  of  $\mathcal{D}$  generates perturbed bilinear and linear forms  $\widehat{\mathcal{B}} = \mathcal{B}(\widehat{\mathcal{D}})$  and  $\widehat{\mathcal{F}} = \mathcal{F}(\widehat{\mathcal{D}}) = \langle f(\widehat{\mathcal{D}}), \cdot \rangle$ , and a perturbation  $\widehat{u} = u(\widehat{\mathcal{D}})$  of  $u$ , which satisfies

$$\widehat{u} \in \mathbb{V} : \quad \widehat{\mathcal{B}}[\widehat{u}, w] = \widehat{\mathcal{F}}[w] \quad \text{for all } w \in \mathbb{W}. \quad (10.51)$$

We assume, as in Section 5.4.2, that a call  $[\widehat{\mathcal{T}}, \widehat{\mathcal{D}}] = \text{DATA}(\mathcal{T}, \mathcal{D}, \tau)$  generates an admissible refinement  $\widehat{\mathcal{T}}$  of  $\mathcal{T}$  and discrete data  $\widehat{\mathcal{D}}$  over  $\widehat{\mathcal{T}}$ , such that

$$\|\mathcal{D} - \widehat{\mathcal{D}}\|_{D(\Omega)} \leq C_{\text{data}} \tau, \quad (10.52)$$

where  $C_{\text{data}} > 0$  depends on data (see Section 7.2). Finally, we associate to any admissible refinement  $\mathcal{T}$  of  $\mathcal{T}_0$ , two finite dimensional spaces  $\mathbb{V}_{\mathcal{T}} \subset \mathbb{V}$  and  $\mathbb{W}_{\mathcal{T}} \subset \mathbb{W}$  of equal dimension, made of piecewise polynomial functions on  $\mathcal{T}$  (typically, this is accomplished by choosing a type of finite element compatible with the pair  $(\mathbb{V}, \mathbb{W})$  and adopting it in any  $\mathcal{T} \in \mathbb{T}$ ).

We are ready to state the assumptions which will rule our forthcoming analysis of AFEM-TS.

**Assumption 10.24 (perturbation estimate)** For any  $\tilde{\mathcal{T}} \in \mathbb{T}$  and  $\varepsilon \leq \varepsilon_0$ , let  $[\hat{\mathcal{T}}, \hat{\mathcal{D}}] = \text{DATA}(\tilde{\mathcal{T}}, \mathcal{D}, \varepsilon)$  and let  $\hat{u} = u(\hat{\mathcal{D}})$  be the solution of (10.51). There exists a constant  $C_D > 0$ , independent of  $\tilde{\mathcal{T}}$  and  $\varepsilon$ , such that

$$\|u - \hat{u}\|_{\mathbb{V}} \leq C_D \|\mathcal{D} - \hat{\mathcal{D}}\|_{D(\Omega)}. \quad (10.53)$$

Note that, concatenating this inequality with (10.52) for  $\tau = \varepsilon$ , we can quantify the effect of a call to DATA on the perturbation of the exact solution; we have indeed

$$\|u - \hat{u}\|_{\mathbb{V}} \leq C_D C_{\text{data}} \varepsilon. \quad (10.54)$$

**Assumption 10.25 (uniform continuity constant)** For any  $\tilde{\mathcal{T}} \in \mathbb{T}$  and  $\varepsilon \leq \varepsilon_0$ , let  $[\hat{\mathcal{T}}, \hat{\mathcal{D}}] = \text{DATA}(\tilde{\mathcal{T}}, \mathcal{D}, \varepsilon)$  and let  $\hat{\mathcal{B}} = \mathcal{B}(\hat{\mathcal{D}})$  be the associated bilinear form. There exists a constant  $C_B \geq \|\mathcal{B}\|$ , independent of  $\tilde{\mathcal{T}}$  and  $\varepsilon$ , such that

$$\|\hat{\mathcal{B}}\| \leq C_B. \quad (10.55)$$

**Assumption 10.26 (uniform inf-sup constant)** For any  $\tilde{\mathcal{T}} \in \mathbb{T}$  and  $\varepsilon \leq \varepsilon_0$ , let  $[\hat{\mathcal{T}}, \hat{\mathcal{D}}] = \text{DATA}(\tilde{\mathcal{T}}, \mathcal{D}, \varepsilon)$ , let  $\hat{\mathcal{B}} = \mathcal{B}(\hat{\mathcal{D}})$  be the associated bilinear form, let  $\mathcal{T}$  be either  $\hat{\mathcal{T}}$  or an admissible refinement of  $\hat{\mathcal{T}}$ , and finally let  $\mathbb{V}_{\mathcal{T}} \subset \mathbb{V}$ ,  $\mathbb{W}_{\mathcal{T}} \subset \mathbb{W}$  be the subspaces built on  $\mathcal{T}$  as above. There exists a constant  $0 < \bar{\beta} \leq \beta$ , independent of  $\tilde{\mathcal{T}}$ ,  $\varepsilon$ , and  $\mathcal{T}$ , such that

$$\inf_{w \in \mathbb{W}_{\mathcal{T}}} \sup_{v \in \mathbb{V}_{\mathcal{T}}} \frac{\hat{\mathcal{B}}[v, w]}{\|v\|_{\mathbb{V}} \|w\|_{\mathbb{W}}} \geq \bar{\beta}. \quad (10.56)$$

The last assumption guarantees the well-posedness of all the discrete variational problems

$$u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}} : \quad \hat{\mathcal{B}}[u_{\mathcal{T}}, w_{\mathcal{T}}] = \hat{\mathcal{F}}[w_{\mathcal{T}}] \quad \forall w_{\mathcal{T}} \in \mathbb{W}_{\mathcal{T}}, \quad (10.57)$$

associated with the successive refinements of the initial mesh  $\mathcal{T}_0$  performed by AFEM-TS.

We want to prove, as in the coercive case (see Proposition 5.27), that the number of iterations performed in any call to GALERKIN inside AFEM-TS (which is finite by Proposition 10.16) is indeed uniformly bounded.

**Proposition 10.27 (computational cost of GALERKIN)** *Let Assumptions 10.2, 10.3, 10.24, 10.25, and 10.26 be valid. For any  $k \in \mathbb{N}$ , the number of subiterations  $J_k$  inside a call to GALERKIN at iteration  $k$  of AFEM-TS is bounded by a constant  $J$  independent of  $k$ .*

*Proof.* Denote by  $\mathcal{T}_{k,j}$  the successive refinements of  $\hat{\mathcal{T}}_k$  defined in GALERKIN at iteration  $k$ , and let  $u_{k,j} \in \mathbb{V}_{k,j} = \mathbb{V}_{\mathcal{T}_{k,j}}$  be the corresponding Galerkin solutions, which are approximations of the solution  $\hat{u}_k \in \mathbb{V}$  of the perturbed problem (10.51) with forms  $\hat{\mathcal{B}} = \hat{\mathcal{B}}_k = \mathcal{B}(\hat{\mathcal{D}}_k)$  and  $\hat{f} = \hat{f}_k = f(\hat{\mathcal{D}}_k)$ . Note as well that we use a posteriori estimators  $\eta_{k,j} = \eta_{k,j}(v)$  defined on  $\mathbb{V}_{k,j}$ , which depend on  $\hat{\mathcal{D}}_k$  via the coefficients of the equation. However, in reference to Assumptions 10.2 and 10.3,

we always suppose the constants in the bounds (10.5) and (10.6) to be independent of  $k$  and  $j$ .

Let us pick  $j := J_k - 1$ . By definition of stopping criterion in GALERKIN, and by (10.5) and (10.7), we get

$$\varepsilon_k < \eta_{k,j}(u_{k,j}) \leq \frac{1}{C_L} \|\widehat{u}_k - u_{k,j}\|_{\mathbb{V}} \leq \frac{c}{C_L} \rho^j \|\widehat{u}_k - u_{k,0}\|_{\mathbb{V}}. \quad (10.58)$$

The norm on the right-hand side can be bounded via Corollary 3.3 (quasi-monotonicity), applied to  $\mathcal{B} := \widehat{\mathcal{B}}_k$ ,  $u := \widehat{u}_k \in \mathbb{V}$ ,  $u_N := u_{k,0} \in \mathbb{V}_N := \mathbb{V}_{\widehat{\mathcal{T}}_k}$ , and  $v := u_k \in \mathbb{V}_M := \mathbb{V}_{\mathcal{T}_k} \subseteq \mathbb{V}_{\widehat{\mathcal{T}}_k}$  (the output of GALERKIN at iteration  $k-1$ ). Using Assumptions 10.25 (uniform continuity constant) and 10.26 (uniform inf-sup constant), we thus have

$$\|\widehat{u}_k - u_{k,0}\|_{\mathbb{V}} \leq \lambda \|\widehat{u}_k - u_k\|_{\mathbb{V}}, \quad (10.59)$$

with  $\lambda = \frac{C_B}{\beta}$ .

At last, we use again the triangle inequality to get

$$\begin{aligned} \|\widehat{u}_k - u_k\|_{\mathbb{V}} &\leq \|\widehat{u}_k - \widehat{u}_{k-1}\|_{\mathbb{V}} + \|\widehat{u}_{k-1} - u_k\|_{\mathbb{V}} \\ &\leq \|u - \widehat{u}_k\|_{\mathbb{V}} + \|u - \widehat{u}_{k-1}\|_{\mathbb{V}} + \|\widehat{u}_{k-1} - u_k\|_{\mathbb{V}}; \end{aligned}$$

then, Assumption 10.24 (perturbation estimate) yields  $\|u - \widehat{u}_k\|_{\mathbb{V}} \leq C_D C_{\text{data}} \omega \varepsilon_k$  and  $\|u - \widehat{u}_{k-1}\|_{\mathbb{V}} \leq C_D C_{\text{data}} \omega \varepsilon_{k-1}$ , whereas the termination test for GALERKIN at iteration  $k-1$  yields  $\|\widehat{u}_{k-1} - u_k\|_{\mathbb{V}} \leq C_U \varepsilon_{k-1}$ . Hence, recalling  $\varepsilon_{k-1} = 2\varepsilon_k$  and  $\omega \leq 1$ , we get

$$\|\widehat{u}_k - u_k\|_{\mathbb{V}} \leq \sigma \varepsilon_k \quad (10.60)$$

with  $\sigma = 3C_D C_{\text{data}} + 2C_U$ . Finally, concatenating (10.58), (10.59) and (10.60), we obtain

$$\rho^j \geq \frac{C_L}{c\lambda\sigma},$$

which implies  $J_k \leq 1 + (\log \frac{C_L}{c\lambda\sigma})(\log \rho)^{-1} =: J$ .  $\square$

The remaining of this section is devoted to investigate the rate-optimality of AFEMs for inf-sup stable problems. Precisely, we aim at establishing the analogue of bound (6.1) for such problems, i.e.,

$$\|u - u_{\mathcal{T}}\|_{\mathbb{V}} \leq C(u, \mathcal{D}) (\#\mathcal{T})^{-s}. \quad (10.61)$$

To this end, we have to introduce approximation classes for the solution and the data, and to study the quasi-optimality properties of mesh refinement and GALERKIN.

#### 10.3.4 Nonlinear approximation classes

The definition of the approximation class  $\mathbb{A}_s = \mathbb{A}_s(\mathbb{V}; \mathcal{T}_0)$  for functions in  $\mathbb{V}$  is identical to that given in Section 6.1.1 for functions in  $H_0^1(\Omega)$  (see Definition 6.1), provided the norm  $|v|_{H_0^1(\Omega)}$  is replaced by the norm  $\|v\|_{\mathbb{V}}$  at all occurrences.

In the rest of the section we will make the following regularity assumption.

**Assumption 10.28 (approximability of  $u$ )** The exact solution  $u \in \mathbb{V}$  of problem (10.36) belongs to the approximation class  $\mathbb{A}_s(\mathbb{V}; \mathcal{T}_0)$  for some  $s = s_u \in (0, \frac{n}{d}]$ .

The approximation classes of a data  $\mathcal{D} \in D(\Omega)$  are defined via discrete approximations  $\mathcal{D}_{\mathcal{T}} \in \mathbb{D}_{\mathcal{T}}$  subordinate to a partition  $\mathcal{T} \in \mathbb{T}$ , which produce the oscillation

$$\text{osc}_{\mathcal{T}}(\mathcal{D}) = \inf_{\mathcal{D}_{\mathcal{T}} \in \mathbb{D}_{\mathcal{T}}} \|\mathcal{D} - \mathcal{D}_{\mathcal{T}}\|_{D(\Omega)}.$$

**Definition 10.29 (approximation classes of  $\mathcal{D}$ )** Let  $\mathbb{D}_s := \mathbb{D}_s(D(\Omega); \mathcal{T}_0)$  be the set of data  $\mathcal{D} \in D(\Omega)$  satisfying

$$|\mathcal{D}|_{\mathbb{D}_s} := \sup_{N \geq \#\mathcal{T}_0} \left( N^s \inf_{\mathcal{T} \in \mathbb{T}_N} \text{osc}_{\mathcal{T}}(\mathcal{D}) \right) < \infty \Rightarrow \inf_{\mathcal{T} \in \mathbb{T}_N} \text{osc}_{\mathcal{T}}(\mathcal{D}) \leq |\mathcal{D}|_{\mathbb{D}_s} N^{-s}. \quad (10.62)$$

The following assumptions on the data of our boundary-value problem will be valid in the rest of the section.

**Assumption 10.30 (approximability of  $\mathcal{D}$ )** The data  $\mathcal{D} \in D(\Omega)$  of problem (10.36) belongs to the approximation class  $\mathbb{D}_s(D(\Omega); \mathcal{T}_0)$  for some  $s = s_{\mathcal{D}} \in (0, \frac{n}{d}]$ .

**Assumption 10.31 (quasi-optimality of DATA)** A call  $[\widehat{\mathcal{T}}, \widehat{\mathcal{D}}] = \text{DATA}(\mathcal{T}, \mathcal{D}, \varepsilon)$  marks a set of elements  $\mathcal{M}_{\mathcal{D}}$  whose cardinality  $N(\mathcal{D}) = \#\mathcal{M}_{\mathcal{D}}$  obeys

$$N(\mathcal{D}) \lesssim |\mathcal{D}|_{\mathbb{D}_s}^{\frac{1}{s_{\mathcal{D}}}} \varepsilon^{-\frac{1}{s_{\mathcal{D}}}}. \quad (10.63)$$

The concept of  $\varepsilon$ -approximation of order  $s$  of  $u \in \mathbb{A}_s(\mathbb{V}; \mathcal{T}_0)$  is identical to the one given in Definition 6.12, and so is the proof of the following result.

**Lemma 10.32 ( $\varepsilon$ -approximation of  $u$  of order  $s$ )** Let  $u \in \mathbb{A}_s(\mathbb{V}; \mathcal{T}_0)$  and  $v \in \mathbb{V}$  satisfy  $\|u - v\|_{\mathbb{V}} \leq \varepsilon$  for some  $0 < \varepsilon \leq \varepsilon_0$ . Then  $v$  is a  $2\varepsilon$ -approximation of order  $s$  to  $u$ .

### 10.3.5 Rate-optimality of GALERKIN

To estimate the growth of the cardinality of the meshes produced inside a call to GALERKIN, which always deals with discrete data, and to relate it to the approximation class of the exact solution  $u$ , we need an additional assumption of the estimators  $\eta_{\mathcal{T}}$ . In the sequel, for any subset  $\mathcal{R} \subset \mathcal{T}$ , we define  $\eta_{\mathcal{T}}(v, \mathcal{S})$  by

$$\eta_{\mathcal{T}}(v, \mathcal{S})^2 = \sum_{T \in \mathcal{S}} \eta_{\mathcal{T}}(v, T)^2.$$

**Assumption 10.33 (discrete reliability of the estimator)** There exists a constant  $c_2 > 0$  such that for any  $\mathcal{T} \in \mathbb{T}$  and any refinement  $\mathcal{T}_* \geq \mathcal{T}$ , if  $\mathcal{R} = \mathcal{R}_{\mathcal{T} \rightarrow \mathcal{T}_*} = \mathcal{T} \setminus \mathcal{T}_*$  is the set of refined elements of  $\mathcal{T}$ , it holds

$$\|u_{\mathcal{T}_*} - u_{\mathcal{T}}\|_{\mathbb{V}} \leq c_2 \eta_{\mathcal{T}}(u_{\mathcal{T}}, \mathcal{R}),$$

where  $u_{\mathcal{T}}$  and  $u_{\mathcal{T}_*}$  (resp.) are the Galerkin solutions in  $\mathbb{V}_{\mathcal{T}}$  and  $\mathbb{V}_{\mathcal{T}_*}$  (resp.).

We recall that the module MARK in GALERKIN implements Dörfler's strategy, i.e., for a fixed  $\theta \in (0, 1]$ , it identifies a subset  $\mathcal{M} \subseteq \mathcal{T}$  of elements undergoing bisection by the condition

$$\eta_{\mathcal{T}}(u_{\mathcal{T}}, \mathcal{M}) \geq \theta \eta_{\mathcal{T}}(u_{\mathcal{T}}). \quad (10.64)$$

The following property is the analogue of the one stated in Lemma 6.16 for coercive problems. Since the proof is similar, we omit it.

**Lemma 10.34 (Dörfler marking)** *Let Assumptions 10.17 and 10.33 be valid. For all  $0 < \mu < \frac{1}{2}$  there exists  $0 < \theta_0 < 1$  such that, if  $\mathcal{T} \in \mathbb{T}$  and  $\mathcal{T}_*$  is a refinement of  $\mathcal{T}$  with refined set  $\mathcal{R} = \mathcal{T} \setminus \mathcal{T}_*$ , and if the Galerkin solutions  $u_{\mathcal{T}} \in \mathbb{V}_{\mathcal{T}}$  and  $u_{\mathcal{T}_*} \in \mathbb{V}_{\mathcal{T}_*}$  satisfy*

$$\eta_{\mathcal{T}_*}(u_{\mathcal{T}_*}) \leq \mu \eta_{\mathcal{T}}(u_{\mathcal{T}}),$$

then

$$\theta_0 \eta_{\mathcal{T}}(u_{\mathcal{T}}) \leq \eta_{\mathcal{T}}(u_{\mathcal{T}}, \mathcal{R}).$$

We are ready to investigate the rate-optimality of the  $k$ -th call to GALERKIN in the two-step AFEM (see Definition 5.1). We denote by  $\mathcal{M}_{k,j} \subseteq \mathcal{T}_{k,j}$  the marked set at the  $j$ -th iteration inside GALERKIN (hereafter, we refer to the notation in the proof of Proposition 10.27). To achieve quasi-optimality, the following assumption is fundamental.

**Assumption 10.35 (minimality of marked sets)** The module MARK selects a set  $\mathcal{M}_{k,j}$  with *minimal* cardinality among those satisfying Dörfler's condition

$$\eta_{k,j}(u_{k,j}, \mathcal{M}) \geq \theta \eta_{k,j}(u_{k,j}) \quad \forall k, j.$$

**Proposition 10.36 (cardinality of marked sets)** *Let Assumptions 10.2, 10.24, 10.25, 10.26, 10.28, 10.17, 10.33 and 10.35 hold true. There exists a constant  $C_0 > 0$  independent of  $k$  and  $j$  such that the cardinality  $N_{k,j}(u)$  of  $\mathcal{M}_{k,j}$  satisfies*

$$N_{k,j}(u) \leq C_0 |u|_{\mathbb{A}_s}^{1/s} \varepsilon_k^{-1/s} \quad (10.65)$$

and

$$N_{k,j}(u) \leq C_0 |u|_{\mathbb{A}_s}^{1/s} \|u - u_{k,j}\|_{\mathbb{V}}^{-1/s}. \quad (10.66)$$

*Proof.* The proof can be easily obtained by slightly adapting to the current abstract setting the proof of Corollary 6.22, keeping also into account Proposition 6.18.  $\square$

Let  $\mathcal{M}_k$  denote the set of marked elements in GALERKIN at iteration  $k$  of AFEM. Since the cardinality  $N_k(u) = \#\mathcal{M}_k$  of  $\mathcal{M}_k$  satisfies  $N_k(u) = \sum_{j=0}^{J_k-1} N_{k,j}(u)$ , we can estimate its cardinality by combining Propositions 10.27 and 10.36.

**Corollary 10.37 (rate-optimality of GALERKIN)** *Under the assumptions of Propositions 10.27 and 10.36, the total number of marked elements  $\mathcal{M}_k$  in GALERKIN at iteration  $k$  of AFEM satisfies*

$$N_k(u) \leq J C_0 |u|_{\mathbb{A}_s}^{1/s} \varepsilon_k^{-1/s}.$$

### 10.3.6 Rate-optimality of AFEM-TS

At last, we focus on the two-step AFEM in Definition 5.1 (AFEM-TS), and prove its rate-optimality, in relation to the nonlinear approximation classes of the solution  $u$  and the problem data  $\mathcal{D}$ .

**Theorem 10.38 (rate-optimality of AFEM-TS)** *Under the same assumptions of Proposition 10.36, plus Assumptions 10.30 and 10.31, there exists a constant  $C_*$  independent of  $u$  and  $\mathcal{D}$  such that the sequence  $(\mathcal{T}_k, \mathbb{V}_{\mathcal{T}_k}, u_{\mathcal{T}_k})$ ,  $k \geq 0$ , produced by AFEM-TS satisfies*

$$\|u - u_{\mathcal{T}_k}\|_{\mathbb{V}} \leq C_* \left( |u|_{\mathbb{A}_{s_u}}^{1/s_u} + |\mathcal{D}|_{\mathbb{D}_{s_{\mathcal{D}}}}^{1/s_{\mathcal{D}}} \right)^s (\#\mathcal{T}_k)^{-s},$$

with  $0 < s = \min(s_u, s_{\mathcal{D}}) \leq \frac{n}{d}$ .

*Proof.* Let us denote by  $\mathcal{M}_\ell^u$  ( $\mathcal{M}_\ell^{\mathcal{D}}$ , resp.) the set of elements marked by GALERKIN (by DATA, resp.) at iteration  $\ell$  of AFEM. By Corollary 10.37 and Assumption 10.31, there exist constants  $D_1, D_2$  independent of  $u$ ,  $\mathcal{D}$  and  $k$  such that

$$N_\ell(u) \leq D_1 |u|_{\mathbb{A}_{s_u}}^{1/s_u} \varepsilon_\ell^{-1/s_u}, \quad N_\ell(\mathcal{D}) \leq D_2 |\mathcal{D}|_{\mathbb{D}_{s_{\mathcal{D}}}}^{1/s_{\mathcal{D}}} \varepsilon_\ell^{-1/s_{\mathcal{D}}}.$$

Then, we conclude as in the proof of Theorem 6.24.  $\square$

## 10.4 The Stokes problem

Here we consider the Stokes problem

$$\begin{aligned} -\Delta \mathbf{u} + \nabla p &= \mathbf{f} && \text{in } \Omega, \\ \operatorname{div} \mathbf{u} &= 0 && \text{in } \Omega, \\ \mathbf{u} &= 0 && \text{on } \partial\Omega, \end{aligned} \tag{10.67}$$

already introduced in Section 2.3. Assuming  $\mathbf{f} \in H^{-1}(\Omega; \mathbb{R}^d) = \mathbb{V}^*$ , its weak formulation is given in (2.15) or, equivalently, in (2.16), where the bilinear form  $\mathcal{B}$  is continuous and inf-sup stable, as a consequence of Brezzi's Theorem 2.11 (Brezzi); see Section 2.4.

A Galerkin discretization of this problem, based on finite dimensional subspaces  $\mathbb{V}_{\mathcal{T}} \subset \mathbb{V} = H_0^1(\Omega; \mathbb{R}^d)$  and  $\mathbb{Q}_{\mathcal{T}} \subset \mathbb{Q} = L_0^2(\Omega)$ , reads as follows: find  $(\mathbf{u}_{\mathcal{T}}, p_{\mathcal{T}}) \in \mathbb{V}_{\mathcal{T}} \times \mathbb{Q}_{\mathcal{T}}$  such that

$$\begin{aligned} a[\mathbf{u}_{\mathcal{T}}, \mathbf{v}] + b[p_{\mathcal{T}}, \mathbf{v}] &= \langle \mathbf{f}, \mathbf{v} \rangle && \forall \mathbf{v} \in \mathbb{V}_{\mathcal{T}}, \\ b[q, \mathbf{u}_{\mathcal{T}}] &= 0 && \forall q \in \mathbb{Q}_{\mathcal{T}}, \end{aligned} \tag{10.68}$$

or equivalently,

$$(\mathbf{u}_{\mathcal{T}}, p_{\mathcal{T}}) \in \mathbb{V}_{\mathcal{T}} \times \mathbb{Q}_{\mathcal{T}} : \quad \mathcal{B}[(\mathbf{u}_{\mathcal{T}}, p_{\mathcal{T}}), (\mathbf{v}, q)] = \langle \mathbf{f}, \mathbf{v} \rangle \quad \forall (\mathbf{v}, q) \in \mathbb{V}_{\mathcal{T}} \times \mathbb{Q}_{\mathcal{T}}.$$

We assume that the pair  $(\mathbb{V}_{\mathcal{T}}, \mathbb{Q}_{\mathcal{T}})$  is *uniformly inf-sup stable* for the form  $b$ , i.e.,



there exists a constant  $\beta > 0$ , independent of the refinement  $\mathcal{T}$ , such that

$$\inf_{q \in \mathbb{Q}_{\mathcal{T}}} \sup_{\mathbf{v} \in \mathbb{V}_{\mathcal{T}}} \frac{b[q, \mathbf{v}]}{\|\mathbf{v}\|_{\mathbb{V}} \|q\|_{\mathbb{W}}} \geq \beta. \quad (10.69)$$

This condition is equivalent to the uniform inf-sup stability of the bilinear form  $\mathcal{B}$  on the product space  $\mathbb{X}_{\mathcal{T}} := \mathbb{V}_{\mathcal{T}} \times \mathbb{Q}_{\mathcal{T}}$ . Then, applying a discrete form of Brezzi's Theorem, we obtain the existence and uniqueness of the solution of (10.68), which satisfies the stability bound

$$\|\mathbf{u}_{\mathcal{T}}\|_{\mathbb{V}} + \|p_{\mathcal{T}}\|_{\mathbb{Q}} \leq C \|\mathbf{f}\|_{\mathbb{V}^*}, \quad (10.70)$$

where  $C$  only depends on the continuity constant  $\|a\|$  and the coercivity constant  $\alpha$  of the form  $a$ , and the inf-sup constant  $\beta$ . Furthermore, we have the quasi-best approximation bounds ([Boffi *et al.* 2013, Proposition 8.2.1])

$$\|\mathbf{u} - \mathbf{u}_{\mathcal{T}}\|_{\mathbb{V}} \leq C_{11} \min_{\mathbf{v} \in \mathbb{V}_{\mathcal{T}}} \|\mathbf{u} - \mathbf{v}\|_{\mathbb{V}} + C_{12} \min_{q \in \mathbb{Q}_{\mathcal{T}}} \|p - q\|_{\mathbb{Q}} \quad (10.71)$$

$$\|p - p_{\mathcal{T}}\|_{\mathbb{V}} \leq C_{21} \min_{\mathbf{v} \in \mathbb{V}_{\mathcal{T}}} \|\mathbf{u} - \mathbf{v}\|_{\mathbb{V}} + C_{22} \min_{q \in \mathbb{Q}_{\mathcal{T}}} \|p - q\|_{\mathbb{Q}} \quad (10.72)$$

where the constants  $C_{ij}$ ,  $1 \leq i, j \leq 2$ , only depend on the quantities  $\|a\|$ ,  $\|b\|$ ,  $\alpha$  and  $\beta$ .

There are many families of finite element spaces that are uniformly inf-sup stable for the Stokes problem; see [Boffi *et al.* 2013, Chapter 8]. Among them, we consider here the Taylor-Hood element [Taylor and Hood 1973] and its higher-order versions. They all use continuous discrete pressures, so they fit in the general form

$$\begin{aligned} \mathbb{V}_{\mathcal{T}} &= \{\mathbf{v} \in H_0^1(\Omega; \mathbb{R}^d) : \mathbf{v}|_T \in \mathbf{V}_T, T \in \mathcal{T}\} \\ \mathbb{Q}_{\mathcal{T}} &= \{q \in L_0^2(\Omega) \cap C^0(\bar{\Omega}) : q|_T \in Q_T, T \in \mathcal{T}\}, \end{aligned}$$

where  $\mathbf{V}_T$  and  $Q_T$  are spaces of polynomials on the element  $T$ . Considering simplicial elements, we have for  $n \geq 2$

$$\mathbf{V}_T = (\mathbb{P}_n(T))^d, \quad Q_T = \mathbb{P}_{n-1}(T). \quad (10.73)$$

The convergence and optimality of an adaptive algorithm for the Stokes problem based on Taylor-Hood elements was first established by Feischl [Feischl 2019] (see also [Feischl 2022, Section 6]). We aim at deriving a similar result using the abstract framework presented in this section.

We start by developing the a posteriori error analysis, and for this we introduce the weak residual

$$\langle \mathcal{R}_{\mathcal{T}}, (\mathbf{v}, q) \rangle := \langle \mathbf{f}, \mathbf{v} \rangle - \mathcal{B}[(\mathbf{u}_{\mathcal{T}}, p_{\mathcal{T}}), (\mathbf{v}, q)] \quad \forall (\mathbf{v}, q) \in \mathbb{V} \times \mathbb{Q},$$

which we represent as  $\mathcal{R}_{\mathcal{T}} = (\mathcal{R}_{\mathcal{T}}^m, \mathcal{R}_{\mathcal{T}}^c)$  according to the momentum and continuity equations; note that  $\mathcal{R}_{\mathcal{T}}^c = \operatorname{div} \mathbf{u}_{\mathcal{T}}$ . The continuity and inf-sup stability properties of the exact Stokes form  $\mathcal{B}$  yield the equivalence

$$\|\mathbf{u} - \mathbf{u}_{\mathcal{T}}\|_{\mathbb{V}} + \|p - p_{\mathcal{T}}\|_{\mathbb{Q}} \approx \|\mathcal{R}_{\mathcal{T}}\|_{\mathbb{V}^* \times \mathbb{Q}^*} \approx \|\mathcal{R}_{\mathcal{T}}^m\|_{\mathbb{V}^*} + \|\operatorname{div} \mathbf{u}_{\mathcal{T}}\|_{L^2(\Omega)}. \quad (10.74)$$

We now apply Corollary 4.6 (star localization of residual norm) to each component of the momentum residual  $\mathcal{R}_{\mathcal{T}}^m$  to get

$$\|\mathcal{R}_{\mathcal{T}}^m\|_{\mathbb{V}^*}^2 \approx \sum_{z \in \mathcal{V}} \|\mathcal{R}_{\mathcal{T}}^m\|_{(H^{-1}(\omega_z))^d}^2,$$

whereas Lemma 4.35 (splitting of local residual norm) yields the equivalence

$$\|\mathcal{R}_{\mathcal{T}}^m\|_{(H^{-1}(\omega_z))^d}^2 \approx \|P_{\mathcal{T}}\mathbf{f} + \Delta \mathbf{u}_{\mathcal{T}} - \nabla p_{\mathcal{T}}\|_{(H^{-1}(\omega_z))^d}^2 + \|\mathbf{f} - P_{\mathcal{T}}\mathbf{f}\|_{(H^{-1}(\omega_z))^d}^2.$$

In view of mesh refinement, we recall Lemma 4.8 (localization re-indexing) and we express the error indicator in terms of elements  $T \in \mathcal{T}$  rather than stars  $\omega_z$ , in analogy with the scalar case (4.52). To this end, define

$$\begin{aligned} \eta_{\mathcal{T}}^m(\mathbf{u}_{\mathcal{T}}, p_{\mathcal{T}}, T)^2 &:= h_T^2 \|P_T \mathbf{f} + \Delta \mathbf{u}_T - \nabla p_T\|_{(L^2(T))^d}^2 \\ &\quad + h_T \sum_{F \subset \partial T \setminus \partial \Omega} \|[(\nabla \mathbf{u}_T) \mathbf{n}_F] - P_F \mathbf{f}\|_{(L^2(F))^d}^2, \quad (10.75) \\ \text{osc}_{\mathcal{T}}(\mathbf{f}, T)_{-1}^2 &:= \|\mathbf{f} - P_T \mathbf{f}\|_{(H^{-1}(\omega_T))^d}^2. \end{aligned}$$

Note that the jump term  $[(\nabla \mathbf{u}_T) \mathbf{n}_F]$  does not contain the pressure contribution, since discrete pressures in  $\mathbb{Q}_{\mathcal{T}}$  are globally continuous. We thus have

$$\|\mathcal{R}_{\mathcal{T}}^m\|_{\mathbb{V}^*}^2 \approx \sum_{T \in \mathcal{T}} \left( \eta_{\mathcal{T}}^m(\mathbf{u}_{\mathcal{T}}, p_{\mathcal{T}}, T)^2 + \text{osc}_{\mathcal{T}}(\mathbf{f}, T)_{-1}^2 \right). \quad (10.76)$$

Recalling (10.74), the full local PDE residual indicator could be defined as

$$\eta_{\mathcal{T}}(\mathbf{u}_{\mathcal{T}}, p_{\mathcal{T}}, T)^2 := \eta_{\mathcal{T}}^m(\mathbf{u}_{\mathcal{T}}, p_{\mathcal{T}}, T)^2 + \|\text{div } \mathbf{u}_{\mathcal{T}}\|_{L^2(T)}^2.$$

However, such a quantity is not guaranteed to strictly reduce under mesh refinement, due to the presence of the divergence term, which is not scaled by a positive power of the meshsize. The following result provides an equivalent expression of the last term, which does reduce. We recall the definition (9.1) of jumps across faces.

**Lemma 10.39 (norm equivalence for divergence)** *It holds*

$$\|\text{div } \mathbf{u}_{\mathcal{T}}\|_{L^2(\Omega)}^2 \approx \sum_{T \in \mathcal{T}} \sum_{F \subset \partial T \setminus \partial \Omega} h_F \|[(\text{div } \mathbf{u}_{\mathcal{T}})]\|_{L^2(F)}^2.$$

*Proof.* The result follows from applying to  $\varphi = \text{div } \mathbf{u}_{\mathcal{T}}$  the equivalence

$$\|\varphi - \Pi_{\mathcal{T}} \varphi\|_{L^2(\Omega)}^2 \approx \sum_{T \in \mathcal{T}} \sum_{F \subset \partial T \setminus \partial \Omega} h_F \|[(\varphi)]\|_{L^2(F)}^2 \quad \forall \varphi \in \mathbb{S}_{\mathcal{T}}^{n-1, -1}$$

(where  $\Pi_{\mathcal{T}}$  is the  $L^2$ -orthogonal projection upon  $\mathbb{S}_{\mathcal{T}}^{n-1, 0}$ ), after observing that  $\Pi_{\mathcal{T}} \varphi = 0$  since  $\mathbf{u}_{\mathcal{T}}$  is discretely divergence-free, i.e., it satisfies the second set of equations in (10.68). To prove the equivalence for arbitrary  $\varphi \in \mathbb{S}_{\mathcal{T}}^{n-1, -1}$ , we use the quasi-interpolation operator  $\mathcal{I}_{\mathcal{T}}^{\text{dG}}$  introduced in Section 9.1.2, which leaves

$\mathbb{S}_{\mathcal{T}}^{n-1,0}$  invariant. Then, it is easily seen that

$$\|\varphi - \Pi_{\mathcal{T}}\varphi\|_{L^2(\Omega)}^2 \approx \|\varphi - \mathcal{I}_{\mathcal{T}}^{\text{dG}}\varphi\|_{L^2(\Omega)}^2,$$

so, it is enough to prove the equivalence with  $\Pi_{\mathcal{T}}$  replaced by  $\mathcal{I}_{\mathcal{T}}^{\text{dG}}$ . But this calculation can be done on patches  $\omega_T$  since  $\mathcal{I}_{\mathcal{T}}^{\text{dG}}$  is quasi-local:

$$\|\varphi - \mathcal{I}_{\mathcal{T}}^{\text{dG}}\varphi\|_{L^2(T)}^2 \lesssim \sum_{F \subset \omega_T} h_F \|\llbracket \varphi \rrbracket\|_{L^2(F)}^2 \lesssim \|\varphi - \mathcal{I}_{\mathcal{T}}^{\text{dG}}\varphi\|_{L^2(\omega_T)}^2 \quad \forall \varphi \in \mathbb{S}_{\mathcal{T}}^{n-1,-1}.$$

The first inequality follows from (9.10) (see also [Bänsch, Morin and Nochetto 2002, Proposition 5.4]). The second inequality results from the fact that if the rightmost term vanishes on  $\omega_T$ , then  $\varphi = \mathcal{I}_{\mathcal{T}}^{\text{dG}}\varphi$ , whence  $\varphi$  is continuous in  $\omega_T$ . This yields  $\llbracket \varphi \rrbracket|_F = 0$  for all internal faces  $F$  of  $\omega_T$  and the middle term vanishes.  $\square$

Applying Lemma 10.39, we are led to define the elemental residual indicator

$$\eta_{\mathcal{T}}(\mathbf{u}_{\mathcal{T}}, p_{\mathcal{T}}, T)^2 := \eta_{\mathcal{T}}^m(\mathbf{u}_{\mathcal{T}}, p_{\mathcal{T}}, T)^2 + h_T \sum_{F \subset \partial T \setminus \partial \Omega} \|\llbracket \operatorname{div} \mathbf{u}_{\mathcal{T}} \rrbracket\|_{L^2(F)}^2. \quad (10.77)$$

Concatenating (10.74), (10.76) and Lemma 10.39, we fulfill Assumption 10.22 (equivalence of error and full estimator). The precise result is as follows.

**Proposition 10.40 (a posteriori error analysis for Stokes)** *There exists constants  $C_U \geq C_L > 0$  such that*

$$C_L \mathcal{E}_{\mathcal{T}}(\mathbf{u}_{\mathcal{T}}, p_{\mathcal{T}}, \mathbf{f}) \leq \|\mathbf{u} - \mathbf{u}_{\mathcal{T}}\|_{\mathbb{V}} + \|p - p_{\mathcal{T}}\|_{\mathbb{Q}} \leq C_U \mathcal{E}_{\mathcal{T}}(\mathbf{u}_{\mathcal{T}}, p_{\mathcal{T}}, \mathbf{f}),$$

where the full estimator is defined by

$$\mathcal{E}_{\mathcal{T}}(\mathbf{u}_{\mathcal{T}}, p_{\mathcal{T}}, \mathbf{f})^2 := \sum_{T \in \mathcal{T}} \mathcal{E}_{\mathcal{T}}(\mathbf{u}_{\mathcal{T}}, p_{\mathcal{T}}, \mathbf{f}, T)^2$$

with  $\mathcal{E}_{\mathcal{T}}(\mathbf{u}_{\mathcal{T}}, p_{\mathcal{T}}, \mathbf{f}, T)^2 := \eta_{\mathcal{T}}(\mathbf{u}_{\mathcal{T}}, p_{\mathcal{T}}, T)^2 + \operatorname{osc}_{\mathcal{T}}(\mathbf{f}, T)_{-1}^2$  introduced in (10.77) and (10.75) and  $\operatorname{osc}_{\mathcal{T}}(\mathbf{f}, T)_{-1} = \|\mathbf{f} - P_{\mathcal{T}}\mathbf{f}\|_{H^{-1}(\omega_T)}$ .  $\square$

Since the Stokes problem has constant coefficients but variable forcing, it is natural to resort to Algorithm 5.16 (AFEM-SW), the one-step AFEM with switch, for its adaptive discretization. With respect to the functional setting of Section 10.3.2, the ambient space  $\mathbb{W}$  is  $\mathbb{V} \times \mathbb{Q}$  and the data projection operator  $P_{\mathcal{T}}$  is

$$P_{\mathcal{T}} := ((P_{\mathcal{T}})^d, \Pi_{\mathcal{T}}^{n-1}) : \mathbb{W}^* \rightarrow (\mathbb{F}_{\mathcal{T}})^d \times \mathbb{S}^{n-1,-1},$$

where  $\mathbb{F}_{\mathcal{T}}$  is the scalar discrete space introduced in Definition 4.17,  $P_{\mathcal{T}}$  is here the scalar projection operator introduced in Definition 4.24, and  $\Pi_{\mathcal{T}}^{n-1}$  is the  $L^2$ -orthogonal projection upon  $\mathbb{S}^{n-1,-1}$ . Furthermore, the norm used to measure data perturbations is  $\|(\mathbf{f}, g)\|_{\mathbb{W}_{\mathcal{T}}^*}^2 = \sum_{T \in \mathcal{T}} (\|\mathbf{f}\|_{(H^{-1}(\omega_T))^d}^2 + \|g\|_{L^2(T)}^2)$ .

It is easily seen that  $\eta_{\mathcal{T}}(\mathbf{u}_{\mathcal{T}}, p_{\mathcal{T}}, T)$  satisfies Assumptions 10.17 (Lipschitz continuity of estimator) and 10.18 (monotonicity of estimator) as well as the hypotheses

of Proposition 10.19 (estimator reduction under Dörfler marking): the estimator is clearly Lipschitz continuous and monotone, and it satisfies condition (10.38) since all its addends appear multiplied by a positive power of the meshsize. In addition, the oscillation  $\text{osc}_{\mathcal{T}}((f, g)) = \|(f, g) - P_{\mathcal{T}}(f, g)\|_{\mathbb{W}_{\mathcal{T}}^*}$  fulfills Assumption 10.20 (quasi-monotonicity of oscillation).

Theorem 10.23 (linear convergence for AFEM-SW) provides sufficient conditions for the linear convergence of the algorithm, and these conditions have been verified along the previous discussion. Therefore, we may summarize our findings in the following theorem.

**Theorem 10.41 (linear convergence for Stokes)** *Consider the Galerkin discretization (10.68) of the Stokes problem which uses Taylor-Hood elements of order  $n \geq 2$ , and let the a posteriori estimator be given in Proposition 10.40. Then, Theorem 10.6 guarantees the linear convergence of Algorithm 5.16 (AFEM-SW) applied to this problem, i.e., it holds for some  $c > 0$  and  $0 \leq \rho < 1$*

$$e_{j+i} \leq c\rho^i e_j \quad \forall i, j \in \mathbb{N},$$

with  $e_j := \|\nabla(\mathbf{u} - \mathbf{u}_j)\|_{\Omega} + \|p - p_j\|_{\Omega}$ . □

In order to assess the optimality of the discretization, we specify the definition of approximation classes for the solution of the Stokes problem. Precisely, given  $(\mathbf{v}, q) \in \mathbb{V} \times \mathbb{Q}$  we let  $\sigma_N(\mathbf{v}, q)$  be the smallest approximation error incurred on  $(\mathbf{v}, q)$  with elements in  $\mathbb{V}_{\mathcal{T}} \times \mathbb{Q}_{\mathcal{T}}$  over meshes belonging to  $\mathbb{T}_N$ :

$$\sigma_N(\mathbf{v}, q) := \inf_{\mathcal{T} \in \mathbb{T}_N} \inf_{(\mathbf{v}_{\mathcal{T}}, q_{\mathcal{T}}) \in \mathbb{V}_{\mathcal{T}} \times \mathbb{Q}_{\mathcal{T}}} (\|\nabla(\mathbf{v} - \mathbf{v}_{\mathcal{T}})\|_{\Omega}^2 + \|q - q_{\mathcal{T}}\|_{\Omega}^2)^{1/2}. \quad (10.78)$$

For  $0 < s \leq n/d$ , the class  $\mathbb{A}_s = \mathbb{A}_s(\mathbb{V} \times \mathbb{Q}; \mathcal{T}_0)$ , relative to the partition  $\mathcal{T}_0$  is the set of functions  $(\mathbf{v}, q) \in \mathbb{V} \times \mathbb{Q}$  such that

$$|(\mathbf{v}, q)|_{\mathbb{A}_s} := \sup_{N \geq \#\mathcal{T}_0} (N^s \sigma_N(\mathbf{v}, q)) < \infty \quad (10.79)$$

By adapting the arguments used in the proof of Theorem 6.20 (rate optimality of one-step AFEMs), we can prove the following result.

**Theorem 10.42 (rate optimality of AFEM-SW for Stokes)** *Let the assumptions of Theorem 10.41 be valid. If  $(\mathbf{u}, p) \in \mathbb{A}_s$ , then the sequence  $\{\mathcal{T}_k, \mathbb{V}_k, (\mathbf{u}_k, p_k)\}_{k \geq 0}$  generated by AFEM-SW are such that*

$$\|\nabla(\mathbf{u} - \mathbf{u}_k)\|_{L^2(\Omega)} + \|p - p_k\|_{L^2(\Omega)} \lesssim |(\mathbf{u}, p)|_{\mathbb{A}_s} (\#\mathcal{T}_k)^{-s}, \quad k \geq 0. \quad (10.80)$$

**Remark 10.43 (limits of the analysis)** Other inf-sup stable elements, such as the Mini element or the Crouzeix-Raviart element (see e.g. [Boffi *et al.* 2013]), do not fit in the present setting of analysis, since their velocities contain elementwise bubble components (which are indeed responsible for the stability of the elements). Unfortunately, a bubble on an element does not restricts to two bubbles when the

element is bisected, preventing the nestedness condition  $\mathbb{V}_{\mathcal{T}} \subset \mathbb{V}_{\mathcal{T}_*}$  to be satisfied when  $\mathcal{T}_*$  is a refinement of  $\mathcal{T}$ .

### 10.5 Mixed FEMs for scalar elliptic PDEs

The diffusion-reaction problem (2.5) can be formulated in mixed form as follows:

$$\begin{aligned} \mathbf{A}^{-1} \boldsymbol{\sigma} &= \nabla u & \text{in } \Omega, \\ -\operatorname{div} \boldsymbol{\sigma} + cu &= f & \text{in } \Omega, \\ u &= 0 & \text{on } \partial\Omega. \end{aligned} \quad (10.81)$$

Introducing the porosity matrix  $\mathbf{K} := \mathbf{A}^{-1}$ , we assume hereafter that data

$$\mathcal{D} = (\mathbf{K}, c, f) \in D(\Omega) := M(\alpha_1, \alpha_2) \times R(c_1, c_2) \times L^2(\Omega), \quad (10.82)$$

where  $M(\alpha_1, \alpha_2)$  and  $R(c_1, c_2)$  are defined in (5.48) and (5.49), respectively. Note that the current parameters  $\alpha_1, \alpha_2$  are the reciprocals of  $\alpha_2, \alpha_1$  in (5.48), but to avoid complicating the notation further, we relabel them hereafter.

*Weak formulation.* To write the weak formulation of these equations, we introduce the Hilbert space

$$H(\operatorname{div}; \Omega) := \{ \boldsymbol{\tau} \in L^2(\Omega; \mathbb{R}^d) : \operatorname{div} \boldsymbol{\tau} \in L^2(\Omega) \} \quad (10.83)$$

equipped with the norm  $\| \boldsymbol{\tau} \|_{H(\operatorname{div}; \Omega)}^2 := \| \boldsymbol{\tau} \|_{\Omega}^2 + \| \operatorname{div} \boldsymbol{\tau} \|_{\Omega}^2$ . Then, we multiply the first equation in (10.81) by  $\boldsymbol{\tau} \in H(\operatorname{div}; \Omega)$  and the second equation by  $v \in L^2(\Omega)$ , we integrate over  $\Omega$  and apply partial integration to the term containing  $\nabla u$ , keeping into account the Dirichlet boundary condition. In this way, we obtain the following variational problem: find  $(\boldsymbol{\sigma}, u) \in \mathbb{V} := H(\operatorname{div}; \Omega) \times L^2(\Omega)$  such that

$$\begin{aligned} \int_{\Omega} \mathbf{K} \boldsymbol{\sigma} \cdot \boldsymbol{\tau} + \int_{\Omega} u \operatorname{div} \boldsymbol{\tau} &= 0 & \forall \boldsymbol{\tau} \in H(\operatorname{div}; \Omega), \\ \int_{\Omega} v \operatorname{div} \boldsymbol{\sigma} - \int_{\Omega} c u v &= - \int_{\Omega} f v & \forall v \in L^2(\Omega). \end{aligned} \quad (10.84)$$

This can be written as: find  $(\boldsymbol{\sigma}, u) \in V \times Q$  such that

$$\begin{aligned} a[\boldsymbol{\sigma}, \boldsymbol{\tau}] + b[u, \boldsymbol{\tau}] &= 0 & \forall \boldsymbol{\sigma} \in V, \\ b[v, \boldsymbol{\sigma}] + m[u, v] &= -\langle f, v \rangle & \forall v \in Q, \end{aligned} \quad (10.85)$$

if we set  $V := H(\operatorname{div}; \Omega)$ ,  $Q := L^2(\Omega)$ , and we introduce the continuous bilinear forms  $a: V \times V \rightarrow \mathbb{R}$ ,  $b: Q \times V \rightarrow \mathbb{R}$  and  $m: Q \times Q \rightarrow \mathbb{R}$  by

$$a[\boldsymbol{\sigma}, \boldsymbol{\tau}] = \int_{\Omega} \mathbf{K} \boldsymbol{\sigma} \cdot \boldsymbol{\tau}, \quad b[v, \boldsymbol{\tau}] = \int_{\Omega} v \operatorname{div} \boldsymbol{\tau}, \quad m[u, v] = - \int_{\Omega} c u v,$$

and the linear form  $\langle f, v \rangle = \int_{\Omega} f v$ . An equivalent formulation, similar to (2.16), is as follows:

$$(\boldsymbol{\sigma}, u) \in V \times Q : \quad \mathcal{B}[(\boldsymbol{\sigma}, u), (\boldsymbol{\tau}, v)] = -\langle f, v \rangle \quad \forall (\boldsymbol{\tau}, v) \in V \times Q, \quad (10.86)$$

with

$$\mathcal{B}[(\sigma, u), (\tau, v)] := a[\sigma, \tau] + b[u, \tau] + b[v, \sigma] + m[u, v].$$

Formulation (10.85) is a generalization of the classical saddle point problem considered in Section 2.4, given by the presence of the third bilinear form  $m$ . According to [Boffi *et al.* 2013, Theorem 4.3.1], the well-posedness of such a problem can be derived from the three following conditions:

- (i) the form  $a$  is coercive on  $V_0 = \{\tau \in V : b[v, \tau] = 0 \text{ for all } v \in Q\}$ ,
- (ii) the form  $b$  satisfies an inf-sup condition on  $V \times Q$ ,
- (iii) the form  $m$  is non-positive on  $Q$ , i.e.,  $m[v, v] \leq 0$  for all  $v \in Q$ .

These conditions are easily checked for our mixed formulation of the Dirichlet problem.

*Discretization.* To define a finite element discretization of this problem, we consider partitions  $\mathcal{T} \in \mathbb{T}$  obtained by conforming bisection refinements of an initial partition  $\mathcal{T}_0$ , and let  $V_{\mathcal{T}} \subset V$  and  $Q_{\mathcal{T}} \subset Q$  be finite dimensional subspaces made of piecewise polynomial functions on  $\mathcal{T}$ . Among the families of *uniformly inf-sup stable* finite element spaces for this problem, we consider the Raviart-Thomas-Nédélec family [Raviart and Thomas 1977, Nédélec 1980], and the Brezzi-Douglas-Marini family [Brezzi, Douglas and Marini 1985] on simplicial elements. They fit into the general definition

$$\begin{aligned} V_{\mathcal{T}} &= \{\tau \in H(\text{div}; \Omega) : \tau|_T \in \mathbf{V}_T, T \in \mathcal{T}\}, \\ Q_{\mathcal{T}} &= \{q \in L^2(\Omega) : q|_T \in Q_T, T \in \mathcal{T}\}. \end{aligned}$$

For the Raviart-Thomas-Nédélec (RTN) family we have

$$\mathbf{V}_T = (\mathbb{P}_{n-1}(T))^d + \mathbf{x} \mathbb{P}_{n-1}(T), \quad Q_T = \mathbb{P}_{n-1}(T), \quad n \geq 1,$$

where  $\mathbf{x} = (x_1, \dots, x_d)$  is the coordinate vector, whereas for the Brezzi-Douglas-Marini (BDM) family we have

$$\mathbf{V}_T = (\mathbb{P}_n(T))^d, \quad Q_T = \mathbb{P}_{n-1}(T), \quad n \geq 1.$$

Note that for any face  $F$  of the triangulation it holds  $\tau|_F \cdot \mathbf{n}_F \in \mathbb{P}_{n-1}(F)$  for the RTN family, and  $\tau|_F \cdot \mathbf{n}_F \in \mathbb{P}_n(F)$  for the BDM family; furthermore,  $\text{div } \mathbf{V}_T = Q_T$ . We refer to [Boffi *et al.* 2013, Sects. 2.3.1 and 7.1.2] for more details.

Due to the presence of variable data, it is natural to perform the adaptive discretization of the problem by adopting Algorithm 5.28 (AFEM-TS), the two-step AFEM. The procedure  $[\widehat{\mathcal{T}}, \widehat{\mathcal{D}}] = \text{DATA}(\mathcal{T}, \mathcal{D}, \tau)$  generates an admissible refinement  $\widehat{\mathcal{T}}$  of  $\mathcal{T}$  and discrete data

$$\widehat{\mathcal{D}} = (\widehat{\mathbf{K}}, \widehat{c}, \widehat{f}) \in \mathbb{D}_{\widehat{\mathcal{T}}} := [\mathbb{S}_{\widehat{\mathcal{T}}}^{n-1, -1}]^{d \times d} \times \mathbb{S}_{\widehat{\mathcal{T}}}^{n-1, -1} \times \mathbb{S}_{\widehat{\mathcal{T}}}^{n-1, -1}$$

over  $\widehat{\mathcal{T}}$ , such that  $\widehat{\mathbf{K}} \in M(\widehat{\alpha}_1, \widehat{\alpha}_2)$ ,  $\widehat{c} \in R(\widehat{c}_1, \widehat{c}_2)$  (see Sects. 7.2.2 and 7.2.3), and

$$\|\mathcal{D} - \widehat{\mathcal{D}}\|_{\widehat{\mathcal{D}}(\Omega)} \leq C_{\text{data}} \tau,$$

where the space  $\widehat{D}(\Omega)$  is defined in (5.55).

The Galerkin discretization with these discrete data reads: find  $(\sigma_{\mathcal{T}}, u_{\mathcal{T}}) \in V_{\mathcal{T}} \times Q_{\mathcal{T}}$  such that

$$\begin{aligned} \widehat{a}[\sigma_{\mathcal{T}}, \tau] + b[u_{\mathcal{T}}, \tau] &= 0 & \forall \tau \in V_{\mathcal{T}}, \\ b[v, \sigma_{\mathcal{T}}] + \widehat{m}[u_{\mathcal{T}}, v] &= -\langle \widehat{f}, v \rangle & \forall v \in Q_{\mathcal{T}}, \end{aligned} \quad (10.87)$$

with

$$\widehat{a}[\sigma, \tau] = \int_{\Omega} \widehat{\mathbf{K}} \sigma \cdot \tau, \quad \widehat{m}[u, v] = - \int_{\Omega} \widehat{c} u v, \quad \langle \widehat{f}, v \rangle = \int_{\Omega} \widehat{f} v,$$

or equivalently,

$$(\sigma_{\mathcal{T}}, u_{\mathcal{T}}) \in V_{\mathcal{T}} \times Q_{\mathcal{T}} : \quad \widehat{\mathcal{B}}[(\sigma_{\mathcal{T}}, u_{\mathcal{T}}), (\tau, v)] = -\langle \widehat{f}, v \rangle \quad \forall (\tau, v) \in V_{\mathcal{T}} \times Q_{\mathcal{T}}.$$

*A posteriori error estimator.* Let us denote by  $(\widehat{\sigma}, \widehat{u}) \in V \times Q$  the exact solution of the perturbed problem

$$\widehat{\mathcal{B}}[(\widehat{\sigma}, \widehat{u}), (\tau, v)] = -\langle \widehat{f}, v \rangle \quad \forall (\tau, v) \in V \times Q;$$

note that the forcing  $\widehat{f}$  appears with a negative sign. Then, by continuity and uniform inf-sup stability of the form  $\widehat{\mathcal{B}}$ , we know that the error

$$\|\widehat{\sigma} - \sigma_{\mathcal{T}}\|_{H(\text{div}; \Omega)} + \|\widehat{u} - u_{\mathcal{T}}\|_{L^2(\Omega)}$$

is equivalent to the quantity

$$\sup_{(\tau, v) \in V \times Q} \frac{\widehat{\mathcal{B}}[(\widehat{\sigma} - \sigma_{\mathcal{T}}, \widehat{u} - u_{\mathcal{T}}), (\tau, v)]}{\|\tau\|_{H(\text{div}; \Omega)} + \|v\|_{L^2(\Omega)}} = \sup_{(\tau, v) \in V \times Q} \frac{\langle \widehat{f}, v \rangle + \widehat{\mathcal{B}}[(\sigma_{\mathcal{T}}, u_{\mathcal{T}}), (\tau, v)]}{\|\tau\|_{H(\text{div}; \Omega)} + \|v\|_{L^2(\Omega)}}.$$

By Galerkin orthogonality, the numerator is equal to

$$\langle \widehat{f}, v - v_{\mathcal{T}} \rangle + \widehat{\mathcal{B}}[(\sigma_{\mathcal{T}}, u_{\mathcal{T}}), (\tau - \tau_{\mathcal{T}}, v - v_{\mathcal{T}})] \quad \forall (\tau_{\mathcal{T}}, v_{\mathcal{T}}) \in V_{\mathcal{T}} \times Q_{\mathcal{T}},$$

that we now proceed to estimate. The term

$$\widehat{\mathcal{B}}[(\sigma_{\mathcal{T}}, u_{\mathcal{T}}), (\tau - \tau_{\mathcal{T}}, 0)]$$

can be analyzed as in Carstensen [Carstensen 1997] (see also [Verfürth 2013, Section 4.8]), by resorting to a stable decomposition of  $H(\text{div}; \Omega)$ : precisely, given  $\tau \in H(\text{div}; \Omega)$ , there exist  $\Phi \in (H^1(\Omega))^d$  and  $u \in (H^1(\Omega))^d$  such that

$$\tau = \Phi + \text{curl } u \quad (10.88)$$

with  $\|\Phi\|_{(H^1(\Omega))^d} + \|u\|_{(H^1(\Omega))^d} \lesssim \|\tau\|_{H(\text{div}; \Omega)}$  (see [Xu, Chen and Nochetto 2009, Section 5.1.3]). Note that if  $\Omega$  is convex, then (10.88) is the Helmholtz decomposition of  $\tau$ , with  $\Phi = \nabla G$  for some  $G \in (H^2(\Omega))^d$ . Using (10.88) and a suitable choice of  $\tau_{\mathcal{T}}$ , one can show that

$$|\widehat{\mathcal{B}}[(\sigma_{\mathcal{T}}, u_{\mathcal{T}}), (\tau - \tau_{\mathcal{T}}, 0)]| \lesssim \eta_{\mathcal{T}, 1}((\sigma_{\mathcal{T}}, u_{\mathcal{T}})) \|\tau\|_{H(\text{div}; \Omega)} \quad (10.89)$$

with

$$\eta_{\mathcal{T},1}((\boldsymbol{\sigma}_{\mathcal{T}}, u_{\mathcal{T}}))^2 = \sum_{T \in \mathcal{T}} \eta_{\mathcal{T},1}((\boldsymbol{\sigma}_{\mathcal{T}}, u_{\mathcal{T}}), T)^2$$

and

$$\begin{aligned} \eta_{\mathcal{T},1}((\boldsymbol{\sigma}_{\mathcal{T}}, u_{\mathcal{T}}), T)^2 &:= h_T^2 \|\widehat{\mathbf{K}} \boldsymbol{\sigma}_{\mathcal{T}} - \nabla u_{\mathcal{T}}\|_{L^2(T)}^2 + h_T^2 \|\operatorname{curl}(\widehat{\mathbf{K}} \boldsymbol{\sigma}_{\mathcal{T}})\|_{L^2(T)}^2 \\ &\quad + h_T \sum_{F \subset \partial T \setminus \partial \Omega} \left\| \left[ (\widehat{\mathbf{K}} \boldsymbol{\sigma}_{\mathcal{T}})_t \right] \right\|_{L^2(F)}^2 + h_T \sum_{F \subset \partial T \cap \partial \Omega} \|(\widehat{\mathbf{K}} \boldsymbol{\sigma}_{\mathcal{T}})_t\|_{L^2(F)}^2, \end{aligned} \quad (10.90)$$

where  $\boldsymbol{\phi}_t = \boldsymbol{\phi} - (\boldsymbol{\phi} \cdot \mathbf{n}_F) \mathbf{n}_F$  denotes the tangential component of the vector field  $\boldsymbol{\phi}$  on  $F$ . On the other hand, the term

$$\langle \widehat{f}, v - v_{\mathcal{T}} \rangle + \widehat{\mathcal{B}}[(\boldsymbol{\sigma}_{\mathcal{T}}, u_{\mathcal{T}}), (0, v - v_{\mathcal{T}})] = \int_{\Omega} (\widehat{f} + \operatorname{div} \boldsymbol{\sigma}_{\mathcal{T}} - \widehat{c} u_{\mathcal{T}})(v - v_{\mathcal{T}})$$

can be bounded as follows. For any  $T \in \mathcal{T}$ , let  $\Pi_T = \Pi_T^{n-1}$  be the  $L^2$ -orthogonal projection upon  $Q_T = \mathbb{P}_{n-1}(T)$ , and let us choose  $(v_{\mathcal{T}})|_T = \Pi_T v$ . Then, noticing that  $\widehat{f} + \operatorname{div} \boldsymbol{\sigma}_{\mathcal{T}} \in Q_T$ , we have

$$\begin{aligned} \int_{\Omega} (\widehat{f} + \operatorname{div} \boldsymbol{\sigma}_{\mathcal{T}} - \widehat{c} u_{\mathcal{T}})(v - v_{\mathcal{T}}) &= \sum_{T \in \mathcal{T}} \int_T (\widehat{f} + \operatorname{div} \boldsymbol{\sigma}_{\mathcal{T}} - \Pi_T(\widehat{c} u_{\mathcal{T}}))(v - \Pi_T v) \\ &\quad - \sum_{T \in \mathcal{T}} \int_T (\widehat{c} u_{\mathcal{T}} - \Pi_T(\widehat{c} u_{\mathcal{T}}))(v - \Pi_T v) \\ &= - \sum_{T \in \mathcal{T}} \int_T (\widehat{c} u_{\mathcal{T}} - \Pi_T(\widehat{c} u_{\mathcal{T}}))v, \end{aligned} \quad (10.91)$$

whence

$$|\langle \widehat{f}, v - v_{\mathcal{T}} \rangle + \widehat{\mathcal{B}}[(\boldsymbol{\sigma}_{\mathcal{T}}, u_{\mathcal{T}}), (0, v - v_{\mathcal{T}})]| \leq \sum_{T \in \mathcal{T}} \|\widehat{c} u_{\mathcal{T}} - \Pi_T(\widehat{c} u_{\mathcal{T}})\|_{L^2(T)} \|v\|_{L^2(T)}.$$

Conversely, it is easily checked that (10.91) implies the bound

$$\|\widehat{c} u_{\mathcal{T}} - \Pi_T(\widehat{c} u_{\mathcal{T}})\|_{L^2(T)} \lesssim \|\operatorname{div} \boldsymbol{\sigma} - \operatorname{div} \boldsymbol{\sigma}_{\mathcal{T}}\|_{L^2(T)} + \|\widehat{c}\|_{L^\infty(T)} \|\widehat{u} - u_{\mathcal{T}}\|_{L^2(T)}.$$

The choice  $n = 1$  yields  $\widehat{c} u_{\mathcal{T}} \in \mathbb{P}_0(T)$ , hence,  $\widehat{c} u_{\mathcal{T}} - \Pi_T(\widehat{c} u_{\mathcal{T}}) = 0$ . For  $n \geq 2$ , we could define as a (squared) local error indicator the quantity

$$\eta_{\mathcal{T},1}((\boldsymbol{\sigma}_{\mathcal{T}}, u_{\mathcal{T}}), T)^2 + \|\widehat{c} u_{\mathcal{T}} - \Pi_T(\widehat{c} u_{\mathcal{T}})\|_{L^2(T)}^2,$$

but the second addend is not guaranteed to reduce under refinement, since it is not scaled by a positive power of the meshsize. However, there is an equivalent quantity which does reduce, as stated in the following result.

**Lemma 10.44 (equivalence of local error indicators)** Assume  $\widehat{c}, u_{\mathcal{T}} \in \mathbb{P}_{n-1}(T)$ ,



for  $n \geq 2$ . Let  $\Pi_T^j$  be the  $L^2$ -orthogonal projection upon  $\mathbb{P}_j(T)$ . Then,

$$\|\widehat{c}u_{\mathcal{T}} - \Pi_T(\widehat{c}u_{\mathcal{T}})\|_{L^2(T)} \approx h_T \sum_{j=1}^n \|u_{\mathcal{T}} - \Pi_T^{n-1-j}u_{\mathcal{T}}\|_{L^2(T)} \|\nabla \widehat{c} - \Pi_T^{j-2}\nabla \widehat{c}\|_{L^\infty(T)}, \quad (10.92)$$

where the constants hidden in the symbol  $\approx$  are independent of  $\widehat{c}$ ,  $u_{\mathcal{T}}$ , and  $T$ .

*Proof.* By the Bramble-Hilbert theorem,

$$\|\widehat{c}u_{\mathcal{T}} - \Pi_T(\widehat{c}u_{\mathcal{T}})\|_{L^2(T)} \lesssim h_T^n |\widehat{c}u_{\mathcal{T}}|_{H^n(T)}$$

and

$$|\widehat{c}u_{\mathcal{T}}|_{H^n(T)} \lesssim \sum_{j=1}^{n-1} |\widehat{c}|_{W_\infty^j(T)} |u_{\mathcal{T}}|_{H^{n-j}(T)}.$$

Moreover,

$$|\widehat{c}|_{W_\infty^j(T)} = |\nabla \widehat{c}|_{W_\infty^{j-1}(T)} = |\nabla \widehat{c} - \Pi_T^{j-2}\nabla \widehat{c}|_{W_\infty^{j-1}(T)}$$

and

$$|u_{\mathcal{T}}|_{H^{n-j}(T)} = |u_{\mathcal{T}} - \Pi_T^{n-1-j}u_{\mathcal{T}}|_{H^{n-j}(T)}.$$

Applying inverse estimates for semi-norms, we obtain the  $\lesssim$  inequality in (10.92).

To get the opposite inequality, it is enough to check that the vanishing of the left-hand side implies the vanishing of the right-hand side, since both quantities are defined on finite-dimensional spaces and their scaling with respect to the element size is the same. Now,  $\widehat{c}u_{\mathcal{T}} = \Pi_T(\widehat{c}u_{\mathcal{T}})$  implies  $\widehat{c}u_{\mathcal{T}} \in \mathbb{P}_{n-1}(T)$ . Let us assume that  $u_{\mathcal{T}} \in \mathbb{P}_{n-1-k}(T)$  for some  $0 \leq k \leq n-1$ , and consequently  $\widehat{c} \in \mathbb{P}_k(T)$ , i.e.,  $\nabla \widehat{c} \in \mathbb{P}_{k-1}(T)$ . Then,  $\Pi_T^{n-1-j}u_{\mathcal{T}} = u_{\mathcal{T}}$  for any  $j \leq k$ , hence, the corresponding differences in the summation on the right-hand side of (10.92) vanish. Conversely, for  $j > k$  we have  $j-2 \geq k-1$ , which implies  $\Pi_T^{j-2}\nabla \widehat{c} = \nabla \widehat{c}$ , that is the corresponding differences in the summation on the right-hand side vanish. In conclusion, all terms in the summation in (10.92) vanish, and the thesis is proven.  $\square$

Summarizing, we have obtained the following result.

**Proposition 10.45 (a posteriori error estimator for mixed methods)** *For every  $T \in \mathcal{T}$ , the following local quantity*

$$\begin{aligned} \eta_{\mathcal{T}}((\sigma_{\mathcal{T}}, u_{\mathcal{T}}), T)^2 := & h_T^2 \|\widehat{\mathbf{K}}\sigma_{\mathcal{T}} - \nabla u_{\mathcal{T}}\|_{L^2(T)}^2 + h_T^2 \|\operatorname{curl}(\widehat{\mathbf{K}}\sigma_{\mathcal{T}})\|_{L^2(T)}^2 \\ & + h_T \sum_{F \subset \partial T \setminus \partial \Omega} \|[(\widehat{\mathbf{K}}\sigma_{\mathcal{T}})_t]_1\|_{L^2(F)}^2 + h_T \sum_{F \subset \partial T \cap \partial \Omega} \|(\widehat{\mathbf{K}}\sigma_{\mathcal{T}})_t\|_{L^2(F)}^2 \\ & + h_T^2 \sum_{j=1}^n \|u_{\mathcal{T}} - \Pi_T^{n-1-j}u_{\mathcal{T}}\|_{L^2(T)}^2 \|\nabla \widehat{c} - \Pi_T^{j-2}\nabla \widehat{c}\|_{L^\infty(T)}^2 \end{aligned} \quad (10.93)$$

is a (squared) *a posteriori* error indicator for the mixed problem (10.87), which gives rise to a global *a posteriori* error estimator  $\eta_{\mathcal{T}}(\boldsymbol{\sigma}_{\mathcal{T}}, u_{\mathcal{T}})$  that satisfies Assumption 10.2.  $\square$

Finally, Assumption 10.3 follows from Proposition 10.19 (estimator reduction under Dörfler marking), since the estimator is clearly Lipschitz continuous and monotone, and it satisfies condition (10.38) since all its addends are scaled by positive powers of the meshsize.

As a consequence, the GALERKIN step in AFEM-TS converges linearly by Theorem 10.6, and the number of sub-iterations in the  $k$ -th call to GALERKIN is bounded by a constant  $J$  independent of  $k$  (Proposition 10.27). Furthermore, Theorem 10.38 guarantees the quasi-optimality of the two-step AFEM.

**Theorem 10.46 (quasi-optimality of AFEM-TS for mixed methods)** *Let the exact solution  $(\boldsymbol{\sigma}, u)$  of the mixed problem (10.84) belong to the approximation class  $\mathbb{A}_{s_u}(V; \mathcal{T}_0)$  and let the data  $(\mathbf{K}, c, f)$  belong to the approximation class  $\mathbb{D}_{s_D}(D(\Omega); \mathcal{T}_0)$ . Let Assumptions 6.14 (marking parameter), 6.21 (size of  $\omega$ ), and 6.19 (initial labeling) be valid. Consider the Galerkin discretization (10.87) based on one of the Raviart-Thomas-Nédélec or Brezzi-Douglas-Marini finite element pairs. There exists a constant  $C_*$  independent of the exact solution  $(\boldsymbol{\sigma}, u)$  and the data  $\mathcal{D} = (\mathbf{K}, c, f)$  such that the sequence  $\{(\mathcal{T}_k, \mathbb{V}_{\mathcal{T}_k} \times \mathbb{Q}_{\mathcal{T}_k}, (\boldsymbol{\sigma}_{\mathcal{T}_k}, u_{\mathcal{T}_k}))\}_{k \geq 0}$  produced by AFEM-TS satisfies for  $k \geq 0$*

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_{\mathcal{T}_k}\|_{H(\operatorname{div}; \Omega)} + \|u - u_{\mathcal{T}_k}\|_{L^2(\Omega)} \leq C_* \left( |(\boldsymbol{\sigma}, u)|_{\mathbb{A}_{s_u}}^{1/s_u} + |\mathcal{D}|_{\mathbb{D}_{s_D}}^{1/s_D} \right)^s (\#\mathcal{T}_k)^{-s},$$

with  $0 < s = \min(s_u, s_D) \leq \frac{n}{d}$ .  $\square$

**Remark 10.47** Another family of uniformly inf-sup stable spaces is the Brezzi-Douglas-Fortin-Marini's (see [Brezzi, Douglas, Fortin and Marini 1987]), where  $\mathbb{V}_T = \{\boldsymbol{\tau} \in (\mathbb{P}_n(T))^d : \boldsymbol{\tau} \cdot \mathbf{n}_F \in \mathbb{P}_{n-1}(F) \forall F \in \mathcal{F} \cap \partial T\}$  and  $\mathbb{Q}_T = \mathbb{P}_n(T)$ ,  $n \geq 1$ . However, the imposed condition on the normal component of vector fields on each face of  $\mathcal{T}$  prevents the inclusion of  $\mathbb{V}_{\mathcal{T}}$  into  $\mathbb{V}_{\mathcal{T}_*}$  to hold if  $\mathcal{T}_*$  is a bisection refinement of  $\mathcal{T}$ .

### 10.6 Proof of Theorem 10.15

This section is devoted to establishing Theorem 10.15, which in turn contributes with Corollary 10.14 to the proof of Theorem 10.6.

It is important to notice that the growth of  $\|\mathbf{U}\|_2$  is dictated by the number of blocks  $N$  rather than the actual dimension  $n_N \gg N$  of  $\mathbf{U}$ . Therefore, we again use the block notation from Section 10.2

$$\mathbf{B} = (\mathbf{B}(i, j))_{i,j=0}^N \in \mathbb{R}^{n_N \times n_N},$$

with lower and upper triangular factors

$$\mathbf{L} = (\mathbf{L}(i, j))_{i,j=0}^N \in \mathbb{R}^{n_N \times n_N}, \quad \mathbf{U} = (\mathbf{U}(i, j))_{i,j=0}^N \in \mathbb{R}^{n_N \times n_N}.$$

We also set

$$\mathbf{A} = (\mathbf{A}(i, j))_{i,j=0}^N := \sqrt{\alpha} \mathbf{B}$$

for a suitable parameter  $\alpha > 0$  defined below.

### 10.6.1 Representation of block inverse matrices

We first show that it suffices to derive the estimates

$$\|\mathbf{U}^{-1}\|_2 \lesssim N^{-1/p} \quad p > 2, \quad (10.94)$$

$$\|\tilde{\mathbf{U}}^{-1}\|_2 \lesssim N^{-1/p} \quad p > 2, \quad (10.95)$$

$$\|\mathbf{D}\|_2 \lesssim 1, \quad (10.96)$$

where  $\mathbf{B}^T = \tilde{\mathbf{L}}\tilde{\mathbf{U}}$  is the normalized block triangular decomposition of  $\mathbf{B}^T$ , and  $\mathbf{D} \in \mathbb{R}^{n_N \times n_N}$  stands for the block diagonal part of  $\mathbf{U}$ ,

In fact, in view of property (P1) (continuity of  $\mathcal{B}$ ), we see that

$$\|\mathbf{B}\|_2 = \|\mathbf{B}^T\|_2 \leq \|\mathcal{B}\|,$$

whence

$$\mathbf{L} = \mathbf{B}\mathbf{U}^{-1} \Rightarrow \|\mathbf{L}\|_2 \leq \|\mathbf{B}\|_2 \|\mathbf{U}^{-1}\|_2 \leq \|\mathcal{B}\| \|\mathbf{U}^{-1}\|_2$$

and, similarly,

$$\|\tilde{\mathbf{L}}\|_2 \leq \|\mathcal{B}\| \|\tilde{\mathbf{U}}^{-1}\|_2.$$

On the other hand, from

$$\mathbf{B} = \mathbf{L}\mathbf{D} (\mathbf{D}^{-1}\mathbf{U}) \quad \Rightarrow \quad \mathbf{B}^T = (\mathbf{U}^T \mathbf{D}^{-1}) \mathbf{D} \mathbf{L}^T$$

we infer that

$$\begin{aligned} \tilde{\mathbf{L}} = \mathbf{U}^T \mathbf{D}^{-1} &\Rightarrow \mathbf{U} = \mathbf{D} \tilde{\mathbf{L}}^T \\ \tilde{\mathbf{U}} = \mathbf{D} \mathbf{L}^T &\Rightarrow \mathbf{L}^{-1} = \mathbf{D} \tilde{\mathbf{U}}^{-T}, \end{aligned} \quad (10.97)$$

which implies

$$\begin{aligned} \|\mathbf{U}\|_2 &\leq \|\mathbf{D}\|_2 \|\tilde{\mathbf{L}}^T\|_2, \\ \|\mathbf{L}^{-1}\|_2 &\leq \|\mathbf{D}\|_2 \|\tilde{\mathbf{U}}^{-T}\|_2. \end{aligned}$$

Therefore, we can focus on proving (10.94) and (10.96), since the proof of (10.95) is identical to that of (10.94). We proceed in several steps. The most delicate estimate is (10.94).

$\square$  *j*-th column of  $\mathbf{U}^{-1}$ . To prove (10.94), it turns out to be convenient to get first an explicit expression for the *j*-th column of  $\mathbf{U}^{-1}$ . We achieve this next.

**Lemma 10.48 (representation of the  $j$ -th column of  $U^{-1}$ )** *We have*

$$U^{-1}(i, j) = \mathbf{B}[j]^{-1}(i, j) \quad \forall 0 \leq i \leq j \leq N. \quad (10.98)$$

*Proof.* We compute the  $(i, j)$  block of  $\mathbf{B}[j]^{-1} = \mathbf{U}[j]^{-1} \mathbf{L}[j]^{-1}$

$$\mathbf{B}[j]^{-1}(i, j) = \sum_{k=0}^j U^{-1}[j](i, k) \mathbf{L}^{-1}[j](k, j) = \mathbf{U}[j]^{-1}(i, j)$$

because  $\mathbf{L}^{-1}(k, j) = \mathbf{0}$  for  $k < j$  and  $\mathbf{L}^{-1}(j, j) = \mathbf{I}(j, j)$ . Moreover, we claim that

$$U^{-1}[j](i, j) = U^{-1}(i, j) \quad i \leq j$$

because  $U^{-1}$  is block upper triangular. To see this, let

$$\mathbf{x}(:, j) = U^{-1}(:, j) \in \mathbb{R}^{n_N \times d_j}$$

be the  $j$ -th block column of  $U^{-1}$ , which satisfies

$$U\mathbf{x}(:, j) = \mathbf{I}(:, j) \in \mathbb{R}^{n_N \times d_j}.$$

Since  $\mathbf{I}(i, j) = \mathbf{0}$  for  $i > j$  and  $U$  is block upper triangular, we have  $\mathbf{x}(i, j) = \mathbf{0}$  for  $i > j$ . Therefore, the matrix

$$\tilde{\mathbf{x}}(:, j) = (\mathbf{x}(i, j))_{i=0}^j \in \mathbb{R}^{n_j \times d_j}$$

with the first  $j$  blocks of  $\mathbf{x}(:, j)$  satisfies the reduced system

$$\begin{aligned} U[j](j, j) \tilde{\mathbf{x}}(j, j) &= \mathbf{I}(j, j) \\ \sum_{k=i}^j U[j](i, k) \tilde{\mathbf{x}}(k, j) &= \mathbf{0} \quad 0 \leq i \leq j. \end{aligned}$$

We thus deduce that  $\tilde{\mathbf{x}}(:, j) = U[j]^{-1}(:, j)$ , as asserted.  $\square$

This lemma justifies dealing with  $\mathbf{B}[j]^{-1}$ .

$\square$  *Representation of  $\mathbf{B}[j]^{-1}$ .* We resort to the Neumann series expansion. We first consider the uniform SPD matrix

$$\mathbf{B}[j] \mathbf{B}[j]^T \in \mathbb{R}^{n_j \times n_j}$$

for which there exists  $\alpha > 0$  such that

$$\|\mathbf{I}[j] - \alpha \mathbf{B}[j] \mathbf{B}[j]^T\|_2 < 1$$

uniformly in  $j$ . In fact, note that for  $\mathbf{x} \in \mathbb{R}^{n_j \times n_j}$

$$\|\mathbf{x} - \alpha \mathbf{B}[j] \mathbf{B}[j]^T \mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2 - 2\alpha \langle \mathbf{x}, \mathbf{B}[j] \mathbf{B}[j]^T \mathbf{x} \rangle + \alpha^2 \|\mathbf{B}[j] \mathbf{B}[j]^T \mathbf{x}\|_2^2$$

as well as

$$\langle \mathbf{x}, \mathbf{B}[j] \mathbf{B}[j]^T \mathbf{x} \rangle = \|\mathbf{B}[j]^T \mathbf{x}\|_2^2 \geq \beta^2 \|\mathbf{x}\|_2^2$$

in view of property (P2) (discrete inf-sup) and (3.2), and

$$\|\mathbf{B}[j]\mathbf{B}[j]^T \mathbf{x}\|_2 \leq \|\mathbf{B}[j]\|_2^2 \|\mathbf{x}\|_2 \leq \|\mathcal{B}\|^2 \|\mathbf{x}\|_2.$$

Consequently

$$\|\mathbf{x} - \alpha \mathbf{B}[j]\mathbf{B}[j]^T \mathbf{x}\|_2^2 \leq (1 - 2\alpha\beta^2 + \alpha^2 \|\mathcal{B}\|^4) \|\mathbf{x}\|_2^2.$$

The quadratic polynomial in  $\alpha$  on the right-hand side is minimized by  $\alpha = \frac{\beta^2}{\|\mathcal{B}\|^4}$ , and gives

$$\|\mathbf{I}[j] - \alpha \mathbf{B}[j]\mathbf{B}[j]^T\|_2^2 \leq 1 - \frac{\beta^4}{\|\mathcal{B}\|^4} =: \rho^2. \quad (10.99)$$

From now on, we fix this value of  $\alpha$  and assume the uniform bound (10.99). Let

$$\mathbf{A}[j] = \sqrt{\alpha} \mathbf{B}[j] \in \mathbb{R}^{n_j \times n_j}$$

be the  $j$ -th principal section of the matrix  $\mathbf{A}$  introduced previously, and let

$$\mathbf{G}[j] := \mathbf{I}[j] - \mathbf{A}[j]\mathbf{A}[j]^T \in \mathbb{R}^{n_j \times n_j}. \quad (10.100)$$

**Lemma 10.49 (representation of  $\mathbf{B}[j]^{-1}$ )** *The following expression is valid*

$$\mathbf{B}[j]^{-1} = \alpha \mathbf{B}[j]^T \sum_{m=0}^{\infty} \mathbf{G}[j]^m \quad \forall 0 \leq j \leq N. \quad (10.101)$$

*Proof.* Since  $\|\mathbf{G}[j]\|_2 \leq \rho < 1$  according to (10.99), the Neumann series theorem guarantees that

$$\mathbf{A}[j]\mathbf{A}[j]^T = \mathbf{I}[j] - \mathbf{G}[j]$$

is invertible and the inverse reads

$$\mathbf{A}[j]^{-T} \mathbf{A}[j]^{-1} = \sum_{m=0}^{\infty} \mathbf{G}[j]^m,$$

where  $\mathbf{G}[j]^0 = \mathbf{I}[j]$ . Multiplying on the left by  $\mathbf{A}[j]^T$ , we obtain

$$\frac{1}{\sqrt{\alpha}} \mathbf{B}[j]^{-1} = \sqrt{\alpha} \mathbf{B}[j] \sum_{m=0}^{\infty} \mathbf{G}[j]^m,$$

which yields the assertion.  $\square$

$\square$  *Representation of  $\mathbf{U}^{-1}$ .* In order to obtain a representation of  $\mathbf{U}^{-1}$ , we now build on (10.98), which gives a formula for the  $j$ -th column of  $\mathbf{U}^{-1}$  in terms of  $\mathbf{B}[j]^{-1}$ , and (10.101), which provides a series representation of  $\mathbf{B}[j]^{-1}$ . To this end, we introduce the block upper triangular matrix  $\mathbf{G}_m \in \mathbb{R}^{n_N \times n_N}$  given by

$$\mathbf{G}_m(i, j) := \begin{cases} \mathbf{G}[j]^m(i, j) & i \leq j \\ 0 & i > j \end{cases}$$

for  $m \geq 1$  and  $\mathbf{G}_0 = \mathbf{I}$ . Hence,

$$\mathbf{U}^{-1}(i, j) = \begin{cases} \alpha \left( \mathbf{B}[j]^T \sum_{m=0}^{\infty} \mathbf{G}_m[j] \right) (i, j) & i \leq j \\ 0 & i > j. \end{cases}$$

To write this expression in compact form, it is convenient to introduce the *block upper triangular truncation operator*  $\mathcal{U} : \mathbb{R}^{n_N \times n_N} \rightarrow \mathbb{R}^{n_N \times n_N}$  defined by

$$\mathcal{U}(\mathbf{M})(i, j) := \begin{cases} \mathbf{M}(i, j) & i \leq j \\ 0 & i > j \end{cases} \quad \forall \mathbf{M} \in \mathbb{R}^{n_N \times n_N}.$$

**Lemma 10.50 (representation of  $\mathbf{U}^{-1}$ )** *There holds*

$$\mathbf{U}^{-1} = \alpha \mathcal{U} \left( \mathbf{B}^T \sum_{m=0}^{\infty} \mathbf{G}_m \right). \quad (10.102)$$

*Proof.* Since  $\mathbf{G}_m$  is block upper triangular for all  $m \geq 0$ , so is the series  $\sum_{m=0}^{\infty} \mathbf{G}_m$ . It thus suffices to check that

$$\left( \mathbf{B}^T \sum_{m=0}^{\infty} \mathbf{G}_m \right) (i, j) = \left( \mathbf{B}[j]^T \sum_{m=0}^{\infty} \mathbf{G}_m[j] \right) (i, j) \quad i \leq j.$$

This shows the desired relation (10.102).  $\square$

$\boxed{4}$  *Recursion.* In order to estimate  $\mathbf{G}_m$ , it is useful to relate  $\mathbf{G}_m$  with  $\mathbf{G}_{m-1}$ . We start with a simple property of the operator  $\mathcal{U}$ : for  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$  and  $1 \leq i \leq j \leq n$ , there holds

$$(\mathbf{A} \mathcal{U}(\mathbf{B}))_{ij} = \sum_{k=1}^n A_{ik} \mathcal{U}(\mathbf{B})_{kj} = \sum_{k=1}^j A_{ik} \mathbf{B}_{kj} = (\mathbf{A}[j] \mathbf{B}[j])(i, j).$$

**Lemma 10.51 (recursion)** *The following is valid for all  $m \geq 1$*

$$\mathbf{G}_m = \mathbf{G}_{m-1} - \mathcal{U} \left( \mathbf{A} \mathcal{U}(\mathbf{A}^T \mathbf{G}_{m-1}) \right), \quad (10.103)$$

with  $\mathbf{G}_0 = \mathbf{I}$ . Therefore, the  $j$ -th column of  $\mathbf{G}_m$  reads

$$\mathbf{G}_m(0 : j, j) = \mathbf{G}[j] \mathbf{G}_{m-1}(0 : j, j) \quad 0 \leq j \leq N. \quad (10.104)$$

*Proof.* Take first  $m = 1$  and apply the proceeding relation for  $0 \leq i \leq j \leq N$  to obtain

$$\begin{aligned} (\mathbf{G}_0 - \mathcal{U}(\mathbf{A} \mathcal{U}(\mathbf{A}^T \mathbf{G}_0)))(i, j) &= \mathbf{I}(i, j) - (\mathbf{A} \mathcal{U}(\mathbf{A}^T))(i, j) \\ &= \mathbf{I}(i, j) - \mathbf{A}[j] \mathbf{A}[j]^T(i, j) \\ &= \mathbf{G}[j](i, j) = \mathbf{G}_1(i, j), \end{aligned}$$

in light of (10.100). Take next  $m > 1$  and  $0 \leq i \leq j$  to arrive at

$$\begin{aligned}
 & (\mathbf{G}_{m-1} - \mathcal{U}(\mathbf{A} \mathcal{U}(\mathbf{A}^T \mathbf{G}_{m-1}))) (i, j) \\
 &= \mathbf{G}_{m-1}(i, j) - \sum_{k, \ell=1}^j \mathbf{A}(i, k) \mathbf{A}^T(k, \ell) \mathbf{G}_{m-1}(\ell, j) \\
 &= \sum_{\ell=1}^j (\mathbf{I}[j] - \mathbf{A}[j] \mathbf{A}[j]^T) (i, \ell) \mathbf{G}_{m-1}(\ell, j) \\
 &= \sum_{\ell=1}^j \mathbf{G}[j](i, \ell) \mathbf{G}[j]^{m-1}(\ell, j) \\
 &= \mathbf{G}[j]^m(i, j) = \mathbf{G}_m(i, j).
 \end{aligned}$$

This is the asserted equality (10.103). The remaining relation (10.104) follows from the last equality upon realizing that

$$\mathbf{G}_m(0 : j, j) = (\mathbf{G}_m(i, j))_{i=0}^j = \mathbf{G}[j] \mathbf{G}[j]^{m-1}(0 : j, j) = \mathbf{G}[j] \mathbf{G}_{m-1}(0 : j, j).$$

This completes the proof.  $\square$

**[5] Schatten norms.** In the view Lemmas 10.51 (recursion) and 10.50 (representation of  $\mathbf{U}^{-1}$ ), we intend to estimate  $\|\mathbf{U}^{-1}\|_2$  in terms of suitable norms of  $\mathbf{G}_m$  that depend on the number  $N$  of blocks rather than the dimension  $n_N$ , because  $n_N \gg N$ . These special norms are called *block Schatten norms*.

However, for the sake of clarity, we start with the definition and properties of the usual Schatten norms. They include the operator 2-norm, the Frobenius norm, and satisfy a Hölder inequality.

**Definition 10.52 (Schatten norms)** Given  $\mathbf{M} \in \mathbb{R}^{n \times n}$  let

$$\sigma_1(\mathbf{M}) \geq \sigma_2(\mathbf{M}) \geq \cdots \geq \sigma_n(\mathbf{M}) \geq 0$$

be the singular values of  $\mathbf{M}$ . Given  $1 \leq p \leq \infty$ , let the  $p$ -Schatten norm be

$$|\mathbf{M}|_p := \left( \sum_{m=1}^n \sigma_m(\mathbf{M})^p \right)^{1/p}.$$

**Remark 10.53** Note that if  $p = \infty$  the Schatten norm reduces to the 2-norm, i.e.

$$|\mathbf{M}|_\infty = \sigma_1(\mathbf{M}) = \|\mathbf{M}\|_2,$$

and if  $p = 2$  it is equivalent to the Frobenius norm,

$$|\mathbf{M}|_2 = \left( \sum_{m=1}^n \sigma_m(\mathbf{M})^2 \right)^{1/2} = \left( \sum_{m=1}^n \mathbf{M}_{ij}^2 \right)^{1/2} = \|\mathbf{M}\|_F.$$

We now list a number of useful properties of these norms.

**Lemma 10.54 (properties of  $|\cdot|_p$ )** *The following properties hold for  $1 \leq p \leq \infty$ :*

- (i)  $\sigma_i(\mathbf{M}^T \mathbf{M}) = \sigma_i(\mathbf{M})^2 \quad \Rightarrow \quad |\mathbf{M}^T \mathbf{M}|_p = |\mathbf{M}|_{2p}^2$
- (ii)  $\sigma_i(\mathbf{M}) = \sigma_i(\mathbf{M}^T) \quad \Rightarrow \quad |\mathbf{M}|_p = |\mathbf{M}^T|_p$
- (iii) *Hölder inequality:* for  $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$ , with  $r, p, q \in [1, \infty]$ ,
$$|\mathbf{M}_1 \mathbf{M}_2|_r \leq |\mathbf{M}_1|_p |\mathbf{M}_2|_q$$
- (iv)  $|\mathcal{U}(\mathbf{M})|_\infty \lesssim \log(n) |\mathbf{M}|_\infty$
- (v)  $|\mathcal{U}(\mathbf{M})|_{2^j} \leq 2^{j-1} |\mathbf{M}|_{2^j}$ .

**Remark 10.55** Properties (i) and (ii) are trivial. We refer to [Dunford and Schwartz 1988, Lemma XI.9.20] for property (iii), to [Bhatia 2000, Eq (15)] for property (iv), and to [Davies 1988] and [Feischl 2022, Lemma 17] for property (v).

To define the block Schatten norms, we consider the subspace  $\mathcal{D}_b$  of  $\mathbb{R}^{n_N \times (N+1)}$  of matrices of the form

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_0 & & & \\ & \mathbf{X}_1 & & \\ & & \ddots & \\ & & & \mathbf{X}_N \end{bmatrix}, \quad \mathbf{X}_j \in \mathbb{R}^{d_j} \quad 0 \leq j \leq N,$$

or equivalently

$$\mathbf{X} \in \mathcal{D}_b \quad \Longleftrightarrow \quad \mathbf{X}_{ij} = \mathbf{0} \quad \forall i \neq n_{j-1} + 1, \dots, n_j.$$

We can represent  $\mathbf{X}$  using block notation as follows:

$$\mathbf{X} = (\mathbf{X}(i, j))_{i,j=0}^N, \quad \mathbf{X}(i, j) \in \mathbb{R}^{d_j \times 1},$$

where

$$\mathbf{X}(i, j) = \begin{cases} \mathbf{X}_j & i = j \\ \mathbf{0} & i \neq j. \end{cases}$$

Given a block matrix  $\mathbf{M} = (\mathbf{M}(i, j))_{i,j=0}^N \in \mathbb{R}^{n_N \times n}$ , we consider

$$\mathbf{M}\mathbf{X} = (\mathbf{M}(i, j)\mathbf{X}_j)_{i,j=0}^N \in \mathbb{R}^{n_N \times (N+1)},$$

namely the  $j$ -th block column of  $\mathbf{M}\mathbf{X}$  is

$$(\mathbf{M}(i, j)\mathbf{X}_j)_{i=0}^N \in \mathbb{R}^{n_N}.$$

**Definition 10.56 (block Schatten norms)** For  $1 \leq p \leq \infty$ , let

$$|\mathbf{M}|_{b,p} := \sup_{\mathbf{X} \in \mathcal{D}_b, |\mathbf{X}|_\infty \leq 1} |\mathbf{M}\mathbf{X}|_p \quad \forall \mathbf{M} \in \mathbb{R}^{n_N \times n_N}.$$



Note the unusual norm  $|X|_\infty$  instead of  $|X|_p$  in this definition of operator norm  $|M|_{b,p}$ . This choice is deliberate and will be useful later; see Remark 10.58. We now list important properties of the block Schatten norms; see [Feischl 2022, Lemmas 15, 16, 17] for proofs.

**Lemma 10.57 (properties of  $|\cdot|_{b,p}$ )** *The following properties hold for all  $M, M_1, M_2 \in \mathbb{R}^{n_N \times n_N}$  and  $1 \leq p \leq \infty$ :*

- (i)  $|M|_{b,p} \leq (N+1)^{1/p} |M|_\infty = (N+1)^{1/p} \|M\|_2$
- (ii)  $|M_1 M_2|_{b,p} \leq |M_1|_\infty |M_2|_{b,p}$
- (iii)  $|M|_\infty \leq |M|_{b,p}$ ,  $|M|_{b,\infty} = |M|_\infty$
- (iv) If  $M_1 \in \mathbb{R}^{n_N \times n_N}$  is block triangular with  $j$ -th block column

$$M_1(0 : j, j) = P_j M_2(0 : j, j), \quad P_j \in \mathbb{R}^{n_j \times n_j}$$

for  $0 \leq j \leq N$ , then

$$|M_1|_{b,2} \leq \max_{0 \leq j \leq N} |P_j|_\infty |M_2|_{b,2}$$

- (v)  $|\mathcal{U}(M)|_{b,2^k} \leq 2^{k-1} |M|_{b,2^k} \quad k = 1, 2$
- (vi)  $|\mathcal{U}(M)|_\infty \leq (\lceil \log_2(N) \rceil + 1) |M|_\infty$ .

**Remark 10.58** To understand the significance of Definition 10.56, we examine the growth of the usual and block  $p$ -Schatten norm relative to the  $\infty$ -Schatten norm for  $1 \leq p < \infty$ . Given  $M \in \mathbb{R}^{n_N \times n_N}$ , we have for the usual  $p$ -norm

$$|M|_p = \left( \sum_{i=1}^{n_N} \sigma_i(M)^p \right)^{1/p} \leq n_N^{1/p} \sigma_1(M) = n_N^{1/p} |M|_\infty = n_N^{1/p} \|M\|_2,$$

whereas for the block  $p$ -norm we get

$$|M|_{b,p} \leq (N+1)^{1/p} |M|_\infty = (N+1)^{1/p} \|M\|_2,$$

according to Lemma 10.57 (i). In fact, given  $X \in \mathcal{D}_b$  with  $|X|_\infty = \|X\|_2 = 1$ , we first note that

$$|MX|_p = \left( \sum_{j=0}^N \sigma_j(MX)^p \right)^{\frac{1}{p}} \leq \sigma_0(MX)(N+1)^{\frac{1}{p}} = \|MX\|_2(N+1)^{\frac{1}{p}},$$

and also that

$$\|MX\|_2 = \sup_{x \in \mathbb{R}^{N+1}} \frac{\|MXx\|_2}{\|x\|_2} \leq \|M\|_2 \sup_{x \in \mathbb{R}^{N+1}} \frac{\|Xx\|_2}{\|x\|_2} \leq \|M\|_2$$

because  $\|M\|_2 = |M|_\infty = 1$ . On the one hand, this explains why it is convenient to have the norm  $|X|_\infty$  rather than  $|X|_p$  in Definition 10.56. On the other hand, this calculation reveals the key point that

$$|M|_{b,p} \ll |M|_p$$

because the growth of  $|\mathbf{M}|_{b,p}$  is dictated by the number of blocks  $N + 1$  whereas that of  $|\mathbf{M}|_p$  is proportional to the dimension  $n_N$  of  $\mathbf{M}$  and  $n_N \gg N$ . This property is essential in the estimate of  $\|\mathbf{U}^{-1}\|_2$  below.

□ *Estimate of  $\|\mathbf{U}^{-1}\|_2$ .* We are now in a position to prove the desired bound (10.94).

**Proposition 10.59 (estimate of  $\|\mathbf{U}^{-1}\|_2$ )** *Let  $\mathbf{B} \in \mathbb{R}^{n_N \times n_N}$  be a block matrix such that*

$$\|\mathbf{B}\|_2 \leq \|\mathcal{B}\|, \quad \max_{0 \leq j \leq N} \|\mathbf{B}[j]^{-1}\|_2 \leq \frac{1}{\beta}.$$

*Then there exist constants  $C_{LU}$  and  $p > 2$  such that the block upper triangular factor  $\mathbf{U}$  of  $\mathbf{B}$  satisfies*

$$\|\mathbf{U}^{-1}\|_2 \leq C_{LU} N^{1/p}. \quad (10.105)$$

*Proof.* We recall (10.102) of Lemma 10.50 (representation of  $\mathbf{U}^{-1}$ )

$$\mathbf{U}^{-1} = \alpha \mathcal{U} \left( \mathbf{B}^T \sum_{m=0}^{\infty} \mathbf{G}_m \right),$$

along with (10.103) of Lemma 10.51 (recursion)

$$\mathbf{G}_m = \mathbf{G}_{m-1} - \mathcal{U} \left( \mathbf{A} \mathcal{U}(\mathbf{A}^T \mathbf{G}_{m-1}) \right) \quad m \geq 1$$

and (10.104) of Lemma 10.51

$$\mathbf{G}(0 : j, j) = \mathbf{G}[j] \mathbf{G}_{m-1}(0 : j, j) \quad 0 \leq j \leq N,$$

with  $\mathbf{G}_0 = \mathbf{I}$ . We use these expressions in conjunction with Lemma 10.57 (properties of  $|\cdot|_{b,p}$ ) to prove (10.105). We proceed in several steps.

(i) *Bound for  $|\mathbf{G}_m|_{b,2}$ .* In light of (10.99) and (10.100)

$$|\mathbf{G}[j]|_{\infty} = \|\mathbf{G}[j]\|_2 \leq \rho = \sqrt{1 - \frac{\beta^4}{\|\mathcal{B}\|^4}} < 1 \quad 0 \leq j \leq N.$$

Applying Lemma 10.57 (iv) to  $\mathbf{G}_m$  yields

$$\begin{aligned} |\mathbf{G}_m|_{b,2} &\leq \max_{0 \leq j \leq N} |\mathbf{G}[j]|_{\infty} |\mathbf{G}_{m-1}|_{b,2} \\ &\leq \rho |\mathbf{G}_{m-1}|_{b,2} \leq \rho^m |\mathbf{I}|_{b,2}. \end{aligned}$$

Recalling Lemma 10.57 (i)

$$|\mathbf{I}|_{b,2} \leq (N+1)^{1/2} \|\mathbf{I}\|_2 = (N+1)^{1/2},$$

whence

$$|\mathbf{G}_m|_{b,2} \leq \rho^m (N+1)^{1/2}.$$

We observe that this bound is not good enough for our purposes because it scales like  $N^{1/2}$  instead of  $N^{1/p}$  for  $p > 2$ . We next improve upon this bound.

- (ii) *Bound for  $|\mathbf{G}_m|_{b,4}$ .* We take  $k = 2$  in Lemma 10.57 (v) and use the triangle inequality to arrive at

$$\begin{aligned} |\mathbf{G}_m|_{b,4} &\leq |\mathbf{G}_{m-1}|_{b,4} + |\mathcal{U}(\mathbf{A} \mathcal{U}(\mathbf{A}^T \mathbf{G}_{m-1}))|_{b,4} \\ &\leq |\mathbf{G}_{m-1}|_{b,4} + 2|\mathbf{A} \mathcal{U}(\mathbf{A}^T \mathbf{G}_{m-1})|_{b,4}. \end{aligned}$$

We further apply Lemma 10.57 (ii) and (v) to obtain

$$|\mathbf{A} \mathcal{U}(\mathbf{A}^T \mathbf{G}_{m-1})|_{b,4} \leq 2|\mathbf{A}|_\infty |\mathbf{A}^T \mathbf{G}_{m-1}|_{b,4} \leq 2|\mathbf{A}|_\infty^2 |\mathbf{G}_{m-1}|_{b,4}.$$

Therefore,

$$|\mathbf{G}_m|_{b,4} \leq (1 + 4|\mathbf{A}|_\infty^2) |\mathbf{G}_{m-1}|_{b,4}$$

but the prefactor on the right-hand side is greater than 1 and so not suitable for iteration. We still have

$$|\mathbf{G}_m|_{b,4} \leq (1 + 4|\mathbf{A}|_\infty^2)^m |\mathbf{I}|_{b,4}.$$

- (iii) *Bound for  $|\mathbf{G}_m|_\infty$ .* We combine the estimates from steps 1 and 2 to exploit their relative merits. Recall from Lemma 10.57 (iii) that

$$|\mathbf{G}_m|_\infty \leq |\mathbf{G}_m|_{b,p} \quad \forall 1 \leq p \leq \infty.$$

Take  $p = 2, 4$  and  $0 < t < 1$  to be chosen later, and write

$$\begin{aligned} |\mathbf{G}_m|_\infty &\leq |\mathbf{G}_m|_{b,2}^{1-t} |\mathbf{G}_m|_{b,4}^t \\ &\leq \left[ \rho^{1-t} (1 + 4|\mathbf{A}|_\infty^2)^t \right]^m |\mathbf{I}|_{b,2}^{1-t} |\mathbf{I}|_{b,4}^t. \end{aligned}$$

Consequently, there exists  $0 < t_0 < 1$  such that

$$q := \rho^{1-t} (1 + 4|\mathbf{A}|_\infty^2)^t < 1 \quad 0 < t < t_0$$

and

$$|\mathbf{G}_m|_\infty \leq q^m |\mathbf{I}|_{b,2}^{1-t} |\mathbf{I}|_{b,4}^t.$$

We now estimate the two terms on the right-hand side relying on Lemma 10.57 (i), namely

$$\begin{aligned} |\mathbf{I}|_{b,2} &\leq (N+1)^{1/2} \|\mathbf{I}\|_2 = (N+1)^{1/2}, \\ |\mathbf{I}|_{b,4} &\leq (N+1)^{1/4} \|\mathbf{I}\|_2 = (N+1)^{1/4}. \end{aligned}$$

We thus obtain

$$|\mathbf{G}_m|_\infty \leq q^m (N+1)^{1/\tilde{p}}$$

with  $\frac{1}{\tilde{p}} = \frac{1-t}{2} + \frac{t}{4} < \frac{1}{2}$  for  $0 < t < t_0$ .

- (iv) *Estimate of  $\|\mathbf{U}^{-1}\|_2$ .* Recalling the expression

$$\mathbf{U}^{-1} = \alpha \mathcal{U} \left( \mathbf{B}^T \sum_{m=0}^{\infty} \mathbf{G}_m \right),$$

and applying Lemma 10.57 (vi), (ii) and (iii), we see that

$$\begin{aligned}\|U^{-1}\|_2 &= |U^{-1}|_\infty \lesssim \log(N) |B|_\infty \sum_{m=0}^{\infty} |G_m|_\infty \\ &\lesssim |B|_\infty (N+1)^{1/\tilde{p}} \log(N) \sum_{m=0}^{\infty} q^m \\ &\lesssim \|B\|_2 (N+1)^{1/\tilde{p}} \log(N).\end{aligned}$$

Finally, for any  $2 < p < \tilde{p}$ , we can absorb the logarithm thereby getting

$$\|U^{-1}\|_2 \lesssim \|B\|_2 (N+1)^{1/p},$$

which is the desired estimate (10.105).

This concludes the proof.  $\square$

$\square$  *Estimate of block diagonal  $D$ .* We recall that  $D = \text{diag } U \in \mathbb{R}^{n_N \times n_N}$  is the block diagonal of  $U$ . We consider the block partitioning of  $B[j]$

$$B[j] = \begin{bmatrix} B[j-1] & R_1 \\ R_2^T & R_3 \end{bmatrix} \in \mathbb{R}^{n_j \times n_j}$$

where

$$\begin{aligned}R_1 &= B[j](1 : j-1, j) \in \mathbb{R}^{n_{j-1} \times d_j}, \\ R_2^T &= B[j](j, 1 : j-1) \in \mathbb{R}^{d_j \times n_{j-1}}, \\ R_3 &= B[j](j, j) \in \mathbb{R}^{d_j \times d_j}.\end{aligned}$$

**Lemma 10.60 (bound of  $\|D\|_2$ )** *There holds*

$$\|D\|_2 \leq \|B\| + \frac{\|B\|^2}{\beta} = C_D. \quad (10.106)$$

*Proof.* Compute the  $LU$  factorization of  $B[j]$

$$B[j] = \begin{bmatrix} I[j-1] & 0 \\ R_2^T B[j-1]^{-1} & 1 \end{bmatrix} \begin{bmatrix} B[j-1] & R_1 \\ 0 & R_3 - R_2^T B[j-1]^{-1} R_1 \end{bmatrix}$$

and realize that

$$U(j, j) = D(j, j) = R_3 - R_2^T B[j-1]^{-1} R_1 \in \mathbb{R}^{d_j \times d_j}.$$

Since

$$|R_i|_\infty = \|R_i\|_2 \leq \|B[j]\|_2 \leq \|B\|_2 = \|B\| \quad i = 1, 2$$

$$|R_3|_\infty = \|R_3\|_2 \leq \|B[j]\|_2 \leq \|B\|,$$

and

$$|\mathbf{B}[j-1]^{-1}|_{\infty} = \|\mathbf{B}[j-1]\|_2 \leq \frac{1}{\beta},$$

according to properties (P1) and (P2) of the bilinear form  $\mathcal{B}$ , we deduce

$$|\mathbf{D}(j, j)|_{\infty} = \|\mathbf{D}(j, j)\|_2 \leq \|\mathcal{B}\| + \frac{\|\mathcal{B}\|^2}{\beta}$$

as asserted.  $\square$

$\square$  *Bound of  $\mathbf{LU}$  factors.* We are finally in the position to prove Theorem 10.15. We combine Proposition 10.59 (estimate of  $\|\mathbf{U}^{-1}\|_2$ ) and  $\mathbf{L} = \mathbf{B}\mathbf{U}^{-1}$  to obtain

$$\|\mathbf{L}\|_2 \leq \|\mathbf{B}\|_2 \|\mathbf{U}^{-1}\|_2 \leq \|\mathcal{B}\| C_{LU} N^{1/p}.$$

Then, invoking (10.97) in conjunction with Proposition 10.59 and Lemma 10.60 (bound of  $\|\mathbf{G}\|_2$ ) as well as the bounds of  $\|\mathbf{U}^{-1}\|_2$  and  $\|\mathbf{L}\|_2$ , yields

$$\begin{aligned} \|\mathbf{U}\|_2 &\leq \|\mathbf{D}\|_2 \|\tilde{\mathbf{L}}^T\|_2 \leq C_D \|\mathcal{B}\| C_{LU} N^{1/p} \\ \|\mathbf{L}^{-1}\|_2 &\leq \|\mathbf{D}\|_2 \|\tilde{\mathbf{U}}^{-T}\|_2 \leq C_D C_{LU} N^{1/p} \end{aligned} \tag{10.107}$$

with  $C_D$  being the constant in (10.106). This completes the proof.  $\square$

## Index

### Algorithms

- DATA: procedure to approximate the data  $\mathcal{D} = (A, c, f)$ , 123, 214
- AFEM-SW: AFEM with switch, 139
- AFEM-TS: two step AFEM successively approximating the data and the Galerkin solution with approximate data, 123, 154
- AFEM-DG-TS: interior penalty version of AFEM-TS, 233
- CONSTRAINT-A : modify an approximation of  $A$  to satisfy the structural constraints, 201
- CONSTRAINT-c : modify an approximation of  $A$  to satisfy the structural constraints, 204
- ESTIMATE : compute the element error indicators and element data oscillations, 128
- GALERKIN: procedure that iterates SOLVE, ESTIMATE , MARK, REFINE , 123, 129
- GALERKIN-DG: discontinuous Galerkin version of GALERKIN, 253
- GREEDY : abstract greedy algorithm for DATA, 191
- MARK: Dörfler marking, 128
- REFINE : refine all marked elements  $b$  times and others necessary to produce a conforming mesh, 129, 226
- REFINE : refine marked elements and others necessary to produce a  $\Lambda$ -admissible mesh, 230
- SOLVE: construct FEM approximation, 127

### Assumptions

- Restrictions on  $\kappa$ ,  $\omega$ , and  $\theta$ , 260
- Abstract cut-off, 74
- Admissible set of parameters for GREEDY , 192
- Approximability of  $u$ , 164
- Approximability of data, 167
- Cardinality of the marked set, 171
- Discrete coefficients and discrete functionals, 88
- Equivalence of error and estimator, 264
- Equivalence of error and full estimator, 282
- Estimator reduction, 264
- Initial labeling, 173
- Lipschitz continuity of estimator, 278
- Marking parameter  $\theta$ , 169
- Monotonicity of estimator, 279
- Monotonicity of local oscillations, 193
- Properties of DATA, 148
- Quasi-monotonicity of oscillation, 279
- Quasi-optimality of DATA, 168

Relaxed quasi-orthogonality, 264  
 Restriction on  $\omega$ , 170  
 Size of  $\omega$ , 174  
 Structural assumption for discrete data, 144  
 Structural assumption for exact data, 144  
 Structure of  $f$ , 214

## Constants

$(C_L, C_U)$ : a-posteriori lower and upper bounds constants, 89, 125, 169  
 $(C_L^{\text{eq}}, C_U^{\text{eq}})$ : lower and upper estimators equivalence constant, 159  
 $(\alpha_1, \alpha_2)$ : lower and upper bounds of the diffusion coefficient spectrum, 14, 144  
 $(\hat{\alpha}_1, \hat{\alpha}_2)$ : lower and upper bounds of the approximate diffusion coefficient spectrum, 144  
 $(\hat{\alpha}_1, \hat{\alpha}_2)$ : lower and upper bounds on the spectrum of  $\hat{\mathbf{A}}$ , 203  
 $(c_1, c_2)$ : lower and upper bounds of the reaction coefficient, 144  
 $(\hat{c}_1, \hat{c}_2)$ : lower and upper bounds of the approximate reaction coefficient, 144  
 $(c_{\mathcal{B}}, C_{\mathcal{B}})$ : norm equivalence constants, 18, 124  
 $(c_{\hat{\mathcal{B}}}, C_{\hat{\mathcal{B}}})$ : norm equivalence constants for the perturbed problem, 145  
 $C_D$ : DATA constant, 148  
 $C_P$ : Poincaré constant, 13  
 $C_{\text{loc}}$ : localization constant, 57  
 $C_{\text{ovrl}}$ : overlay constant, 57  
 $C_{\text{cea}}$ : best approximation constant, 25  
 $C_{\text{osc}}$ : oscillation quasi-monotonicity constant, 100  
 $C_{\text{Lip}}$ : estimator Lipschitz property constant, 99, 127  
 $L$ : threshold parameter for constrained approximation, 197  
 $C_{\text{BA}}$ : best approximation constant of  $\Pi_T$ , 149  
 $C_{\text{BA}}$ : quasi-monotonicity constant of  $\Pi_T$ , 149  
 $D$ : complexity of REFINE constant, 43, 53, 216  
 $\tilde{D}$ : modified complexity of REFINE constant, 173  
 $C_{\text{ctr}}$ : constrain upper bound amplification constant, 144  
 $C_{\text{data}}$ : DATA approximation constant, 148, 152  
 $C_{\text{Lip}}$ : estimator Lipschitz property constant, 98  
 $\Lambda$ :  $\Lambda$ -admissibility constant, 49  
 $\Lambda_{\text{data}}$ : DATA quasi-optimality constant, 148, 165  
 $\alpha$ : inf sup constant, 17  
 $\alpha$ : contraction constant, 140  
 $\alpha$ : contraction constant, 130  
 $\sigma$ : shape regularity constant of  $\mathbb{T}$ , 26  
 $\theta$ : Dörfler marking parameter, 128  
 $\tilde{C}_U$ : localized upper bound constant, 93, 169

## Definitions

$\varepsilon$ -approximation of order  $s$ , 168  
 $\mathcal{T}$ -meshed subdomain, 69  
 faced-connected, 31  
 Global index of a node, 48  
 Interior vertex property, 96  
 Sobolev number  $\text{sob}(W_p^k)$ , 11

## Error Estimators

$E_{\mathcal{T}}(v)_q$ : generic total error, 191  
 $R_{\mathcal{T}}$ : residual in  $H^{-1}(\Omega)$ , 55  
 $\mathcal{E}_{\mathcal{T}}^{\text{abs}}(z)$ : abstract total estimator, 83  
 $\mathcal{E}_{\mathcal{T}}$ : total estimator, 89  
 $\mathcal{E}_{\mathcal{T}}(T), \mathcal{E}_{\mathcal{T}}(u_{\mathcal{T}}, f, T)$ : local total estimator, 89  
 $\mathcal{E}_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, f, T)$ : standard local indicators, 68  
 $\mathcal{E}_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, \mathcal{D})$ : standard residual estimator, 61  
 $\mathcal{E}_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, \mathcal{D}, T)$ : standard local indicators, 61  
 $\eta_{\mathcal{T}}^{\text{abs}}(z)$ : abstract PDE estimator, 83  
 $\eta_{\mathcal{T}}(u_{\mathcal{T}}, T)$ : PDE local estimator, 89  
 $\eta_{\mathcal{T}}(u_{\mathcal{T}}, f)$ : PDE estimator, 91  
 $\eta_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, T)$ : standard local PDE indicators, 68  
 $\text{osc}_{\mathcal{T}}^{\text{abs}}(R_{\mathcal{T}}, z)$ : abstract oscillation, 83  
 $\text{osc}_{\mathcal{T}}(\mathbf{A}, T)_r$ : local surrogate for the diffusion coefficient approximation error, 150  
 $\text{osc}_{\mathcal{T}}(\mathcal{D})$ : surrogate for the data error, 147  
 $\text{osc}_{\mathcal{T}}(c, T)_2$ : super-convergent local surrogate for the reaction coefficient approximation error, 150  
 $\text{osc}_{\mathcal{T}}(c, T)_{\infty}$ : super-convergent local surrogate for the reaction coefficient approximation error, 150  
 $\text{osc}_{\mathcal{T}}(c, T)_q$ : local surrogate for the reaction coefficient approximation error, 150  
 $\text{osc}_{\mathcal{T}}(f), \text{osc}_{\mathcal{T}}(f)_{-1}$ : oscillation for the load function, 91  
 $\text{osc}_{\mathcal{T}}(f, T), \text{osc}_{\mathcal{T}}(f, T)_{-1}$ : local oscillation for the load function, 89, 91, 151  
 $\text{osc}_{\mathcal{T}}(v, T)_p$ : generic surrogate for data error, 150  
 $\text{osc}_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, \mathcal{D})$ : standard oscillation, 64  
 $\text{osc}_{\mathcal{T}}^{\text{std}}(u_{\mathcal{T}}, \mathcal{D}, T)$ : standard local oscillation, 64  
 $\text{osc}_{\mathcal{T}}(\mathcal{D})$ : total data error estimator, 151  
 $\text{osc}_{\mathcal{T}}(\mathbf{A})_r$ : oscillation for the diffusion coefficient, 151  
 $\text{osc}_{\mathcal{T}}(c)_q$ : oscillation for the reaction coefficient, 151  
 $\text{osc}_{\mathcal{T}}(f), \text{osc}_{\mathcal{T}}(f)_{-1}$ : oscillation for the load function, 151  
 $\tilde{E}_{\mathcal{T}}(f)_{-1}^2$ : generic surrogate estimator for the approximation of the load term, 207  
 $j(u_{\mathcal{T}})$ : jump residual, 62



$r_{\mathcal{T}}(u_{\mathcal{T}}), r(u_{\mathcal{T}})$ : element residual, 62

### Functional Spaces

$B_{p,q}^s(\Omega)$ : Besov spaces, 182

$D(\Omega)$ : metric space for the data perturbation, 147

$\widehat{D}(\Omega)$ : temporary metric space for the data perturbation, 145

$M(\alpha_1, \alpha_2)$ : Admissible set for  $\mathbf{A}$ , 144

$R(c_1, c_2)$ : Admissible set for  $c$ , 144

$W_p^k(\Omega)$ : Sobolev spaces, 11

$W_q^{-s}(\Omega)$ : dual of  $W_{q^*}^s(\Omega)$  with  $q^* = \frac{q}{q-1}$ , 145

$X_p^s(\Omega)$ : abstract functional spaces, 180

$\mathbb{M}_s$ : approximation classes of  $\mathbf{A}$ , 165

$\mathbb{C}_s$ : approximation classes of  $c$ , 166

$\mathbb{F}_s$ : approximation classes of  $f$ , 166

$\mathbb{D}$ : data, 122

$\mathbb{D}_{\widehat{\mathcal{T}}}$ : discrete data subordinate to  $\widehat{\mathcal{T}}$ , 122

$\mathbb{F}(\mathcal{T}_\omega)$ : local discrete functionals, 69

$\mathbb{F}_{\mathcal{T}}, \mathbb{F}(\mathcal{T})$ : discrete functionals, 69

$\mathbb{V}^+(\mathcal{T}), \mathbb{V}_{\mathcal{T}}^+$ : test space for discrete functionals, 75

$\mathbb{V}^+(\mathcal{T}_\omega)$ : local test space for discrete functionals, 75

$\mathbb{V}_{\mathcal{T}}$ : conforming finite element space, 26

$\mathbb{A}_s$ : approximation class for  $u$ , 162

$\mathbb{A}_s^{-1}$ : approximation classes for  $v$  for the discontinuous Galerkin norm, 239

$\mathbb{E}_{\mathcal{T}}$ : broken  $H^1$  space, 235

$\mathbb{V}_{\mathcal{T}}^{-1}$ : non-conforming finite element space, 234

$\text{Lip}_p^{n+1}(\Omega)$ : Lipschitz spaces, 180

$\mathbb{S}_{\mathcal{T}}^{n,-1}$ : piecewise polynomials of degree  $\leq n$ , 26

$\mathbb{S}_{\mathcal{T}}^{n,0}$ : globally continuous piecewise polynomials of degree  $\leq n$ , 26

### Functions

$\phi_z$ : Lagrange basis of  $\mathbb{S}_{\mathcal{T}}^{1,0}$ , 26

$\psi_z$ : Lagrange basis of  $\mathbb{S}_{\mathcal{T}}^{n,0}$ , 27

$\widehat{u}$ : solution to the perturbed problem (5.5), 124

$u$ : solution to weak formulation (2.7), 15

$u_{\mathcal{T}}$ : Galerkin approximation, 54

### Meshes

$T_d$ : reference element, 25

$\mathcal{P}$ : proper nodes, 48, 51

$\mathcal{F}, \mathcal{F}_{\mathcal{T}}$ : interior faces, 55

$\mathcal{F}_\omega$ : faces interior to  $\omega$ , 69

$\mathcal{F}_z, \mathcal{F}_{\omega_z}$ : faces interior to  $\omega_z$ , 69

$\gamma_z$ : skeleton of  $\omega_z$ , 26

- $\mathcal{T} \leq \mathcal{T}_*$ : refinement relation, 124
- $\mathcal{T}_1 \oplus \mathcal{T}_2$ : mesh overlay, 43
- $\mathcal{T}_\omega$ : triangulated submesh, 69
- $\mathcal{T}_z, \mathcal{T}_{\omega_z}$ : elements forming  $\omega_z$ , 69
- $\mathcal{T}_z$ : star of elements sharing the vertex  $z$ , 57
- $\mathbb{T}$ : set of all conforming refinements of  $\mathcal{T}_0$ , 38
- $\mathbb{T}^\Lambda$ : set of all  $\Lambda$ -admissibility refinements of  $\mathcal{T}_0$ , 49
- $\mathbb{T}_N$ : set of all conforming refinement of  $\mathcal{T}_0$  with no more than  $N$  elements, 162
- $[[\cdot]]$ : jump across faces, 234
- $[[\mathbf{g}]] \cdot \mathbf{n}_F$ : normal jump across  $F$ , 62
- $\lambda(x)$ : global index of a node  $x \in \mathcal{N}$ , 48
- $\{\{.\}\}$ : average on faces, 234
- $\mathcal{N}$ : Lagrange nodes of order  $n$ , 27
- $\omega_F$ : region of elements containing the face  $F$ , 59
- $\omega_T, \omega_{\mathcal{T}}(T)$ : region of elements intersecting  $T$ , 28, 60
- $\widetilde{\omega}_T, \widetilde{\omega}_{\mathcal{T}}(T)$ : elements sharing a face with  $T$ , 28
- $\omega_{\mathcal{T}}(P)$ : domain of influence of a proper node  $P$ , 50
- $\omega_z$ : region made of elements sharing the vertex  $z$ , 26, 57
- $\mathbf{n}_F$ : normal to the face  $F$ , 62
- $\mathcal{V}$ : set of vertices, 26
- $g(T)$ : generation of  $T$ , 40

## Norms

- $\|v\|_{a,\mathcal{T}}$ : discontinuous Galerkin norm, 235
- $\|\cdot\|_\Omega$ : energy norm with exact coefficients, 124
- $\|\cdot\|_\Omega$ : energy norm with perturbed coefficients, 124

## Operators

- $I_{\mathcal{T}}$ : quasi-interpolation operator, 29
- $P_T, P_F$ : polynomial densities of  $P_{\mathcal{T}}$ , 75
- $P_{\mathcal{T}}$ : projection operator from  $H^{-1}(\Omega)$  into  $\mathbb{F}_{\mathcal{T}}$ , 75
- $\mathcal{I}_{\mathcal{T}}^{\text{dG}}$ : discontinuous Galerkin quasi-interpolant, 236
- $\Pi_K, \Pi_K^m$ :  $L^2$  projection onto  $\mathbb{P}_m(K)$ , 64, 148

## References

- R. A. Adams and J. J. F. Fournier (2003), *Sobolev spaces*, Vol. 140 of Pure and Applied Mathematics (Amsterdam), second edition, Elsevier/Academic Press, Amsterdam.
- M. Ainsworth (2010), A framework for obtaining guaranteed error bounds for finite element approximations, *J. Comput. Appl. Math.* **234**(9), 2618–2632.
- M. Ainsworth and J. T. Oden (2000), *A posteriori error estimation in finite element analysis*, Pure and Applied Mathematics (New York), Wiley-Interscience [John Wiley & Sons], New York.
- D. N. Arnold, F. Brezzi, B. Cockburn and L. D. Marini (2002), Unified analysis of discontinuous galerkin methods for elliptic problems, *SIAM journal on numerical analysis* **39**(5), 1749–1779.
- I. Babuška (1971), Error-bounds for finite element method, *Numerische Mathematik* **16**(4), 322–333.
- I. Babuška (1971), The finite element method for elliptic differential equations, in *Numerical Solution of Partial Differential Equations, II (SYNSPADE 1970) (Proc. Sympos., Univ. of Maryland, College Park, Md., 1970)*, Academic Press, New York-London, pp. 69–106.
- I. Babuška and A. K. Aziz (1972), Survey lectures on the mathematical foundations of the finite element method, in *The mathematical foundations of the finite element method with applications to partial differential equations (Proc. Sympos., Univ. Maryland, Baltimore, Md., 1972)*, Academic Press, New York-London, pp. 1–359. With the collaboration of G. Fix and R. B. Kellogg.
- I. Babuška and A. Miller (1987), A feedback finite element method with a posteriori error estimation. I. The finite element method and some basic properties of the a posteriori error estimator, *Comput. Methods Appl. Mech. Engrg.* **61**(1), 1–40.
- I. Babuška and W. C. Rheinboldt (1978), Error estimates for adaptive finite element computations, *SIAM J. Numer. Anal.* **15**(4), 736–754.
- I. Babuška, R. B. Kellogg and J. Pitkäranta (1979), Direct and inverse error estimates for finite elements with mesh refinements, *Numer. Math.* **33**(4), 447–471.
- E. Bänsch, P. Morin and R. H. Nochetto (2002), An adaptive Uzawa FEM for the Stokes problem: convergence without the inf-sup condition, *SIAM J. Numer. Anal.* **40**(4), 1207–1229.
- L. Beirão da Veiga, C. Canuto, R. H. Nochetto, G. Vacca and M. Verani (2023), Adaptive VEM: stabilization-free a posteriori error analysis and contraction property, *SIAM J. Numer. Anal.* **61**(2), 457–494.
- J. Bergh and J. Löfström (1976), *Interpolation spaces. An introduction*, Vol. No. 223 of Grundlehren der Mathematischen Wissenschaften, Springer-Verlag, Berlin-New York.
- C. Bernardi and V. Girault (1998), A local regularization operator for triangular and quadrilateral finite elements, *SIAM Journal on Numerical Analysis* **35**(5), 1893–1916.
- R. Bhatia (2000), Pinching, trimming, truncating, and averaging of matrices, *Amer. Math. Monthly* **107**(7), 602–608.
- P. Binev (2018), Tree approximation for hp-adaptivity, *SIAM J. Numer. Anal.* **56**(6), 3346–3357.
- P. Binev and R. DeVore (2004), Fast computation in adaptive tree approximation, *Numer. Math.* **97**(2), 193–217.
- P. Binev, W. Dahmen and R. DeVore (2004), Adaptive finite element methods with convergence rates, *Numer. Math.* **97**(2), 219–268.

- P. Binev, W. Dahmen, R. DeVore and P. Petrushev (2002), Approximation classes for adaptive methods, Vol. 28, pp. 391–416. Dedicated to the memory of Vassil Popov on the occasion of his 60th birthday.
- P. Binev, F. Fierro and A. Veiser (2023), Near-best adaptive approximation on conforming meshes, *Constr. Approx.* **57**(2), 327–349.
- J. Blechta, J. Málek and M. Vohralík (2020), Localization of the  $W_q^{-1}$ -norm for local a posteriori efficiency, *IMA Journal of Numerical Analysis* **40**(2), 914–950.
- D. Boffi, F. Brezzi and M. Fortin (2013), *Mixed finite element methods and applications*, Vol. 44 of Springer Series in Computational Mathematics, Springer, Heidelberg.
- A. Bonito and D. Devaud (2015), Adaptive finite element methods for the Stokes problem with discontinuous viscosity, *Mathematics of Computation* **84**(295), 2137–2162.
- A. Bonito and R. H. Nochetto (2010), Quasi-optimal convergence rate of an adaptive discontinuous galerkin method, *SIAM Journal on Numerical Analysis* **48**(2), 734–771.
- A. Bonito, J. M. Cascón, K. Mekchay, P. Morin and R. H. Nochetto (2016), High-order AFEM for the Laplace-Beltrami operator: convergence rates, *Found. Comput. Math.* **16**(6), 1473–1539.
- A. Bonito, J. M. Cascón, P. Morin and R. H. Nochetto (2013a), AFEM for geometric PDE: the Laplace-Beltrami operator, *Analysis and numerics of partial differential equations* pp. 257–306.
- A. Bonito, R. A. DeVore and R. H. Nochetto (2013b), Adaptive finite element methods for elliptic problems with discontinuous coefficients, *SIAM J. Numer. Anal.* **51**(6), 3106–3134.
- A. Bonito, R. H. Nochetto and D. Ntongkas (2021), Dg approach to large bending plate deformations with isometry constraint, *Mathematical Models and Methods in Applied Sciences* **31**(01), 133–175.
- F. Bornemann, B. Erdmann and R. Kornhuber (1996), A Posteriori Error Estimates for Elliptic Problems in Two and Three Space Dimensions, *SIAM Journal on Numerical Analysis* **33**(3), 1188–1204.
- D. Braess (2007), *Finite elements*, third edition, Cambridge University Press, Cambridge. Theory, fast solvers, and applications in elasticity theory, Translated from the German by Larry L. Schumaker.
- D. Braess, V. Pillwein and J. Schöberl (2009), Equilibrated residual error estimates are p-robust, *Computer Methods in Applied Mechanics and Engineering* **198**(13–14), 1189–1197.
- S. C. Brenner (2003), Poincaré–Friedrichs inequalities for piecewise  $h^1$  functions, *SIAM Journal on Numerical Analysis* **41**(1), 306–324.
- S. C. Brenner and L. R. Scott (2008), *The mathematical theory of finite element methods*, Vol. 15 of Texts in Applied Mathematics, third edition, Springer, New York.
- F. Brezzi (1974), On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers, *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge* **8**(R-2), 129–151.
- F. Brezzi, J. Douglas, Jr. and L. D. Marini (1985), Two families of mixed finite elements for second order elliptic problems, *Numer. Math.* **47**(2), 217–235.
- F. Brezzi, J. Douglas, Jr., M. Fortin and L. D. Marini (1987), Efficient rectangular mixed finite elements in two and three space variables, *RAIRO Modél. Math. Anal. Numér.* **21**(4), 581–604.

- F. Brezzi, G. Manzini, D. Marini, P. Pietra and A. Russo (2000), Discontinuous galerkin approximations for elliptic problems, *Numerical Methods for Partial Differential Equations: An International Journal* **16**(4), 365–378.
- C. Canuto and D. Fassino (2023), Higher-order adaptive virtual element methods with contraction properties, *Mathematics in Engineering* **5**(6), 1–33.
- R. Carroll, G. Duff, J. Friberg, J. Gobert, P. Grisvard, J. Nečas and R. Seeley (1966), *Equations aux dérivées partielles*, Vol. 19 of Séminaire de mathématiques supérieures, Les Presses de l'Université de Montréal'.
- C. Carstensen (1997), A posteriori error estimate for the mixed finite element method, *Math. Comp.* **66**(218), 465–476.
- C. Carstensen, M. Feischl, M. Page and D. Praetorius (2014), Axioms of adaptivity, *Comput. Math. Appl.* **67**(6), 1195–1253.
- J. M. Cascón and R. H. Nochetto (2012), Quasioptimal cardinality of AFEM driven by nonresidual estimators, *IMA J. Numer. Anal.* **32**(1), 1–29.
- J. M. Cascón, C. Kreuzer, R. H. Nochetto and K. G. Siebert (2008), Quasi-optimal convergence rate for an adaptive finite element method, *SIAM J. Numer. Anal.* **46**(5), 2524–2550.
- P. G. Ciarlet (2002), *The finite element method for elliptic problems*, Vol. 40 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. Reprint of the 1978 original [North-Holland, Amsterdam; MR0520174 (58 #25001)].
- P. Clément (1975), Approximation by finite element functions using local regularization, *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique* **9**(R2), 77–84.
- A. Cohen, R. DeVore and R. H. Nochetto (2012), Convergence rates of AFEM with  $H^{-1}$  data, *Found. Comput. Math.* **12**(5), 671–718.
- P. Daniel and M. Vohralik (2023), Guaranteed contraction of adaptive inexact  $hp$ -refinement strategies with realistic stopping criteria, *ESAIM Math. Model. Numer. Anal.* **57**(1), 329–366.
- E. B. Davies (1988), Lipschitz continuity of functions of operators in the Schatten classes, *J. London Math. Soc. (2)* **37**(1), 148–157.
- S. Dekel and D. Leviatan (2004), Whitney estimates for convex domains with applications to multivariate piecewise polynomial approximation, *Found. Comput. Math.* **4**(4), 345–368.
- R. A. DeVore (1998), Nonlinear approximation, in *Acta numerica, 1998*, Vol. 7 of Acta Numer., Cambridge Univ. Press, Cambridge, pp. 51–150.
- R. A. DeVore and G. G. Lorentz (1993), *Constructive approximation*, Vol. 303 of Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer-Verlag, Berlin.
- R. A. DeVore and V. A. Popov (1988), Interpolation of Besov spaces, *Trans. Amer. Math. Soc.* **305**(1), 397–414.
- L. Diening, L. Gehring and J. Storn (2023), Adaptive mesh refinement for arbitrary initial triangulations, Technical report, arXiv:2304.02674.
- L. Diening, C. Kreuzer and R. Stevenson (2016), Instance optimality of the adaptive maximum strategy, *Found. Comput. Math.* **16**(1), 33–68.
- Z. Ditzian (1988), On the Marchaud-type inequality, *Proc. Amer. Math. Soc.* **103**(1), 198–202.

- W. Dörfler (1996), A convergent adaptive algorithm for Poisson's equation, *SIAM J. Numer. Anal.* **33**(3), 1106–1124.
- N. Dunford and J. T. Schwartz (1988), *Linear operators. Part II*, Wiley Classics Library, John Wiley & Sons, Inc., New York. Spectral theory. Selfadjoint operators in Hilbert space, With the assistance of William G. Bade and Robert G. Bartle, Reprint of the 1963 original, A Wiley-Interscience Publication.
- T. Dupont and R. Scott (1980), Polynomial approximation of functions in Sobolev spaces, *Math. Comp.* **34**(150), 441–463.
- A. Ern, I. Smears and M. Vohralik (2017), Discrete p-robust H(div)-liftings and a posteriori estimates for elliptic problems with H-1-source terms, *Calcolo* **54**(3), 1009–1025.
- L. C. Evans (2010), *Partial differential equations*, Vol. 19 of Graduate Studies in Mathematics, second edition, American Mathematical Society, Providence, RI.
- B. Faermann (2000), Localization of the Aronszajn-Slobodeckij norm and application to adaptive boundary element methods. I. The two-dimensional case, *IMA J. Numer. Anal.* **20**(2), 203–234.
- B. Faermann (2002), Localization of the Aronszajn-Slobodeckij norm and application to adaptive boundary element methods. II. The three-dimensional case, *Numer. Math.* **92**(3), 467–499.
- M. Feischl (2019), Optimality of a standard adaptive finite element method for the Stokes problem, *SIAM J. Numer. Anal.* **57**(3), 1124–1157.
- M. Feischl (2022), Inf-sup stability implies quasi-orthogonality, *Math. Comp.* **91**(337), 2059–2094.
- F. Fierro and A. Veiser (2003), A Posteriori Error Estimators for Regularized Total Variation of Characteristic Functions, *SIAM Journal on Numerical Analysis* **41**(6), 2032–2055.
- S. Funken, D. Praetorius and P. Wissgott (2011), Efficient implementation of adaptive P1-FEM in Matlab, *Comput. Methods Appl. Math.* **11**(4), 460–490.
- G. P. Galdi (1994), *An introduction to the mathematical theory of the Navier-Stokes equations. Vol. I*, Vol. 38 of Springer Tracts in Natural Philosophy, Springer-Verlag, New York. Linearized steady problems.
- F. D. Gaspoz and P. Morin (2014), Approximation classes for adaptive higher order finite element approximation, *Math. Comp.* **83**(289), 2127–2160.
- F. D. Gaspoz and P. Morin (2017), Errata to “Approximation classes for adaptive higher order finite element approximation” [MR3223327], *Math. Comp.* **86**(305), 1525–1526.
- D. Gilbarg and N. S. Trudinger (2001), *Elliptic partial differential equations of second order*, Classics in Mathematics, Springer-Verlag, Berlin. Reprint of the 1998 edition.
- P. Grisvard (1985), *Elliptic Problems in Nonsmooth Domains*, Vol. 24 of Monographs and Studies in Mathematics, Pitman (Advanced Publishing Program), Boston, MA.
- P. Grisvard (2011), *Elliptic problems in nonsmooth domains*, Vol. 69 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. Reprint of the 1985 original [MR0775683], With a foreword by Susanne C. Brenner.
- W. Hackbusch (1992), *Elliptic differential equations. Theory and numerical treatment*, Vol. 18 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin.
- P. Houston, D. Schötzau and T. P. Wihler (2004), Mixed hp-discontinuous galerkin finite element methods for the Stokes problem in polygons, in *Numerical Mathematics and Advanced Applications: Proceedings of ENUMATH 2003 the 5th European Conference on Numerical Mathematics and Advanced Applications Prague, August 2003*, Springer, pp. 493–501.

- P. Houston, D. Schötzau and T. P. Wihler (2007), Energy norm a posteriori error estimation of hp-adaptive discontinuous galerkin methods for elliptic problems, *Mathematical Models and Methods in Applied Sciences* **17**(01), 33–62.
- D. Jerison and C. E. Kenig (1995), The inhomogeneous Dirichlet problem in Lipschitz domains, *J. Funct. Anal.* **130**(1), 161–219.
- J.-P. Kahane (1961), *Teoria constructiva de funciones*, Universidad de Buenos Aires, Buenos Aires.
- O. A. Karakashian and F. Pascal (2007), Convergence of adaptive discontinuous galerkin approximations of second-order elliptic problems, *SIAM Journal on Numerical Analysis* **45**(2), 641–665.
- R. B. Kellogg (1974/75), On the Poisson equation with intersecting interfaces, *Applicable Anal.* **4**, 101–129.
- C. Kreuzer and K. G. Siebert (2011), Decay rates of adaptive finite elements with Dörfler marking, *Numer. Math.* **117**(4), 679–716.
- C. Kreuzer and A. Veese (2019), Convergence of adaptive finite element methods with error-dominated oscillation, in *Numerical mathematics and advanced applications—ENUMATH 2017*, Vol. 126 of Lect. Notes Comput. Sci. Eng., Springer, Cham, pp. 471–479.
- C. Kreuzer and A. Veese (2021), Oscillation in a posteriori error estimation, *Numer. Math.* **148**(1), 43–78.
- C. Kreuzer, A. Veese and P. Zanotti (n.d.), ‘Accurate error bounds for finite element methods’. In preparation.
- P. D. Lax and A. N. Milgram (1954), Parabolic equations, in *Contributions to the theory of partial differential equations*, Annals of Mathematics Studies, no. 33, Princeton University Press, Princeton, N.J., pp. 167–190.
- G. Leoni (2009), *A first course in Sobolev spaces*, Vol. 105 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI.
- R. Luce and B. Wohlmuth (2004), A Local A Posteriori Error Estimator Based on Equilibrated Fluxes, *SIAM Journal on Numerical Analysis* **42**(4), 1394–1414.
- J. M. Maubach (1995), Local bisection refinement for  $n$ -simplicial grids generated by reflection, *SIAM J. Sci. Comput.* **16**(1), 210–227.
- N. G. Meyers (1963), An  $L^p$ -estimate for the gradient of solutions of second order elliptic divergence equations, *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (3)* **17**, 189–206.
- W. F. Mitchell (1989), A comparison of adaptive refinement techniques for elliptic problems, *ACM Trans. Math. Software* **15**(4), 326–347 (1990).
- P. Morin, R. Nochetto and K. Siebert (2003), Local problems on stars: A posteriori error estimators, convergence, and performance, *Mathematics of Computation* **72**(243), 1067–1097.
- P. Morin, R. H. Nochetto and K. G. Siebert (2000), Data oscillation and convergence of adaptive FEM, *SIAM J. Numer. Anal.* **38**(2), 466–488.
- P. Morin, R. H. Nochetto and K. G. Siebert (2002), Convergence of adaptive finite element methods, *SIAM Rev.* **44**(4), 631–658 (2003). Revised reprint of “Data oscillation and convergence of adaptive FEM” [*SIAM J. Numer. Anal.* **38** (2000), no. 2, 466–488 (electronic); MR1770058 (2001g:65157)].
- P. Morin, K. G. Siebert and A. Veese (2008), A basic convergence result for conforming adaptive finite elements, *Math. Models Methods Appl. Sci.* **18**(5), 707–737.
- J.-C. Nédélec (1980), Mixed finite elements in  $\mathbf{R}^3$ , *Numer. Math.* **35**(3), 315–341.

- J. Nečas (1962), Sur une méthode pour résoudre les équations aux dérivées partielles du type elliptique, voisine de la variationnelle, *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (3)* **16**, 305–326.
- R. H. Nochetto and A. Veese (2012), Primer of adaptive finite element methods, in *Multiscale and adaptivity: modeling, numerics and applications*, Vol. 2040 of Lecture Notes in Math., Springer, Heidelberg, pp. 125–225.
- R. H. Nochetto, K. G. Siebert and A. Veese (2009), Theory of adaptive finite element methods: an introduction, in *Multiscale, nonlinear and adaptive approximation*, Springer, Berlin, pp. 409–542.
- L. E. Payne and H. F. Weinberger (1960), An optimal Poincaré inequality for convex domains, *Arch. Rational Mech. Anal.* **5**, 286–292 (1960).
- I. Perugia and D. Schötzau (2003), The hp-local discontinuous galerkin method for low-frequency time-harmonic maxwell equations, *Mathematics of Computation* **72**(243), 1179–1214.
- P.-A. Raviart and J. M. Thomas (1977), A mixed finite element method for 2nd order elliptic problems, in *Mathematical aspects of finite element methods (Proc. Conf., Consiglio Naz. delle Ricerche (C.N.R.), Rome, 1975)*, Lecture Notes in Math., Vol. 606, Springer, Berlin-New York, pp. 292–315.
- R. Sacchi and A. Veese (2006), Locally efficient and reliable a posteriori error estimators for Dirichlet problems, *Math. Models Methods Appl. Sci.* **16**(3), 319–346.
- L. R. Scott and S. Zhang (1990), Finite element interpolation of nonsmooth functions satisfying boundary conditions, *Math. Comp.* **54**(190), 483–493.
- K. Siebert and A. Veese (2007), A Unilaterally Constrained Quadratic Minimization with Adaptive Finite Elements, *SIAM Journal on Optimization* **18**(1), 260–289.
- K. G. Siebert (2012), Mathematically founded design of adaptive finite element software, in *Multiscale and adaptivity: modeling, numerics and applications*, Vol. 2040 of Lecture Notes in Math., Springer, Heidelberg, pp. 227–309.
- R. Stevenson (2007), Optimality of a standard adaptive finite element method, *Found. Comput. Math.* **7**(2), 245–269.
- R. Stevenson (2008), The completion of locally refined simplicial partitions created by bisection, *Math. Comp.* **77**(261), 227–241.
- D. B. Szyld (2006), The many proofs of an identity on the norm of oblique projections, *Numerical Algorithms* **42**(3-4), 309–323.
- F. Tantardini, A. Veese and R. Verfürth (n.d.), ‘Best error localization in the approximation of functionals with piecewise polynomials’. In preparation.
- C. Taylor and P. Hood (1973), A numerical solution of the Navier-Stokes equations using the finite element technique, *Internat. J. Comput. & Fluids* **1**(1), 73–100.
- C. T. Traxler (1997), An algorithm for adaptive mesh refinement in  $n$  dimensions, *Computing* **59**(2), 115–137.
- H. Triebel (2010), *Theory of function spaces*, Modern Birkhäuser Classics, Birkhäuser/Springer Basel AG, Basel. Reprint of 1983 edition [MR0730762], Also published in 1983 by Birkhäuser Verlag [MR0781540].
- A. Veese (2002), Convergent adaptive finite elements for the nonlinear Laplacian, *Numerische Mathematik* **92**(4), 743–770.
- A. Veese (2016), Approximating gradients with continuous piecewise polynomial functions, *Found. Comput. Math.* **16**(3), 723–750.



- A. Veerer and R. Verfürth (2009), Explicit Upper Bounds for Dual Norms of Residuals, *SIAM Journal on Numerical Analysis* **47**(3), 2387–2405.
- R. Verfürth (2013), *A posteriori error estimation techniques for finite element methods*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford.
- J. Xu and L. Zikatanov (2003), Some observations on Babuška and Brezzi theories, *Numer. Math.* **94**(1), 195–202.
- J. Xu, L. Chen and R. H. Nochetto (2009), Optimal multilevel methods for  $H(\text{grad})$ ,  $H(\text{curl})$ , and  $H(\text{div})$  systems on graded and unstructured grids, in *Multiscale, nonlinear and adaptive approximation*, Springer, Berlin, pp. 599–659.