University of Amsterdam

MSc Mathematics

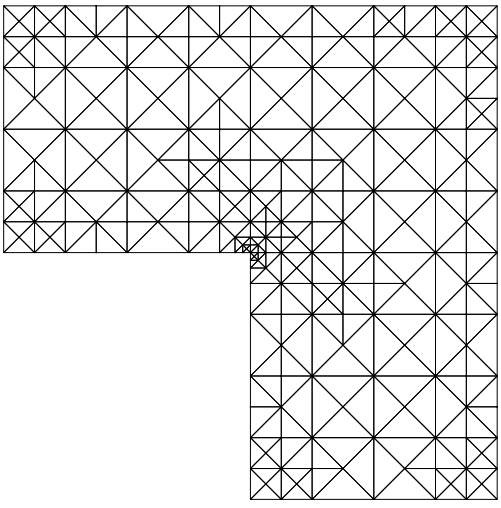
Master Thesis

Equilibrated flux estimator for the adaptive finite element method

Author:
Raymond van Venetië

Supervisor:
prof. dr. R.P. Stevenson

Examination date:
August 31, 2016



Korteweg-de Vries Institute for
Mathematics

Abstract

We investigate the equilibrated flux estimator for finite element solutions. This estimator provides a constant-free reliability bound and — more importantly — yields an efficiency estimate with a constant independent of the polynomial degree used in the finite element space. We prove that the adaptive finite element method driven by this estimator provides an optimal convergence rate. Performance of the equilibrated flux estimator is numerically compared with other well-established estimators.

Title: Equilibrated flux estimator for the adaptive finite element method
Cover image: An optimal triangulation found using AFEM with the equilibrated flux estimator
Author: Raymond van Venetië, raymond.van.venetie@gmail.com, 10004627
Supervisor: prof. dr. R.P. Stevenson
Second Examiner: dr. J.H. Brandts
Examination date: August 31, 2016

Korteweg-de Vries Institute for Mathematics
University of Amsterdam
Science Park 105-107, 1098 XG Amsterdam
`http://kdvi.uva.nl`

Contents

Introduction **4**

1. Theoretical Background **6**
1.1. Weak formulation 6
1.2. Finite element space 9
1.3. Interpolant 11
1.4. Error bounds 12
1.5. A posteriori error estimation 13
1.6. Adaptive finite element method 16

2. Equilibrated flux estimator **18**
2.1. Prager and Synge 18
2.2. Equilibration 20
2.3. Efficiency 23
2.4. Equivalence with the classical residual estimator 26
2.5. Discretization and oscillation 27

3. Optimality of the adaptive finite element method **33**
3.1. Design of the adaptive finite element method 33
3.2. Optimality conditions 35
3.3. Contraction property 38
3.4. Optimality of AFEM 41
3.5. Discussion 45

4. Practical aspects **48**
4.1. Ern and Vohralík's construction 48
4.2. Raviart-Thomas space 51
4.3. Explicit lowest order basis 54
4.4. Implementation of the lowest order estimator 55
4.5. Higher order Raviart-Thomas basis 56

5. Numerical results **60**
5.1. Exact error 60
5.2. Error estimators 61
5.3. Uniform refinements 61
5.4. Adaptive finite element method 66
5.5. Mixed finite element solution 66
5.6. Zienkiewicz-Zhu error estimator 67

5.7. Discussion 70

Popular summary **72**

Appendices **73**

A. Notation **74**

B. Definitions and Reference theorems **76**
B.1. Sobolev spaces 76
B.2. Poincaré-Friedrichs inequality 77
B.3. Raviart-Thomas elements 77
B.4. Auxiliary results 78

Bibliography **80**

Introduction

Many years have passed since the first appearance of the *finite element method*. For a large class of boundary value problems, it has proved to be the approximation method of choice. Let Ω be the domain of the boundary value problem of interest. A general finite element method can be characterized as follows. First, the domain Ω is partitioned into a set of *elements*. Each of these elements is equipped with a (local) function space. Assembling these local spaces together results in a function space $\mathbb{V}$ on the entire domain Ω. A finite element *approximation* U is then computed in $\mathbb{V}$, such that U is close to the exact solution in some sense; typically U is a projection of the exact solution on $\mathbb{V}$.

In this work, we concentrate on second-order partial differential equations, with $\mathbb{V}$ a space of element-wise polynomials of *fixed* degree. Write u for the exact solution of the associated boundary value problem. In general, one wants to construct a sequence of approximations that converge to the exact solution. In the original finite element method such a sequence is produced by taking approximations on repeated subdivisions of the partition.

For smooth solutions u, it has been shown that an optimal convergence rate is obtained by subdividing each element of the partition. In the two-dimensional case with triangular elements, this could correspond to subdividing each triangle into four congruent subtriangles. All of this theory is well established and summarized in various books, e.g. [11, 44].

For less smooth u, this simplistic refinement strategy unfortunately does not provide an optimal convergence rate. Instead of uniformly subdividing each element, one could consider a more thoughtful refinement strategy. An obvious modification would be refining only those elements for which the current approximation error is large in some sense. This idea is formalised by the *adaptive* finite element method. An essential ingredient of the adaptive method is an *error estimator*, i.e. a function that estimates the error between u and U for every element in the partition — without knowing u.

Under some assumptions, refining those elements for which the error estimator is relatively large leads to an optimal convergence rate for non-smooth solutions u as well. The existence of such an error estimator is quite fascinating if one thinks about it: we are somehow estimating the error, without actually knowing the exact solution u. A proof of this optimality and an overview of the various contributions that lead to this proof can be found in [13].

The first optimality proofs for the adaptive finite element method used the standard *residual* estimator [6, 33, 13]. Although the estimator provides optimality, it is not favourable in other aspects. For one, the bounds provided by the error estimator contain unknown constants. These constants make it harder to determine the quality of approximations U in practice. Fortunately, a variety of alternative (nonresidual) estimators are

documented in the literature [38].

This work will focus on of these estimators in particular: the *equilibrated flux* estimator [10, 9, 19]. This estimator provides a true upper bound for the local error without an unknown constant, eliminating one of the shortcomings of the standard residual estimator. Braess, Pillwein and Schöberl [9] showed another — arguably more important — property of this estimator: *polynomial-robustsness*. That is, the constant appearing in the lower bound is independent of the polynomial degree used in $\mathbb{V}$. This latter condition might enable one to prove optimality of yet another version of finite element method, the so-called *hp*-version. In this extension of the adaptive method, one lets the polynomial degree vary per element as well. Recent experiments show that this can lead to an exponential rate of convergence [16].

These convenient properties provide reason for an in-depth study of this estimator. In this work, we will prove that the adaptive finite element method driven by the equilibrated flux estimator exhibits an optimal convergence rate. This proof will be based on the general results provided by Cascón and Nochetto in [12]. We will provide details for some of the unproven claims stated in [12].

The advantages of the equilibrated flux estimator come at the price of implementational complexity. We shall therefore discuss an equivalent — and easier to implement — construction of the equilibrated flux estimator as presented by Ern and Vohralík in [19]. The equilibrated flux estimator for the lowest order finite element space is implemented using this alternative construction. The implementation will be used to analyze and compare the equilibrated flux estimator with the standard residual estimator.

This work is organized as follows. In Chapter 1 we give a summary of the finite element theory and introduce some notation. The equilibrated flux estimator is formally introduced in Chapter 2. The main result — optimality of the adaptive finite element method driven by the equilibrated flux estimator — is given in Chapter 3. An equivalent construction of the equilibrated flux estimator, and some implementational issues are discussed in Chapter 4. Numerical results and a comparison with the standard residual estimator is presented in Chapter 5. An overview of the notation used and some reference theorems are given in the appendices.

1. Theoretical Background

This chapter contains a summary of the finite element method — a method for approximating solutions of partial differential equations. It is mainly included for self-containedness; a complete mathematical derivation of the theory is well documented in the literature, e.g. [11, 44]. We shall focus on second-order elliptic boundary value problems. For a bounded Lipschitz domain $\Omega \subset \mathbb{R}^n$, the general boundary value problem with Dirichlet boundary conditions is solving u from

$$\begin{aligned} -\nabla \cdot (\mathbf{A}\nabla u) + \underline{b} \cdot \nabla u + c &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{1.1}$$

The theory works equally well for other boundary conditions like Neumann, or mixed; we omit these types for ease of presentation and calculations. Under mild conditions on $\mathbf{A}, \underline{b}, c$ and f, we can approximate the solution u of this problem using the finite element method.

1.1. Weak formulation

The first step is translating the boundary value problem (1.1) into a *weak formulation*. That is, we will define a space V of test functions and consider the problem of finding u such that (1.1) holds in a distributional sense. Suppose that u solves (1.1) with $\mathbf{A}, \underline{b}, c, f$ sufficiently smooth functions, then for $v \in V$ we calculate

$$\begin{aligned} \langle f, v\rangle_\Omega =& \langle -\nabla \cdot (\mathbf{A}\nabla u) + \underline{b} \cdot \nabla u + c, v\rangle_\Omega \\ =& \int_\Omega \left(-\nabla \cdot \left(\mathbf{A}(x)\nabla u(x)\right)\right) v(x) + \left(\underline{b}(x) \cdot \nabla u(x)\right) v(x) + c(x)v(x) \, \mathrm{d}x \\ =& \int_{\partial\Omega} - \left(\mathbf{A}(x)\nabla u(x) \cdot n\right) v(x) \, \mathrm{d}x \\ &+ \int_\Omega \left(\mathbf{A}(x)\nabla u(x)\right) \cdot \nabla v(x) + \left(\underline{b}(x) \cdot \nabla u(x)\right) v(x) + c(x)v(x) \, \mathrm{d}x, \end{aligned}$$

where the third equality follows from the Divergence theorem, see B.2. We write $\langle \cdot, \cdot \rangle_\omega$ to mean the standard — possibly vector-valued — $L^2(\omega)$-inner product on any domain ω. Suppose that all test functions v vanish on the boundary $\partial\Omega$, the boundary integral then disappears and the above simplifies to

$$F(v) := \langle f, v\rangle_\Omega = \langle \mathbf{A}\nabla u, \nabla v\rangle_\Omega + \langle \underline{b} \cdot \nabla u + cu, v\rangle_\Omega =: a(u, v). \tag{1.2}$$

A weak solution of (1.1) is a function u that solves the above equation for all $v \in V$. It turns out that there are conditions under which there is a unique weak solution u. To

see this, we must formalise the previous calculations. First, we will tackle the definition of V.

The weak formulation uses the $L^2(\Omega)$-inner product, so we naturally require $V \subset L^2(\Omega)$. All the partial derivatives of v are taken in a distributive sense. For these to be correctly defined, we need that v is at least weakly differentiable. By combining both requirements we deduce that V should lie in the suitable space $W_2^1(\Omega) = H^1(\Omega)$ (conveniently, this is a Hilbert space). We must check that the calculations, in particular the Divergence Theorem, are still valid for $v \in H^1(\Omega)$; of course this is the case, cf. [11, Ch. 5]. Formalizing the vanishing boundary requirement can be done using the trace operator[1] T, i.e. we should have $v \in H^1(\Omega)$ such that $\|Tv\|_{L^2(\partial\Omega)} = 0$. All together this yields

$$V := \Big\{ v \in H^1(\Omega) : v|_{\partial\Omega} = 0 \Big\} = H_0^1(\Omega).$$

The Lax-Milgram theorem can be used to show that certain weak formulations have a unique solution.

Theorem 1.1 (Lax-Milgram, [11, §2])**.** *Given a Hilbert space $(V, \langle\cdot,\cdot\rangle)$ and a bilinear form $a(\cdot,\cdot)$ on V which is*

$$\begin{aligned} \textit{Continuous:} \quad & \exists C < \infty \quad \textit{s.t.} \quad |a(v,w)| \leqslant C\|v\|\|w\| \quad \forall v, w \in V; \\ \textit{Coercive:} \quad & \exists \alpha > 0 \quad \textit{s.t.} \quad a(v,v) \geqslant \alpha\|v\|^2 \quad \forall v \in V. \end{aligned}$$

For every continuous linear functional $F \in V'$, there exists a unique $u \in V$ such that

$$a(u,v) = F(v) \quad \forall v \in V.$$

Before we can apply Lax-Milgram to our weak formulation (1.2) we need restrictions on the parameters $\mathbf{A}, \underline{b}, c$, and f. We assume that:

(a) $\mathbf{A}(x) = (a_{ij}(x))_{1\leqslant i,j\leqslant n}$ is symmetric — $a_{ij} = a_{ji}$ — with $a_{ij} \in L^\infty(\Omega)$. Furthermore, we suppose that the matrix is uniformly elliptic, i.e. there is a positive constant α such that

$$\zeta \cdot (\mathbf{A}(x)\zeta) \geqslant \alpha\|\zeta\|^2 \quad \forall \zeta \in \mathbb{R}^n \quad \text{a.e. in } \Omega;$$

(b) $\underline{b}(x) = \big(b_1(x), \ldots, b_n(x)\big)^\top$ with $b_i \in L^\infty(\Omega)$ such that the weak divergence vanishes, i.e. $\nabla \cdot \underline{b} = 0$ in Ω;

(c) $c \in L^\infty(\Omega)$ is non-negative;

(d) $f \in L^2(\Omega)$.

Under these assumptions, Lax-Milgram can be applied to the weak formulation (1.2): the test space $V = H_0^1(\Omega)$ is a Hilbert space, the form $a(\cdot,\cdot)$ is bilinear, the functional F is linear, and continuity of F follows from Cauchy-Schwarz, i.e.

$$|F(v)| = |\langle f, v\rangle_\Omega| \leqslant \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \leqslant \|f\|_{L^2(\Omega)}\|v\|_{H^1(\Omega)}.$$

We are left to check the continuity and coercivity of the bilinear form $a(\cdot,\cdot)$.

[1] Here we need that Ω has a Lipschitz boundary, cf. [11, Ch. 1.6].

Continuity Using Hölder's inequality we find

$$\begin{aligned}
|a(v,w)| &\leqslant \int_\Omega \left| \sum_{i,j} a_{ij} \tfrac{\partial v}{\partial x_j} \tfrac{\partial w}{\partial x_i} + \sum_i b_i \tfrac{\partial v}{\partial x_i} w + cvw \right| \\
&\leqslant \sum_{i,j} \|a_{ij}\|_{L^\infty(\Omega)} \left\| \tfrac{\partial v}{\partial x_j} \right\|_{L^2(\Omega)} \left\| \tfrac{\partial w}{\partial x_i} \right\|_{L^2(\Omega)} + \sum_i \|b_i\|_{L^\infty(\Omega)} \left\| \tfrac{\partial v}{\partial x_i} \right\|_{L^2(\Omega)} \|w\|_{L^2(\Omega)} \\
&\quad + \|c\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \|w\|_{L^2(\Omega)} \\
&\leqslant C \|v\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)},
\end{aligned}$$

where C is a finite constant because $a_{ij}, b_i, c \in L^\infty(\Omega)$. So $a(\cdot,\cdot)$ is indeed continuous.

Coercivity This is where our extra assumptions come into play. Since $\underline{b}$ has a vanishing weak divergence, a density argument[2] shows that for $v \in H^1_0(\Omega)$ we have

$$\int_\Omega (\underline{b} \cdot \nabla v)\, v = \int_\Omega \underline{b} \cdot \left(\frac{1}{2} \nabla v^2 \right) = - \int_\Omega (\nabla \cdot \underline{b}) \, \frac{1}{2} v^2 = 0.$$

By non-negativity of c, we have $cv^2 \geqslant 0$ on Ω. The Poincaré-Friedrich inequality — stated in §B.2 — provides us with a constant C_Ω such that $\|v\|_{L^2(\Omega)} \leqslant C_\Omega \|\nabla v\|_{L^2(\Omega)}$ for $v \in H^1_0(\Omega)$. Combining this with the uniform ellipticity of $\mathbf{A}$ yields,

$$\begin{aligned}
a(v,v) &= \int_\Omega (\mathbf{A} \nabla v) \cdot \nabla v + (\underline{b} \cdot \nabla v)\, v + cv^2 \\
&\geqslant \int_\Omega \big(\mathbf{A}(x) \nabla v(x) \big) \cdot \nabla v(x) \, \mathrm{d}x \\
&\geqslant \int_\Omega \alpha \|\nabla v(x)\|^2 \, \mathrm{d}x = \alpha \|\nabla v\|^2_{L^2(\Omega)} \\
&\geqslant \frac{\alpha}{2} \left(\frac{1}{C_\Omega^2} \|v\|^2_{L^2(\Omega)} + \|\nabla v\|^2_{L^2(\Omega)} \right) \geqslant \beta \|v\|^2_{H^1(\Omega)}.
\end{aligned}$$

This shows the coercivity of $a(\cdot,\cdot)$.

The above shows that we may invoke Lax-Milgram to obtain the desired result, namely that there is a unique $u \in V$ that solves the weak formulation (1.2).

We cannot expect to (easily) solve the weak formulation in its current form, since this is still an infinite dimensional problem. Instead one generally solves the Galerkin approximation problem:

$$\text{Given a finite-dimensional } \mathbb{V} \subset V\text{: find } U \in \mathbb{V} \text{ s.t. } a(U,v) = F(v) \quad \forall v \in \mathbb{V}. \tag{1.3}$$

This is a finite dimensional problem, and carefully choosing subspaces $\mathbb{V}$ can lead to increasing accuracy of the approximation U — the *discrete* solution. Another application of Lax-Milgram applied to $\mathbb{V}$ shows that this system has a unique solution as well.

[2] Recall that $C^\infty_c(\Omega)$ is dense in $H^1_0(\Omega)$.

1.2. Finite element space

Construction of subspaces $\mathbb{V} \subset V$ can be done in a systematic way using *finite elements*. The general idea is to split the domain Ω into elements K, each with its own function space. The space $\mathbb{V}$ is then constructed by glueing these spaces together. In this section, we will summarize the definitions and results presented in [11, Ch 3].

Definition 1.1. A (unisolvent) finite element is a triplet $(K, \mathcal{P}, \mathcal{N})$ where

(a) $K \subset \mathbb{R}^n$ is the *element domain*: a bounded closed set with nonempty interior and piecewise smooth boundary;

(b) $\mathcal{P}$ is the space of *shape functions*: a finite-dimensional space of functions on K;

(c) $\mathcal{N} = \{N_1, \ldots, N_k\}$ is the set of *nodal variables*: a basis for $\mathcal{P}'$.

Definition 1.2. Let an element $(K, \mathcal{P}, \mathcal{N})$ be given. The basis $\{\phi_1, \ldots, \phi_k\}$ of $\mathcal{P}$ dual to $\mathcal{N}$ — meaning $N_i(\phi_j) = \delta_{ij}$ — is called the *nodal basis* of $\mathcal{P}$.

Lemma 1.2. *Let $\mathcal{P}$ be a k-dimensional vector space with $\mathcal{N} = \{N_1, \ldots, N_k\} \subset \mathcal{P}'$. The set $\mathcal{N}$ is a basis for $\mathcal{P}'$ if and only if $\mathcal{N}$ determines $\mathcal{P}$, i.e. if $v \in \mathcal{P}$ with $N(v) = 0$ for all $N \in \mathcal{N}$ implies that $v = 0$.*

Informally, a finite element space $\mathbb{V}$ can now be introduced. Consider a partition $\mathcal{T}$ of the domain Ω into finite elements, i.e. $\overline{\Omega} = \cup_{K \in \mathcal{T}} K$ where $(K, \mathcal{P}_K, \mathcal{N}_K)$ is a finite element. We can simply define $\mathbb{V}$ to be the space of functions in V that coincide with the shape functions on each element, i.e. $v \in \mathbb{V}$ if $v \in V$ and $v|_K \in \mathcal{P}_K$ for all $K \in \mathcal{T}$. Without further conditions on the partition we have no way of giving smoothness properties for the subspace $\mathbb{V}$ — we do not even know if it is correctly defined[3]. A piecewise polynomial on $\mathcal{T}$ is in $H^1(\Omega)$ if and only if it is continuous over all interior edges of $\mathcal{T}$. It is therefore reasonable to take a subspace $\mathbb{V}$ consisting of continuous functions.

In the literature, a variety of finite elements are documented, each element having its pros and cons. For now we will be using the linear *Lagrange element*.

Example 1.1. For a triangle K, the lowest order Lagrange element — also linear Lagrange — is given by $(K, \mathcal{P}_1(K), \{N_1, N_2, N_3\})$, with $\mathcal{P}_1(K)$ the linear polynomials on K, and functionals $N_i(v) = v(z_i)$, the evaluation of v in the vertices z_i of K.

This element can be generalised to a Lagrange element of arbitrary order p, whose shape functions are given by $\mathcal{P}_p(K)$: polynomials of degree p on K.

We need more regularity of the partition to ensure continuity of the subspace $\mathbb{V}$. In this work we shall consider subdivisions of the domain into simplices (triangles in $\mathbb{R}^2$, tetrahedra in $\mathbb{R}^3$). Implicitly we therefore also require Ω to have a polyhedral boundary. Because of the continuity requirement, we need the partition to be *conforming*.

[3] In general, two adjacent elements have a non-finite intersection. Therefore $v \in \mathbb{V}$ might be incorrectly defined on this intersection in a classical sense.

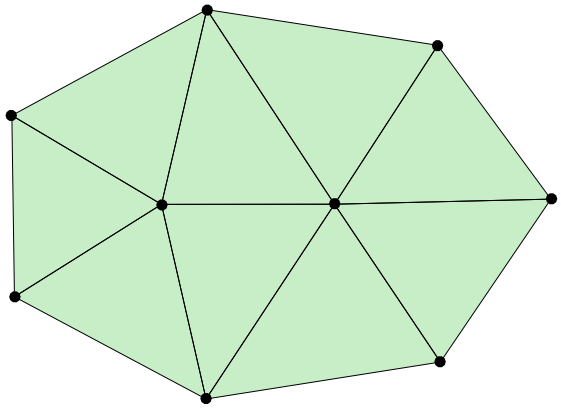


Figure 1.1.: Example of a triangulation for some two-dimensional domain. It has 7 boundary vertices, and 2 interior vertices.

Definition 1.3. A triangulation $\mathcal{T}$ is a conforming partition of the domain Ω into a finite family of simplices. Formally,

- $\overline{\Omega} = \cup_{K \in \mathcal{T}} K$;
- $\mathring{K}_i \neq \emptyset, \mathring{K}_i \cap \mathring{K}_j = \emptyset \quad \forall K_i, K_j \in \mathcal{T}, K_i \neq K_j$;
- If $F = K_i \cap K_j$ for $K_i \neq K_j$, then F is a common (lower dimensional) face of K_i and K_j.

An example of a triangulation is given in Figure 1.1.

Consider a triangulation $\mathcal{T}$ and suppose that $\mathbb{V}$ consists of continuous functions v that are linear polynomials when restricted to each triangle $K \in \mathcal{T}$. Formally

$$\mathbb{V} = \mathbb{V}(\mathcal{T}) := \left\{ v \in C(\overline{\Omega}) : v|_K \in \mathcal{P}_1(K), \quad \forall K \in \mathcal{T} \right\}. \tag{1.4}$$

Why would one look at finite elements $(K, \mathcal{P}, \mathcal{N})$? Instead, one could just try to find a basis for this subspace, without ever needing to define finite elements. It appears however, that for analysis of the approximation error these finite elements are extremely useful.

Linear Lagrange elements are an obvious choice for constructing the continuous piecewise linear subspace $\mathbb{V}$. Every function $v \in \mathbb{V}$ is completely determined by its value in the vertices of $\mathcal{T}$, because of continuity and piecewise linearity. The evaluation of v at a vertex coincides with the choice of the nodal variables for a linear Lagrange element. In other words, for a piecewise polynomial v to be continuous, one requires that $N(v)$ agrees for all nodal variables N that are associated to the same vertex in $\mathcal{T}$. In a sense, one can glue together the nodal variables associated to the same vertex, which results in a set of global variables $\mathcal{N}_\Omega$ — a basis for the bigger space $\mathbb{V}'$.

As a result, we may equivalently define $\mathbb{V}$ to be the span of the global nodal basis with respect to $\mathcal{N}_\Omega$. In this construction we ignored the boundary condition, i.e. $v \in \mathbb{V}$ must

vanish on the boundary. This is easily mended; we can just consider the (non-trivial) nodal variables restricted to $C_0(\overline{\Omega})'$. In the linear Lagrange case this corresponds to removing nodal variables associated with boundary vertices.

In short, a basis for the space $\mathbb{V}'$ can be found by (correctly) glueing together the nodal variables, where we restrict the these global nodal variables to the space $C_0(\overline{\Omega})'$ to ensure the boundary condition. For the Galerkin approximation (1.3) we need more than just continuity; we must actually have $\mathbb{V} \subset H_0^1(\Omega)$. This weak differentiability condition is satisfied if one uses Lagrange elements, as summarized in the next lemma.

Lemma 1.3 ([11, p. 3.3.17])**.** *Let $\mathcal{T}$ be a triangulation of the domain Ω, with $\mathbb{V}$ the finite element space generated by using Lagrange elements. Then a function $v \in \mathbb{V}$ is continuous, weakly differentiable and vanishes on the boundary, i.e.*

$$\mathbb{V} = \left\{ v \in C(\overline{\Omega}) \cap H_0^1(\Omega) : v|_K \in \mathcal{P}_1(K), \quad \forall K \in \mathcal{T} \right\}.$$

1.3. Interpolant

In the literature, the finite element space is frequently defined in terms of the *interpolant* (cf. [11, Ch 3]). This interpolant is also an important tool in deriving error bounds.

Definition 1.4. Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element with nodal basis $\{\phi_1, \ldots, \phi_k\}$. Then for every function v for which all of the nodal variables $N \in \mathcal{N}$ are defined, the local interpolant is given by

$$I_K v = \sum_{i=1}^{k} N_i(v)\phi_i.$$

Definition 1.5. Given a triangulation $\mathcal{T}$ where each $K \in \mathcal{T}$ has an associated finite element triplet $(K, \mathcal{P}, \mathcal{N})$. Then for every v that is in the domain of each local interpolant the global interpolant $I_{\mathcal{T}} v$ reads as

$$I_{\mathcal{T}} v|_K = I|_K v \quad \forall K \in \mathcal{T}.$$

We would like this global interpolant to preserve some regularity. Consider linear Lagrange elements again; one easily sees that the global interpolant can be expressed using the global nodal variables, i.e.

$$I_{\mathcal{T}} v = \sum_{i=1}^{m} N_i(v)\Phi_i \quad \text{for } \mathcal{N}_\Omega = \{N_1, \ldots, N_m\} \text{ and its global nodal basis } \Phi_1, \ldots, \Phi_m.$$

From Lemma 1.3 we then see that for $g \in C(\overline{\Omega})$ one has $I_{\mathcal{T}} g \in C^0(\overline{\Omega}) \cap H_0^1(\Omega)$. We say that the interpolant has *continuity order 0*, and the finite element space $\mathbb{V}$ is said to be a C^0 finite element space.

1.4. Error bounds

Given the finite element space $\mathbb{V}$, one can determine the Galerkin approximation (1.3). The immediate question is of course "what is the quality of the approximation?". Can we find (tight) bounds on the approximation error?

The first such bound is provided by Céa's lemma (cf. [11, Thm 2.8.1]).

Lemma 1.4. *Suppose the assumptions of the Lax-Milgram theorem 1.1 hold. Then for the discrete solution $U \in \mathbb{V}$ of the finite Galerkin approximation* (1.3) *we have*

$$\|u - U\|_{H^1(\Omega)} \leqslant \frac{C}{\alpha} \min_{v \in \mathbb{V}} \|u - v\|_{H^1(\Omega)},$$

with C the continuity constant and α the coercivity constant of $a(\cdot, \cdot)$ on $\mathbb{V}$.

The above lemma tells us that the finite element solution U is a quasi-best approximation from the subspace $\mathbb{V}$. For an upper bound on the approximation error, we are only left to find an element in $v \in \mathbb{V}$ for which we can bound $\|u - v\|_{H^1(\Omega)}$. This is where the interpolant comes into play, because $I_{\mathcal{T}} u \in \mathbb{V}$ if we assume that $u \in C(\overline{\Omega})$.

The general idea is to bound $\|v - I_{\mathcal{T}} v\|_{H^1(\Omega)}$ in terms of local error bounds for each element $K \in \mathcal{T}$. First, an error bound is derived for a *reference element* $\hat{K}$. Then, under certain conditions, this bound can be used to estimate the interpolation error on each of the elements $K \in \mathcal{T}$. The idea of first considering a reference element is often used in finite element analysis and implementation. These extra conditions are formalised by the following definitions.

Definition 1.6. A finite element $(K, \mathcal{P}, \mathcal{N})$ is affine-interpolation equivalent to $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ iff there exists an affine map $F(\hat{x}) = B\hat{x} + c$ (B non singular) such that for $f = \hat{f} \circ F$,

(i) $F(K) = \hat{K}$;

(ii) $\mathcal{P} = \{f : \hat{f} \in \hat{\mathcal{P}}\}$;

(iii) $I_{\hat{\mathcal{N}}} \hat{f} = (I_{\mathcal{N}} f) \circ F$ for all sufficiently smooth $\hat{f}$.

Definition 1.7. Denote $h_K := \operatorname{diam}(K)$ and $p_K := \sup\{\operatorname{diam}(S) : S \text{ ball}, S \subset K\}$. A family of finite elements $(K, \mathcal{P}, \mathcal{N})$ is *uniformly shape regular* if

$$\sup_K h_K / p_K < \infty.$$

In words, h_K is the largest ball containing K, whilst p_K is the largest ball inside K. Shape regularity therefore ensures that triangles are not relatively *degenerate*. Using the Sobolev inequalities, the Bramble-Hilbert lemma and the Transformation lemma, one can prove the following (see [15, Ch 3] or [34] for a concise overview).

Theorem 1.5. *Let $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ be a finite element and suppose that $m - n/2 > l$,*

$$\mathcal{P}_{m-1}(\hat{K}) \subset \hat{\mathcal{P}} \subset H^m(\hat{K}) \quad \textit{and} \quad \hat{N} \subset C^l(\hat{K})'.$$

Then there exists a constant $C(\hat{K}, \hat{P}, \hat{N})$ *such that for all* $(K, \mathcal{P}, \mathcal{N})$ *affine-interpolation equivalent to* $(\hat{K}, \hat{P}, \hat{N})$*, and for* $0 \leqslant s \leqslant m$*,*

$$|v - I_K v|_{H^s(K)} \leqslant C \frac{h_K^m}{p_K^s} |v|_{H^m(K)} \quad \forall v \in H^m(\Omega).$$

If in addition a family $(K, \mathcal{P}, \mathcal{N})$ *is also uniformly shape regular the above is reduced to,*

$$|v - I_K v|_{H^s(K)} \leqslant C h_K^{m-s} |v|_{H^m(K)} \quad \forall v \in H^m(\Omega).$$

This theorem provides us with local error bounds. As a corollary we can now give an error bound for the global interpolant.

Theorem 1.6. *Let* $(\mathcal{T})$ *be a family of uniformly shape regular triangulations of a domain* $\Omega \subset \mathbb{R}^n$*. Suppose that all the finite elements are affine-interpolation equivalent to a reference element* $(\hat{K}, \hat{P}, \hat{N})$*, then under the conditions of the previous theorem we find for* $0 \leqslant s \leqslant m$ *and* $h := \max_{K \in \mathcal{T}} h_K$*,*

$$\left(\sum_{K \in \mathcal{T}} \|v - I_K v\|_{H^s(K)}^2 \right)^{1/2} \leqslant C h^{m-s} |v|_{H^m(\Omega)} \quad \forall v \in H^m(\Omega).$$

In case the global interpolant satisfies $I_{\mathcal{T}} C^l(\overline{\Omega}) \subset C^{m-1}(\overline{\Omega})$*, the above left hand side becomes a norm, i.e.*

$$\|v - I_{\mathcal{T}} v\|_{H^s(\Omega)} \leqslant C h^{m-s} |v|_{H^m(\Omega)} \quad \forall v \in H^m(\Omega).$$

Approximate the elliptic problem (1.1) for a family $\mathcal{T}$ of linear Lagrange triangulations and denote U for respective the discrete solutions. Additionally, if the exact u satisfies $u \in H_0^1(\Omega) \cap H^2(\Omega)$, then Céa's lemma combined with the above theorem tells us that[4]

$$\|u - U\|_{H^1(\Omega)} \leqslant \frac{C_{cea}}{\alpha} \min_{v \in \mathbb{V}} \|u - v\|_{H^1(\Omega)} \leqslant \frac{C_{cea}}{\alpha} \|u - I_{\mathcal{T}} u\|_{H^1(\Omega)} \leqslant \frac{C_{cea}}{\alpha} C h |u|_{H^2(\Omega)}. \quad (1.5)$$

This provides us with a concrete convergence proof.

1.5. A posteriori error estimation

This last bound (1.5) shows convergence of the finite element method: the approximation error decreases when solving U from spaces $\mathbb{V}$ with decreasing h. This so-called *a priori* error estimation (1.5) provides an error bound using problem specific information, e.g. $u \in H^2(\Omega)$. It does not capture any local information; it does not tell *where* the approximation error is substantial.

In practice one often does not have the smoothness required to use (1.5), e.g. $u \notin H^2(\Omega)$ if the domain Ω has a re-entrant corner, which is the case for all non-convex polygon

[4] Here we also used that the global interpolant preserves the Dirichlet boundary condition so that $I_{\mathcal{T}} u$ is actually an element of $\mathbb{V} \subset H_0^1(\Omega)$.

domains. Moreover, one expects a higher order convergence rate when using higher order (Lagrange) elements. Indeed, for the p-th degree finite element space we could obtain the error estimation $\|u-U\|_{H^1(\Omega} \leqslant Ch^p|u|_{H^{p+1}(\Omega)}$ from Theorem 1.6. Unfortunately, this requires even more smoothness of the exact solution, i.e. $u \in H^{p+1}(\Omega)$. This weakness can — to some extent — be overcome using an alternative refinement strategy.

Rather than globally reducing h, it might be more efficient to consider refinements of $\mathcal{T}$ with the diameter only locally reduced. This is where *a posteriori* error estimators come into play. Given a discrete solution U, one can use an a posteriori estimator to indicate on which elements $K \in \mathcal{T}$ the error $\|u-U\|_K$ is (relatively) large.

We follow the terminology used in [34]. For simplicity we restrict ourself to the two-dimensional Poisson problem, i.e. given a polyhedral domain $\Omega \subset \mathbb{R}^2$ and $f \in L^2(\Omega)$:

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{1.6}$$

The bilinear form of the weak formulation is given by $a(v,w) = \langle \nabla v, \nabla w \rangle_\Omega$ with test and trial space $V = H_0^1(\Omega)$. This bilinear form induces the energy seminorm,

$$|||v|||_\Omega := a(v,v)^{1/2} = \langle \nabla v, \nabla v \rangle_\Omega^{1/2} = |v|_{H^1(\Omega)} \quad \forall v \in V,$$

which is a norm on V thanks to the Poincaré-Friedrichs inequality (cf. §B.2).

Let $\mathcal{T}$ be a uniformly shape regular triangulation of Ω. We solve the discrete solution U from $\mathbb{V}(\mathcal{T})$, the Lagrange finite element space of degree p, i.e.

$$\mathbb{V}(\mathcal{T}) = \left\{ v \in H_0^1(\Omega) \cap C(\overline{\Omega}) : v|_K \in \mathcal{P}_p(K) \, \forall K \in \mathcal{T} \right\}.$$

The classical residual error estimator based on the residual $r \in V'$. In bracket notation for functionals, the latter is defined by

$$\langle r, v \rangle := a(u-U,v) = \langle f, v \rangle_\Omega - a(U,v).$$

The residual is closely related to the approximation error; one easily deduces that

$$|||u-U|||_\Omega = \sup_{0 \neq v \in V} \frac{a(u-U,v)}{|v|_{H^1(\Omega)}} = \sup_{0 \neq v \in V} \frac{\langle r, v \rangle}{|v|_{H^1(\Omega)}}. \tag{1.7}$$

For any $\tilde{v} \in \mathbb{V}(\mathcal{T})$, we can bound the residual using Galerkin orthogonality:

$$\begin{aligned} a(u-U,v) &= a(u-U, v-\tilde{v}) \\ &= \sum_{K \in \mathcal{T}} \int_K f(v-\tilde{v}) + \nabla U \cdot \nabla (v-\tilde{v}) \\ &= \sum_{K \in \mathcal{T}} \left[\int_K (f + \Delta U)(v-\tilde{v}) + \int_{\partial K} (\nabla U \cdot n)(v-\tilde{v}) \right] \\ &= \sum_{K \in \mathcal{T}} \left[\int_K (f + \Delta U)(v-\tilde{v}) \right] + \sum_{e \in \mathcal{E}^{int}} \left[\int_e [\![\nabla U]\!] (v-\tilde{v}) \right] \\ &\leqslant \sum_{K \in \mathcal{T}} \|f + \Delta U\|_K \|v-\tilde{v}\|_K + \sum_{e \in \mathcal{E}^{int}} \|[\![\nabla U]\!]\|_e \|v-\tilde{v}\|_e. \end{aligned} \tag{1.8}$$

Here $[\![\nabla U]\!]$ is the jump of $\nabla U \cdot n$ over an interface in the direction of the unit normal n,

$$[\![\nabla U]\!](x) := \lim_{\epsilon \to 0} \nabla U(x + \epsilon n) \cdot n - \nabla U(x - \epsilon n) \cdot n,$$

and $\mathcal{E}^{int}$ is the set of interior edges of $\mathcal{T}$. The equalities follow from integration by parts and the fact that v vanishes on the boundary edges of $\mathcal{T}$. Suppose that we can pick $\tilde{v}$ such that $\|v - \tilde{v}\|$ is bound by its seminorm $|v|_{H^1}$. Then combining the above with (1.7) yields an error estimation for $|||u - U|||$ in terms of localized errors on elements and edges.

We cannot use the interpolants from the previous section to construct $\tilde{v}$ since they require smoothness of v beyond being in $H^1_0(\Omega)$. The so-called *Scott-Zhang* interpolant [32] does the trick. That is, let $I_{\mathcal{T}} : H^1_0(\Omega) \to \mathbb{V}(\mathcal{T})$ be the Scott-Zhang interpolant onto $\mathbb{V}(\mathcal{T})$ as defined in [32, p. 2.13]. This local interpolant has the following pleasant local properties

$$\|v - I_\tau v\|_K \leqslant C h_K |v|_{H^1(\omega_K)}, \quad \|v - I_\tau v\|_e \leqslant C h_{K_e}^{1/2} |v|_{H^1(\omega_{K_e})},$$

for all elements $K \in \mathcal{T}$ and edges $e \in \mathcal{E}^{int}$. Here K_e is an element adjacent to e, and ω_K is the union of elements touching K. Inserting $\tilde{v} = I_\tau v$ into the residual reveals

$$\begin{aligned}
a(u - U, v) &\leqslant \sum_{K \in \mathcal{T}} \|f + \Delta U\|_K \, C h_K |v|_{H^1(\omega_K)} + \sum_{e \in \mathcal{E}^{int}} \|[\![\nabla U]\!]\|_e \, C h_{K_e}^{1/2} |v|_{H^1(\omega_{K_e})} \\
&\leqslant C \left(\sum_{K \in \mathcal{T}} h_K^2 \|f + \Delta U\|_K^2 \right)^{1/2} \left(\sum_{K \in \mathcal{T}} |v|^2_{H^1(\omega_K)} \right)^{1/2} \\
&\quad + C \left(\sum_{e \in \mathcal{E}^{int}} h_K \|[\![\nabla U]\!]\|_e^2 \right)^{1/2} \left(\sum_{e \in \mathcal{E}^{int}} |v|^2_{H^1(\omega_{K_e})} \right)^{1/2} \\
&\leqslant \tilde{C} |v|_{H^1(\Omega)} \left(\sum_{K \in \mathcal{T}} h_K^2 \|f + \Delta U\|_K^2 + \sum_{e \in \mathcal{E}^{int}} h_K \|[\![\nabla U]\!]\|_e^2 \right)^{1/2},
\end{aligned}$$

with the second inequality following from Cauchy-Schwarz for sequences. Together with (1.7), this yields a desired error estimation in terms of localized errors:

$$|||u - U|||_\Omega \leqslant \tilde{C} \left(\sum_{K \in \mathcal{T}} h_K^2 \|f + \Delta U\|_K^2 + \sum_{e \in \mathcal{E}^{int}} h_K \|[\![U]\!]\|_e^2 \right)^{1/2}.$$

This gives rise to the following definitions.

Definition 1.8. For $K \in \mathcal{T}$ and $v \in \mathbb{V}(\mathcal{T})$ the residual *error indicator* for v on K reads

$$\eta^2(v, K) := h_K^2 \|f + \Delta v\|_K^2 + h_K \|[\![\nabla v]\!]\|^2_{\partial K \setminus \partial \Omega}.$$

The *oscillation term* of f on K is given by

$$\operatorname{osc}^2(f, K) := h_K^2 \|f - P_K^r f\|_K^2 \quad \text{for some } r \geqslant p - 2,$$

with P_K^r the $L^2(K)$-orthogonal projector on polynomials of degree r on K.

For a subset $\mathcal{M} \subset \mathcal{T}$, the above terms are defined as the sum over $K \in \mathcal{M}$, i.e.

$$\eta^2(v, \mathcal{M}) := \sum_{K \in \mathcal{M}} \eta^2(v, K), \quad \text{and} \quad \operatorname{osc}^2(f, \mathcal{M}) := \sum_{K \in \mathcal{M}} \operatorname{osc}^2(f, K).$$

Denote $\mathcal{T}_\star \geqslant \mathcal{T}$ if $\mathcal{T}_\star$ is a refinement of $\mathcal{T}$. That is, every element in $\mathcal{T}_\star$ is either contained in $\mathcal{T}$ or can be obtained by applying a finite number of bisections to an element in $\mathcal{T}$. Write $R_{\mathcal{T} \to \mathcal{T}_\star}$ for the set of refined elements, i.e. $R_{\mathcal{T} \to \mathcal{T}_\star} = \mathcal{T} \setminus \mathcal{T}_\star$. To distinguish between finite element solutions, denote $U_\star$ for the Galerkin approximation in $\mathbb{V}(\mathcal{T}_\star)$. The following theorem shows *reliability* and *efficiency* of the residual estimator.

Theorem 1.7 ([34, 13])**.** *There exists a constant C_1 such that for $\mathcal{T}_\star \geqslant \mathcal{T}$ we have reliability and discrete reliability:*

$$|||u - U|||_\Omega^2 \leqslant C_1 \eta^2(U, \mathcal{T}) \quad \textit{and} \quad |||U_\star - U|||_\Omega^2 \leqslant C_1 \eta^2(U, R_{\mathcal{T} \to \mathcal{T}_\star}).$$

Similarly, we have efficiency for a constant C_2:

$$\eta^2(U, \mathcal{T}) \leqslant C_2 \left[|||u - U|||_\Omega^2 + \operatorname{osc}^2(f, \mathcal{T}) \right].$$

Note that the oscillation is dominated by the error estimator, i.e. $\operatorname{osc}^2(f, K) \leqslant \eta^2(v, K)$. The above lemma therefore shows that the estimator $\eta^2(U, \mathcal{T})$ provides an upper- and lower bound for the *total error* $|||u - U|||_\Omega^2 + \operatorname{osc}^2(f, \mathcal{T})$, that is, the error estimator is proportional to the total error. In practice the $\operatorname{osc}^2(f, \mathcal{T})$ is often order of magnitudes smaller than $\eta^2(U, \mathcal{T})$; if this the case, the estimator $\eta^2(U, \mathcal{T})$ itself is proportional to the approximation error.

1.6. Adaptive finite element method

How can one use these local error estimators $\eta(U, K)$? Here we notice that $\eta(U, K)$ consists of the *known* quantities f and U. Therefore, one can actually compute the estimators $\eta(U, K)$ to find out *where* the approximation error is big. Instead of globally refining every element, one can just refine the elements for which $\eta(U, K)$ is large. This leads to the very intuitive algorithm, called the *adaptive finite element method*:

$$\texttt{SOLVE} \to \texttt{ESTIMATE} \to \texttt{MARK} \to \texttt{REFINE}.$$

Solve a discrete solution U, estimate local errors using $\eta(U, K)$, select a subset of elements for which the error is large and refine these elements. Later, in Chapter 3, a more detailed description of this method will be given.

Marking — making the selection of triangles to be refined — is done using Dörfler marking, named after its inventor Dörfler in [17]. That is, one selects the *minimal* cardinality subset $\mathcal{M} \subset \mathcal{T}$ such that $\eta(U, \mathcal{M}) \geqslant \theta \eta(U, \mathcal{T})$, for some marking parameter $\theta \in (0, 1)$. We let `REFINE` select the smallest conforming refinement $\mathcal{T}_\star \geqslant \mathcal{T}$ such that all

triangles in $\mathcal{M}$ are bisected. This can be accomplished using the *newest vertex bisection* algorithm [36, 11].

Write $\{\mathcal{T}_k, U_k, \mathcal{M}_k\}_{k\geqslant 0}$ for the sequence of results calculated by the adaptive method. Using the upper bound from Theorem 1.7, the Dörfler marking property, and Galerkin orthogonality, allows one to proof the following contraction property [17, 24, 13].

Theorem 1.8. *There exists constants $\gamma > 0$ and $\alpha \in (0,1)$, such that*

$$|||u - U_{k+1}|||_\Omega^2 + \gamma\eta^2\left(U_{k+1}, \mathcal{T}_{k+1}\right) \leqslant \alpha\left(|||u - U_k|||_\Omega^2 + \gamma\eta^2(U_k, \mathcal{T}_k)\right)$$

This shows that the *quasi-error* $|||u - U_k|||_\Omega^2 + \gamma\eta^2\left(U_k, \mathcal{T}_k\right)$ is reduced at every step. Since $\eta^2(U_k, \mathcal{T}_k)$ is proportional to the total error, and $|||u - U_k|||$ is a non-increasing sequence, this proves convergence of AFEM. One can even prove that this error decays the best possible rate. To formalise this, we introduce an approximation class.

Definition 1.9. For $s > 0$ define the approximation class $\mathcal{A}^s$ by

$$\mathcal{A}^s := \{u \in H_0^1(\Omega) : \Delta u \in L^2(\Omega),$$
$$|u|_{\mathcal{A}^s} := \sup_{N\in\mathbb{N}} (N+1)^s \min_{\{\mathcal{T}\in\mathbb{T}:\#\mathcal{T}-\#\mathcal{T}_0\leqslant N\}} \sqrt{|||u - U_\mathcal{T}|||_\Omega^2 + \operatorname{osc}^2(f,\mathcal{T})} < \infty\}.$$

Suppose that $u \in \mathcal{A}^s$. For $\widetilde{\mathcal{T}}_N$, the best partition in $N + \#\mathcal{T}_0$ triangles, the total error satisfies $\sqrt{|||u - U_N|||_\Omega^2 + \operatorname{osc}^2(f, \widetilde{\mathcal{T}}_N)} \leqslant (N+1)^{-s}|u|_{\mathcal{A}^s}$. The number of triangles is proportional to the number of degrees of freedom; it therefore provides a representative quantity for comparing the error decay.

We have the following celebrated optimality theorem. A proof of this theorem as well as a historic overview of contributions that lead to this result can be found in [34, 13].

Theorem 1.9. *Let C_1, C_2 be as in Theorem 1.7. Ensure that $\theta^2 < \left(C_1(C_2+1)\right)^{-1}$. Let $u \in \mathcal{A}^s$ for some $s > 0$, then there is a constant C independent of u such that*

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \leqslant C|u|_{\mathcal{A}^s}^{1/s}\left(\sqrt{|||u - U_k|||_\Omega^2 + \operatorname{osc}^2(\mathcal{T}_k, f)}\right)^{-1/s}.$$

In words, the sequence of triangulations produced by AFEM provide the same convergence rate — in terms of the degrees of freedom — as the best possible sequence of triangulations. In Chapter 3 we will prove something very similar, but for a different estimator. Details of the adaptive finite element method and of the optimality proof will then become clear.

2. Equilibrated flux estimator

In the previous chapter, the classical (or standard) residual error estimator is introduced as the driver for the adaptive finite element method (AFEM). This results in an optimal algorithm, i.e. the adaptive meshes generated by this method provide the highest possible convergence rate. Unfortunately, this asymptotic result is a bit inconvenient for practical purposes due to the unknown constants. In practice, one would like to know *when* to stop iterating, i.e. when the approximation error is small enough — say below some threshold.

Another problem of the classical residual error estimator is that it is not polynomial-degree-robust: the constants depend on p, the degree of the finite element solutions. This becomes an issue when considering *hp*-AFEM, a version of AFEM where the polynomial degree can also vary per element. For *hp*-AFEM, one would like an estimator with constants independent of the polynomial degree used on each of the elements.

Conveniently, there are error estimators that suffer less from these constants problems. Recently, quite some research interest has been shown for such constant-free estimators. One of these estimators is the *equilibrated residual error estimator*, also called the *equilibrated flux estimator* or the *Braess-Schöberl estimator* [10, 9, 19]. In this chapter, we will introduce this estimator — which we will refer to as the equilibrated flux estimator — and prove some of its properties.

For simplicity, we restrict ourself to the Poisson problem (1.6) on a two-dimensional polygonal domain $\Omega \subset \mathbb{R}^2$, with a right hand side $f \in L^2(\Omega)$. For some conforming triangulation $\mathcal{T}$ of the domain, we consider $\mathbb{V} := \mathbb{V}(\mathcal{T})$ — the (Lagrange) finite element space of degree $p \geqslant 1$. We assume each triangulation to be uniformly shape regular, i.e. $\sup_{K\in\mathcal{T}} h_K/p_K \leqslant \kappa$ for some shape regularity constant κ. The Galerkin approximation (discrete solution) is denoted by $U \in \mathbb{V}(\mathcal{T})$.

2.1. Prager and Synge

The so-called equilibrated flux estimators are based on the fundamental theorem of Prager and Synge [27], for which, we need the space $H(\mathrm{div};\Omega)$ as defined in B.1. Vector-valued functions will be underlined.

Theorem 2.1 (Prager and Synge)**.** *Let $u \in H^1_0(\Omega)$ be the exact solution of the Poisson problem with a right hand side $f \in L^2(\Omega)$. For a flux $\underline{\sigma} \in H(\mathrm{div};\Omega)$ satisfying the equilibrium condition $\mathrm{div}\,\underline{\sigma} + f = 0$ in L^2-sense, there holds*

$$\|\nabla u - \nabla v\|^2_\Omega + \|\nabla u - \underline{\sigma}\|^2_\Omega = \|\nabla v - \underline{\sigma}\|^2_\Omega \quad \forall v \in H^1_0(\Omega).$$

Proof. Application of the divergence theorem (B.2) yields

$$\begin{aligned}&\int_\Omega (\nabla u - \underline{\sigma}) \cdot \nabla (u - v) \,\mathrm{d}x \\ &= -\int_\Omega (u - v) \,\operatorname{div}(\nabla u - \underline{\sigma}) \,\mathrm{d}x + \int_{\partial\Omega} (u - v)\,(\nabla u \cdot n - \underline{\sigma} \cdot n) \,\mathrm{d}s = 0,\end{aligned}$$

since from the assumptions $\Delta u = -f = \operatorname{div} \underline{\sigma}$ in Ω, and $u - v = 0$ on $\partial\Omega$. From this orthogonality relation and Pythagoras' identity we may conclude that

$$\|\nabla(u - v)\|_\Omega^2 + \|\nabla u - \underline{\sigma}\|_\Omega^2 = \|-\nabla(u - v) + \nabla u - \underline{\sigma}\|_\Omega^2,$$

which equals the asserted. □

For a flux $\underline{\sigma}$ satisfying the equilibrium condition, we obtain an *reliable* constant-free estimator by replacing v with the discrete solution U in the previous theorem:

$$|||u - U|||_\Omega^2 = \|\nabla u - \nabla U\|_\Omega^2 \leqslant \|\nabla U - \underline{\sigma}\|_\Omega^2. \tag{2.1}$$

The question now arises how to construct $\underline{\sigma}$. In general one would like an estimator to be proportional to the error. For this, we need the estimator to also be *efficient*: it should provide a lower bound for $|||u - U|||_\Omega$, up to a constant and possibly an oscillation term. The following construction will provide such efficiency.

From now until §2.5, we suppose that f is a piecewise polynomial of degree at most $p - 1$ on $\mathcal{T}$, i.e.

$$f \in \mathcal{P}_{p-1}^{-1}(\mathcal{T}) := \left\{ f \in L^2(\Omega) : f|_K \in \mathcal{P}_{p-1}(K) \quad \forall K \in \mathcal{T} \right\},$$

and consider the p-th order *Raviart-Thomas* [28] space $\mathcal{RT}_p(\mathcal{T})$ defined by:

$$\begin{aligned}\mathcal{RT}_p(K) &:= \left[\mathcal{P}_p(K)\right]^2 + \mathcal{P}_p(K)\underline{x}, \\ \mathcal{RT}_p^{-1}(\mathcal{T}) &= \left\{ \underline{\sigma} \in \left[L^2(\Omega)\right]^2 : \underline{\sigma}|_K \in \mathcal{RT}_p(K) \quad \forall K \in \mathcal{T} \right\}, \\ \mathcal{RT}_p(\mathcal{T}) &:= H(\operatorname{div};\Omega) \cap \mathcal{RT}_p^{-1}(\mathcal{T}).\end{aligned}$$

Evidently fluxes $\underline{\sigma} \in \mathcal{RT}_p(K)$ satisfy $\operatorname{div} \underline{\sigma} \in \mathcal{P}_p^{-1}(\mathcal{T})$; the divergence mapping is even surjective [7, Prop 2.3.3], and thus $\mathcal{RT}_p(K)$ contains equilibrated fluxes. From (2.1) we then see that the sharpest estimator in the Raviart-Thomas space is found by minimizing $\|\nabla U - \sigma\|$ over all fluxes $\underline{\sigma} \in \mathcal{RT}_p(\mathcal{T})$ that are in equilibrium. However, this global minimization procedure — equivalent to the mixed finite element solution [8] — is too expensive for computation of an error estimate.

To overcome this problem, Braess and Schöberl [10] propose minimizing local problems instead, a procedure called *equilibration*.

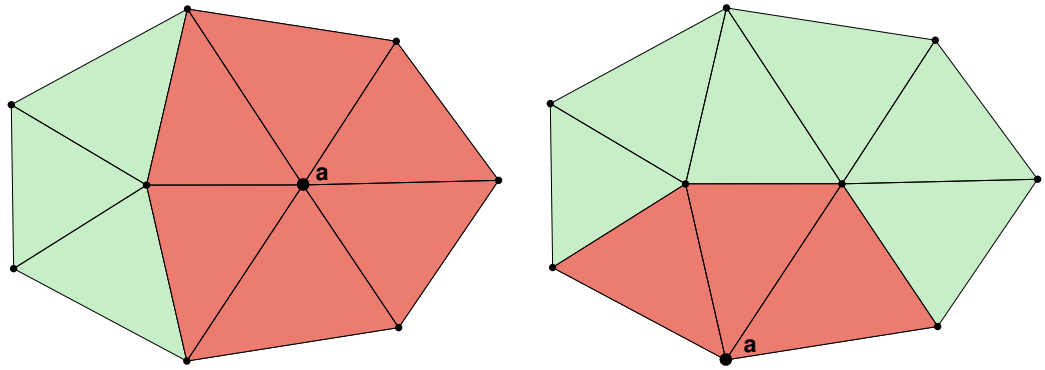


Figure 2.1.: Two different type of patches highlighted. The left image displays a patch ω_a associated to an interior vertex a. The right image displays a typical patch associated to a boundary vertex.

2.2. Equilibration

Instead of directly constructing the flux $\underline{\sigma}$, the difference $\underline{\sigma}^{\triangle} := \nabla U - \underline{\sigma}$ is considered. We assume that f belongs to the broken polynomial space $\mathcal{P}^{-1}_{p-1}(\mathcal{T})$ to avoid the effect of data oscillation. The flux $\underline{\sigma}$ will be constructed in $\mathcal{RT}_p(\mathcal{T})$; the correction $\underline{\sigma}^{\triangle}$ must therefore belong to the broken Raviart-Thomas space $\mathcal{RT}^{-1}_p(\mathcal{T})$.

Following [19], we use $\mathcal{V}, \mathcal{E}$ to denote the set of vertices and edges in the triangulation $\mathcal{T}$. Superscripts ^{int} or ^{bdr} are added to indicate restrictions of $\mathcal{V}, \mathcal{E}$ to the interior or the boundary. For each vertex $a \in \mathcal{V}$, we write ψ_a for the hat function at vertex a, i.e. the unique function in the linear finite element space that takes value 1 at a and vanishes at the other vertices. These hat functions form a partition of unity: $\sum_{a\in\mathcal{V}} \psi_a \equiv \mathbb{1}$. The local problems are solved on patches ω_a, given by the star at a vertex $a \in \mathcal{V}$, also being the support of the hat function ψ_a — see Figure 2.1 for an illustration. We denote γ_a for the union of interior edges touching a. So for a vertex $a \in \mathcal{V}$ we have

$$\omega_a := \omega(\mathcal{T}, a) = \bigcup\{K \in \mathcal{T} : a \in \partial K\}, \quad \gamma_a := \gamma(\mathcal{T}, a) = \bigcup\{e \in \mathcal{E}^{int} : a \in e\}.$$

Often we are interested in the set of elements $K \in \mathcal{T}$ that make up ω_a. We abuse the notation and intuitively write $K \subset \omega_a$ under sums as shorthand for $\{K \in \mathcal{T} : a \in \partial K\}$; similarly we write $e \subset \gamma_a$ to mean $\{e \in \mathcal{E}^{int} : a \in e\}$.

For each patch ω_a, we solve a local problem resulting in a local flux $\underline{\sigma_a}$. The total difference flux $\underline{\sigma}^{\triangle}$ is then taken as the sum of local fluxes over these patches,

$$\underline{\sigma}^{\triangle} := \sum_{a\in\mathcal{V}} \underline{\sigma_a}.$$

As proposed in [9], local fluxes are found by decomposing the residual using the partition of unity. Write $r \in H^1_0(\Omega)'$ for the usual residual defined in §1.5, i.e. for $v \in H^1_0(\Omega)$

$$\langle r, v\rangle := a(u - U, v) = \langle \nabla(u - U), \nabla v\rangle_\Omega = \langle f, v\rangle_\Omega - \langle \nabla U, \nabla v\rangle_\Omega.$$

For every vertex $a \in \mathcal{V}$, the local residual at a is defined as

$$\langle r_a, v\rangle := \langle r, \psi_a v\rangle = a(u - U, \psi_a v)_{\omega_a}.$$

The local flux $\underline{\sigma_a}$ is taken from the broken Raviart-Thomas space $\mathcal{RT}_p^{-1}(\omega_a)$ such that

$$\langle r_a, v\rangle = -\langle \underline{\sigma_a}, \nabla v\rangle_{\omega_a} \quad \forall v \in H^1(\omega_a) \cap H_0^1(\Omega). \tag{2.2}$$

Existence of such a flux $\underline{\sigma_a}$ will be shown at the end of this section. The residual can be expressed in triangle and edge related terms by applying the divergence theorem, cf. (1.8). This leads to the decomposition

$$\langle r_a, v\rangle = \langle f, \psi_a v\rangle_{\omega_a} - \langle \nabla U, \nabla(\psi_a v)\rangle_{\omega_a} = \sum_{K \subset \omega_a} \langle r_a^T, v\rangle_K + \sum_{e \subset \gamma_a} \langle r_a^e, v\rangle_e, \tag{2.3}$$

with L^2-inner products over elements K and (lower dimensional) edges e, and

$$r_a^T := \psi_a \left[f + \Delta U\right], \quad r_a^e := \psi_a [\![\nabla U]\!].$$

Similarly rewriting $\langle \underline{\sigma_a}, \nabla v\rangle_{\omega_a}$ in triangle and edge related terms yields

$$-\langle \underline{\sigma_a}, \nabla v\rangle_{\omega_a} = \sum_{K \subset \omega_a} \langle \operatorname{div} \underline{\sigma_a}, v\rangle_K + \sum_{e \subset \gamma_a} \langle [\![\underline{\sigma_a}]\!], v\rangle + \langle \underline{\sigma_a} \cdot n, v\rangle_{\partial\omega_a \setminus \partial\Omega}.$$

After expanding both sides, we see that (2.2) holds if and only if

$$\begin{aligned} \operatorname{div} \underline{\sigma_a} &= \psi_a \left[f + \Delta U\right] && \text{in } K \subset \omega_a, \\ [\![\underline{\sigma_a}]\!] &= \psi_a [\![\nabla U]\!] && \text{on } e \subset \gamma_a, \\ \underline{\sigma_a} \cdot n &= 0 && \text{on } \partial\omega_a \setminus \partial\Omega. \end{aligned} \tag{2.4}$$

To prove that system (2.4) has a solution we make the following observation. For an interior vertex $a \in \mathcal{V}^{int}$, the hat function ψ_a belongs to the finite element space $\mathbb{V}(\mathcal{T})$. Using that U is the Galerkin approximation therefore gives us $\langle r_a, \mathbb{1}\rangle = \langle f, \psi_a\rangle_\Omega - \langle \nabla U, \nabla \psi_a\rangle_\Omega = 0$. That is, the local residual vanishes on constant functions.

Theorem 2.2. *There exists a flux $\underline{\sigma_a} \in \mathcal{RT}_p^{-1}(\omega_a)$ that solves the above system* (2.4)*.*

Proof. We use a (well known) result for Raviart-Thomas elements proved in Lemma B.5, namely, for an element K there exists a function $\underline{\tau} \in \mathcal{RT}_p(K)$ such that

$$\operatorname{div} \underline{\tau} = P_K \in \mathcal{P}_p(K), \quad \underline{\tau} \cdot n = P_e \in \mathcal{P}_p(e) \quad \forall e \subset \partial K,$$

if the polynomials satisfy the compatibility constraint $\int_K P_K = \int_{\partial K} P_e$. Notice that the polynomials in (2.4) are of degree p, and are thus eligible in this theorem.

Fix an interior vertex $a \in \mathcal{V}^{int}$. In this case, ω_a is of the shape depicted in Figure 2.1. Let $\{K_1, \ldots, K_n\}$ be a numbering of the n distinct elements such that K_i and K_{i+1} share an edge, i.e. $K_i \cap K_{i+1} = e_i \in \gamma_a$ for $1 \leqslant i \leqslant n-1$. The first and last element also share an edge $K_1 \cap K_n = e_n$, because a is an interior vertex. The edges of K_i are thus

given by e_i, e_{i-1} and a (patch) boundary edge $\tilde{e}_i$. For K_1, we pick a $\underline{\sigma_1} \in \mathcal{RT}_p(K_1)$ that solves

$$\begin{aligned} \operatorname{div} \underline{\sigma_1} &= \psi_a \left[f + \Delta U \right] && \text{in } K_1 \subset \omega_a, \\ \underline{\sigma_1} \cdot n &= -\psi_a [\![\nabla U]\!] && \text{on } e_n \subset \gamma_a, \\ \underline{\sigma_1} \cdot n &= 0 && \text{on } \tilde{e}_1 \subset \partial\omega_a, \\ \underline{\sigma_1} \cdot n &= p_1 && \text{on } e_1 \subset \gamma_a, \end{aligned} \tag{2.5}$$

with p_1 a polynomial chosen such that the compatibility condition holds, i.e.

$$\int_{K_1} \psi_a \left[f + \Delta U \right] = \int_{e_n} -\psi_a [\![\nabla U]\!] + \int_{e_1} p_1.$$

Existence of $\underline{\sigma_1}$ then follows from Lemma B.5. Similarly, we solve $\underline{\sigma_2} \in \mathcal{RT}_p(K_2)$ from

$$\begin{aligned} \operatorname{div} \underline{\sigma_2} &= \psi_a \left[f + \Delta U \right] && \text{in } K_2 \subset \omega_a, \\ \underline{\sigma_2} \cdot n &= -\psi_a [\![\nabla U]\!] - p_1 && \text{on } e_1 \subset \gamma_a, \\ \underline{\sigma_2} \cdot n &= 0 && \text{on } \tilde{e}_2 \subset \partial\omega_a, \\ \underline{\sigma_2} \cdot n &= p_2 && \text{on } e_2 \subset \gamma_a, \end{aligned}$$

for some polynomial p_2 that ensures the compatibility condition. Per construction the trace jump of $\underline{\sigma_1}$ and $\underline{\sigma_2}$ over e_1 equals the required $\psi_a [\![\nabla U]\!]$.

We can repeat this process until we arrive at K_n. At this point we are no longer free to pick the 'next' edge polynomial p_n, because this is already determined by the first element in (2.5). To complete the system (2.4) we would like to solve $\underline{\sigma_n} \in \mathcal{RT}_p(K_n)$ from

$$\begin{aligned} \operatorname{div} \underline{\sigma_n} &= \psi_a \left[f + \Delta U \right] && \text{in } K_n \subset \omega_a, \\ \underline{\sigma_n} \cdot n &= -\psi_a [\![\nabla U]\!] - p_{n-1} && \text{on } e_{n-1} \subset \gamma_a, \\ \underline{\sigma_n} \cdot n &= 0 && \text{on } \tilde{e}_n \subset \partial\omega_a, \\ \underline{\sigma_n} \cdot n &= 0 && \text{on } e_n \subset \gamma_a. \end{aligned} \tag{2.6}$$

We will show that the compatibility condition is satisfied. From orthogonality on constant functions we deduce

$$\langle r_a, \mathbb{1} \rangle = 0 \implies \sum_{i=1}^{n} \int_{K_i} \psi_a \left[f + \Delta U \right] + \sum_{i=1}^{n} \int_{e_i} \psi_a [\![\nabla U]\!] = 0.$$

Combining this with the compatibility conditions of $\underline{\sigma_i}$ for $1 \leqslant i \leqslant n-1$ shows

$$\begin{aligned} \int_{K_n} \psi_a \left[f + \Delta U \right] &= -\sum_{i=1}^{n-1} \int_{K_i} \psi_a \left[f + \Delta U \right] - \sum_{i=1}^{n} \int_{e_i} \psi_a [\![\nabla U]\!] \\ &= -\int_{e_1} p_1 - \sum_{i=2}^{n-1} \int_{K_i} \psi_a \left[f + \Delta U \right] - \sum_{i=1}^{n-1} \int_{e_i} \psi_a [\![\nabla U]\!] \\ &= \cdots = -\int_{e_{n-1}} \psi_a [\![\nabla U]\!] - \int_{e_{n-1}} p_{n-1}. \end{aligned}$$

We see that the compatibility also holds for this last system (2.6), thereby proving existence of $\underline{\sigma_n}$. Defining $\underline{\sigma_a}|_{K_i} := \underline{\sigma_i}$ for $1 \leqslant i \leqslant n$ then provides us with a flux $\underline{\sigma_a} \in \mathcal{RT}^{-1}_{p,0}(\omega_a)$ that satisfies (2.4).

The process is similar for a boundary vertex $a \in \mathcal{V}^{bdr}$. We can again find a numbering $\{K_1, \ldots, K_n\}$ of the elements inside the patch ω_a such that $K_i \cap K_{i+1} = e_i$. This time the elements K_1 and K_n do not share a boundary, since a is a boundary vertex (cf. Figure 2.1). The system (2.4) does not prescribe conditions on edges of the domain boundary. We can therefore simply apply the method described above without having a conflict at the last element K_n. This process results in a $\underline{\sigma_a}$ that solves the system (2.4). □

From this proof it is clear that we can restrict ourselves to finding local flux $\underline{\sigma_a}$ in the subspace $\mathcal{RT}^{-1}_{p,0}(\omega_a)$ defined by

$$\mathcal{RT}^{-1}_{p,0}(\omega_a) := \begin{bmatrix} \left\{\underline{\sigma} \in \mathcal{RT}^{-1}_{p}(\omega_a) : \underline{\sigma} \cdot n = 0 \quad \text{on } \partial\omega_a\right\} & a \in \mathcal{V}^{int}, \\ \left\{\underline{\sigma} \in \mathcal{RT}^{-1}_{p}(\omega_a) : \underline{\sigma} \cdot n = 0 \quad \text{on } \partial\omega_a \setminus \partial\Omega\right\} & a \in \mathcal{V}^{bdr}. \end{bmatrix} \tag{2.7}$$

Since there is no unique solution that solves the system (2.4), a *minimal* L_2-norm solution is chosen. In summary, the locally equilibrated flux $\underline{\sigma_a}$ is given by

$$\underline{\sigma_a} \in \mathcal{RT}^{-1}_{p,0}(\omega_a) \quad \text{s.t. } \underline{\sigma_a} \text{ satisfies (2.2) and } \|\underline{\sigma_a}\|_{\omega_a} = \min_{\underline{\sigma} \in \mathcal{RT}^{-1}_{p,0}(\omega_a) : \underline{\sigma} \text{ satisfies (2.2)}} \|\underline{\sigma}\|_{\omega_a}.$$

These locally equilibrated fluxes are lifted to the entire space by taking $\underline{\sigma}^{\triangle} = \sum_{a \in \mathcal{V}} \underline{\sigma_a}$. Finally set $\underline{\sigma} = \nabla U - \underline{\sigma}^{\triangle}$, then for $v \in H^1_0(\Omega)$ we have

$$\begin{aligned} \langle \underline{\sigma}, \nabla v \rangle_\Omega &= \langle \nabla U, \nabla v \rangle_\Omega - \sum_{a \in \mathcal{V}} \langle \underline{\sigma_a}, \nabla v \rangle_{\omega_a} \\ &= \langle \nabla U, \nabla v \rangle_\Omega + \sum_{a \in \mathcal{V}} \langle r_a, v \rangle \\ &= \langle \nabla U, \nabla v \rangle_\Omega + \langle r, v \rangle = \langle f, v \rangle_\Omega. \end{aligned}$$

Since this holds for all $v \in H^1_0(\Omega)$, we know that $\underline{\sigma}$ has a weak divergence that is given by $\operatorname{div} \underline{\sigma} = -f$. Invoking Prager and Synge then gives the desired upper bound.

Theorem 2.3. *The above described process provides a reliable estimator, i.e.*

$$|||u - U|||_\Omega \leqslant \|\nabla U - \underline{\sigma}\|_\Omega = \|\underline{\sigma}^{\triangle}\|_\Omega = \left\| \sum_{a \in \mathcal{V}} \underline{\sigma_a} \right\|_\Omega.$$

2.3. Efficiency

The next (more involved) step is showing efficiency of the equilibrated flux estimator. We will derive patch-wise efficiency results, because of the localized nature of the estimator.

These results will be proven using the relation between $\underline{\sigma_a}$ and the local residual r_a. Recall that for $a \in \mathcal{V}^{int}$, the functional r_a vanishes on constant functions, and therefore we may interpret r_a as a functional on the mean zero space. Again in the notation of [19], we define for $a \in \mathcal{V}$,

$$H^1_\star(\omega_a) := \begin{cases} \left\{ v \in H^1(\omega_a) : \langle v, \mathbb{1} \rangle_{\omega_a} =: v_{\omega_a} = 0 \right\} & a \in \mathcal{V}^{int}, \\ \left\{ v \in H^1(\omega_a) : v = 0 \text{ on } \partial\omega_a \cap \partial\Omega \right\} & a \in \mathcal{V}^{bdr}. \end{cases}$$

We equip this space with the norm $\|\nabla \cdot\|_{\omega_a}$, which is a norm since

$$\|\nabla v\|_{\omega_a} = 0 \implies v = C_{st}, \quad v \in H^1_\star(\omega_a) \implies C_{st} = 0.$$

Even stronger, this norm is equivalent to the H^1-norm by Poincaré-Friedrichs inequality (cf. §B.2). With these definitions we are able to proof local efficiency in terms of r_a. To avoid unnecessary confusion, we will write $a \lesssim b$ if there holds $a \leqslant Cb$ for some constant C. Similarly, we write $a \gtrsim b$ if we have $b \lesssim a$, and $a \eqsim b$ if both inequalities hold.

Lemma 2.4. *For every vertex $a \in \mathcal{V}$, we have the following local efficiency on ω_a,*

$$\|r_a\|_{H^1_\star(\omega_a)'} \lesssim |||u - U|||_{\omega_a},$$

for a constant only dependent on the shape regularity of the triangulation — and thus independent of the polynomial-degree p.

Proof. By definition and the Cauchy-Schwarz inequality we find

$$\begin{aligned} \|r_a\|_{H^1_\star(\omega_a)'} &= \sup_{\{v \in H^1_\star(\omega_a) : \|\nabla v\|_{\omega_a} = 1\}} \left| \langle r_a, v \rangle \right| \\ &= \sup_{\{v \in H^1_\star(\omega_a) : \|\nabla v\|_{\omega_a} = 1\}} \left| \langle \nabla (u - U), \nabla (\psi_a v) \rangle_\Omega \right| \\ &\leqslant \|\nabla u - \nabla U\|_{\omega_a} \sup_{\{v \in H^1_\star(\omega_a) : \|\nabla v\|_{\omega_a} = 1\}} \|\nabla (\psi_a v)\|_{\omega_a}. \end{aligned}$$

In order to get rid of the ψ_a term we first apply the product rule,

$$\begin{aligned} \|\nabla (\psi_a v)\|_{\omega_a} &= \|v \nabla \psi_a + \psi_a \nabla v\|_{\omega_a} \\ &\leqslant \|v \nabla \psi_a\|_{\omega_a} + \|\psi_a \nabla v\|_{\omega_a} \\ &\leqslant \|v\|_{\omega_a} \|\nabla \psi_a\|_{\infty, \omega_a} + \|\psi_a\|_{\infty, \omega_a} \|\nabla v\|_{\omega_a}. \end{aligned}$$

For the ψ_a terms we note that $\|\psi_a\|_{\infty,\omega_a} = 1$, whereas $\nabla \psi_a$ is constant and bounded on each triangle by ρ_K, and thus $\|\nabla \psi_a\|_{\infty,\omega_a} \leqslant C h_{\omega_a}^{-1}$ for some constant only depending on the shape regularity κ. The $\|v\|_{\omega_a}$-term can be estimated using Poincaré-Friedrich inequalities (cf §B.2):

$$\|v\|_{\omega_a} = \begin{cases} \|v - v_{\omega_a}\| \leqslant C_{P,\omega_a} h_{\omega_a} \|\nabla v\|_{\omega_a} & a \in \mathcal{V}^{int}, \\ \|v\| \leqslant C_{F,\omega_a} h_{\omega_a} \|\nabla v\|_{\omega_a} & a \in V^{bdr}, \end{cases}$$

with the constants depending only on the shape regularity κ of the triangulation. Combining these inequalities and using that $\|\nabla v\|_{\omega_a} = 1$, we arrive at the asserted. $\square$

This efficiency bound can be related to the local flux $\underline{\sigma_a}$. Recall that $\underline{\sigma_a} \in \mathcal{RT}_{p,0}(\omega_a)$ was found to be the minimal norm function satisfying (2.2). The fact that this flux is of minimum norm enabled Braess et al. [9] to prove the following powerful theorem.

Theorem 2.5. *For $a \in \mathcal{V}$, let $\underline{\sigma_a}$ be found as described above. The following efficiency bound holds for a constant depending on the shape regularity, but is independent of the polynomial-degree p used in the finite element space $\mathbb{V}$:*

$$\|\underline{\sigma_a}\|_{\omega_a} \lesssim \|r_a\|_{H^1_\star(\omega_a)'},$$

Proof. The proof of this theorem is very involved. A constructive proof is given in [9, Theorem 7]. The theorem formulation given in [9] is only applicable to interior vertices $a \in \mathcal{V}^{int}$. By the nature of the proof, however, this result easily extends to boundary vertices $a \in \mathcal{V}^{bdr}$ as well (cf. Theorem 2.2). $\square$

Theorem 2.6. *Combining (local) efficiency and reliability proves that the equilibrated flux estimator is proportional to the approximation error, i.e.*

$$|||u - U|||^2_\Omega \eqsim \sum_{a\in\mathcal{V}} \|\underline{\sigma_a}\|^2_{\omega_a},$$

for a constant independent of the polynomial-degree p.

Proof. The reliability follows from Theorem 2.3: write $\mathcal{V}_K$ for the vertices of K, then

$$\begin{aligned}
|||u - U|||^2_\Omega \leqslant \|\underline{\sigma}^{\triangle}\|^2_\Omega &= \sum_{K\in\mathcal{T}} \left\| \sum_{a\in\mathcal{V}_K} \underline{\sigma_a} \right\|^2_K \\
&\leqslant \sum_{K\in\mathcal{T}} \left[\sum_{a\in\mathcal{V}_K} \|\underline{\sigma_a}\|_K \right]^2 \\
&\leqslant 3 \sum_{K\in\mathcal{T}} \sum_{a\in\mathcal{V}_K} \|\underline{\sigma_a}\|^2_K = 3 \sum_{a\in\mathcal{V}} \|\underline{\sigma_a}\|^2_{\omega_a}.
\end{aligned}$$

The second inequality follows from the fact that every triangle has three vertices and that $(x+y+z)^2 \leqslant 3x^2 + 3y^2 + 3z^2$ for positive variables.

Global efficiency follows from summarizing the local bound given in the Lemma 2.4 and the previous Theorem, i.e.

$$\sum_{a\in\mathcal{V}} \|\underline{\sigma_a}\|^2_{\omega_a} \lesssim \sum_{a\in\mathcal{V}} \|r_a\|^2_{H^1_\star(\omega_a)'} \lesssim \sum_{a\in\mathcal{V}} |||u - U|||^2_{\omega_a} = 3\,|||u - U|||^2_\Omega\,.$$

The last equality follows from the notion that every triangle is contained in exactly three patches. The hidden constants depend on the shape regularity, but are independent of the polynomial-degree p. $\square$

2.4. Equivalence with the classical residual estimator

The above theorem shows that the equilibrated flux estimator is proportional to the approximation error. When using this estimator in an adaptive finite element method, an obvious refine strategy would consist of refining those patches for which $\|\underline{\sigma_a}\|_{\omega_a}$ is large. Later we will prove that this is indeed an optimal choice. For this optimality proof we require discrete reliability and discrete efficiency (cf. Theorem 3.1).

Cascón and Nochetto [12] note that discrete reliability and efficiency follows from a patch-wise equivalence between the equilibrated flux estimator and the classical residual estimator from Definition 1.8:

$$\|\underline{\sigma_a}\|_{\omega_a} \eqsim h_{\omega_a}\|f + \Delta U\|_{\omega_a} + h_{\omega_a}^{1/2}\|[\![\nabla U]\!]\|_{\gamma_a}\,,$$

where ΔU is to be interpreted as the element-wise Laplacian of U. Unfortunately, neither of these claims — local equivalence and the discrete bounds — are proven in [12]. We will overcome this shortcoming by providing proofs, starting with the patch-wise equivalence. The inequalities that make up this equivalence are separately shown in the following two Lemmas.

Lemma 2.7. *The estimator $\|\underline{\sigma_a}\|_{\omega_a}$ is bounded by the classical estimator on the patch ω_a,*

$$\|\underline{\sigma_a}\|_{\omega_a} \lesssim h_{\omega_a}\|f + \Delta U\|_{\omega_a} + h_{\omega_a}^{1/2}\|[\![\nabla U]\!]\|_{\gamma_a}\,,$$

for a constant only depending on the shape regularity.

Proof. From Theorem 2.5 we know that $\|\underline{\sigma_a}\|_{\omega_a} \lesssim \|r_a\|_{H^1_\star(\omega_a)'}$. Let $v \in H^1_\star(\omega_a)$ and decompose the local residual r_a in element and edge related terms as in (2.3),

$$\begin{aligned}\langle r_a, v\rangle &= \sum_{K\subset\omega_a}\langle \psi_a\left[f+\Delta U\right], v\rangle_K + \sum_{e\subset\gamma_a}\langle \psi_a[\![\nabla U]\!], v\rangle_e \\ &\leqslant \sum_{K\subset\omega_a}\|\psi_a\|_{\infty,K}\|f+\Delta U\|_K\|v\|_K + \sum_{e\subset\gamma_a}\|\psi_a\|_{\infty,e}\|[\![\nabla U]\!]\|_e\|v\|_e \\ &\leqslant \sum_{K\subset\omega_a}\|f+\Delta U\|_K\|v\|_K + \sum_{e\subset\gamma_a}\|[\![\nabla U]\!]\|_e\|v\|_e\,.\end{aligned} \tag{2.8}$$

For the element-related terms, we invoke Cauchy-Schwarz, and use the Poincaré-Friedrichs inequality (cf. §B.2) to find

$$\begin{aligned}\sum_{K\subset\omega_a}\|f+\Delta U\|_K\|v\|_K &\leqslant \sqrt{\sum_{K\subset\omega_a}\|f+\Delta U\|_K^2}\sqrt{\sum_{K\subset\omega_a}\|v\|_K^2} \\ &= \|v\|_{\omega_a}\|f+\Delta U\|_{\omega_a} \\ &\leqslant h_{\omega_a}C_{PF,\omega_a}\|\nabla v\|_{\omega_a}\|f+\Delta U\|_{\omega_a}\,.\end{aligned}$$

Next, we tackle the edge terms. Let $e\subset\gamma_a$ be an edge with K_e an adjoint triangle. Using the trace theorem and transformation lemma [34, Lem 1.2], one can show that

$$\|v\|_e \leqslant \|v\|_{\partial K_e} \lesssim h_{K_e}^{-1/2}\|v\|_{K_e} + h_{K_e}^{1/2}\|\nabla v\|_{K_e}\,,$$

for a constant depending only on the shape regularity. Sum this inequality over all edges, where we pick K_e such that every edge $e \subset \gamma_a$ induces a different element:

$$\begin{aligned}\sum_{e\subset\gamma_a} \|v\|_e &\lesssim \sum_{e\subset\gamma_a} \left[h_{K_e}^{-1/2}\|v\|_{K_e} + h_{K_e}^{1/2}\|\nabla v\|_{K_e}\right] \\ &\lesssim \sqrt{\sum_{e\subset\gamma_a} h_{K_e}^{-1}}\sqrt{\sum_{e\subset\gamma_a}\|v\|_{K_e}^2} + \sqrt{\sum_{e\subset\gamma_a} h_{K_e}}\sqrt{\sum_{e\subset\gamma_a}\|\nabla v\|_{K_e}^2} \\ &\lesssim h_{\omega_a}^{-1/2}\|v\|_{\omega_a} + h_{\omega_a}^{1/2}\|\nabla v\|_{\omega_a} \\ &\lesssim h_{\omega_a}^{1/2}\|\nabla v\|_{\omega_a}\,.\end{aligned}$$

Here we used that maximum number of triangles in a patch is universally bounded[1], and that $v \in H^1_\star(\omega_a)$ to apply Poincaré-Friedrichs inequality. For the edge terms in (2.8) we now have

$$\sum_{e\subset\gamma_a} \|[\![\nabla U]\!]\|_e\|v\|_e \leqslant \|[\![\nabla U]\!]\|_{\gamma_a} \sum_{e\subset\gamma_a}\|v\|_e \lesssim h_{\omega_a}^{1/2}\|[\![\nabla U]\!]\|_{\gamma_a}\|\nabla v\|_{\omega_a}\,.$$

Combined these inequalities yield the asserted upper bound. Carefully examining the constants show that they only depend on shape regularity. □

Lemma 2.8. *The estimator $\|\underline{\sigma_a}\|_{\omega_a}$ is also bounded from below by the classical estimator,*

$$\|\underline{\sigma_a}\|_{\omega_a} \gtrsim h_{\omega_a}\|f + \Delta U\|_{\omega_a} + h_{\omega_a}^{1/2}\|[\![\nabla U]\!]\|_{\gamma_a}\,,$$

for a constant depending on the shape regularity and the polynomial-degree p used in $\mathbb{V}$.

Proof. In contrast to [12], this inequality will not be needed to prove discrete reliability. The proof is therefore not included here, but can be found in §B.4. □

Corollary. *The estimator $\|\underline{\sigma_a}\|_{\omega_a}$ is equivalent to the classical estimator on the patch ω_a,*

$$\|\underline{\sigma_a}\|_{\omega_a} \eqsim h_{\omega_a}\|f + \Delta U\|_{\omega_a} + h_{\omega_a}^{1/2}\|[\![\nabla U]\!]\|_{\gamma_a}\,,$$

for a constant depending on the shape regularity and the polynomial-degree p used in $\mathbb{V}$.

2.5. Discretization and oscillation

Thus far we have assumed the right hand side f to be a broken polynomial of degree at most $p-1$. In practice one often has a more general $f \in L^2(\Omega)$, and thus we must alter the above method. A straightforward approach is to replace the exact f with a polynomial approximation. For an element K, let Π_p denote the $L^2(K)$-orthogonal projector onto polynomials of degree p on K. Similarly, write Π_p^a for the $L^2(\omega_a)$-orthogonal projection

[1] If the number of triangles would be unbounded, then some triangles must have arbitrary small angles. The latter condition can not hold for a family of uniformly shape regular elements.

on the broken polynomial space $\mathcal{P}_p^{-1}(\omega_a)$. We will replace the (exact) local residuals r_a by discretized local residuals $\tilde{r}_a \in H_0^1(\Omega)'$, defined by

$$\begin{aligned}\langle \tilde{r}_a, v\rangle &:= \langle \Pi_p^a(\psi_a f), v\rangle_{\omega_a} - \langle \nabla U, \nabla(\psi_a v)\rangle_{\omega_a} \\ &= \sum_{K\subset\omega_a} \langle \Pi_p\left(\psi_a\left[f + \Delta U\right]\right), v\rangle_K + \sum_{e\subset\gamma_a} \langle \psi_a [\![\nabla U]\!], v\rangle_e.\end{aligned}$$

Since $\mathbb{1} \in \mathcal{P}_p^{-1}(\omega_a)$ and Π_p^a is an *orthogonal* projector we find

$$\langle \Pi_p^a(\psi_a f), \mathbb{1}\rangle_{\omega_a} = \langle \psi_a f, \Pi_p^a(\mathbb{1})\rangle_{\omega_a} = \langle \psi_a f, \mathbb{1}\rangle.$$

From Galerkin orthogonality it then follows that $\tilde{r}_a$ also vanishes on constant functions for interior vertices $a \in \mathcal{V}^{int}$. We can therefore interpret $\tilde{r}_a$ as a functional on the space $H_\star^1(\omega_a)$. Hereafter, $\underline{\sigma_a}$ will be defined in terms of the discretized local residual.

Definition 2.1 (Equilibrated flux)**.** For each vertex $a \in \mathcal{V}$, the equilibrated flux is defined as the minimal L^2-norm flux $\underline{\sigma_a} \in \mathcal{RT}_{p,0}^{-1}(\omega_a)$ that satisfies

$$\langle \tilde{r}_a, v\rangle = -\langle \underline{\sigma_a}, \nabla v\rangle_{\omega_a} \quad \forall v \in H^1(\omega_a) \cap H_0^1(\Omega). \tag{2.9}$$

The proof that such a flux $\underline{\sigma_a}$ exists is identical to the proof of Theorem 2.2, because $\tilde{r}_a$ also vanishes on constant functions for interior vertices $a \in \mathcal{V}^{int}$.

2.5.1. Reliability and efficiency

Of course, this discretization comes at a price. The constant-free error estimate provided by Prager and Synge in Theorem 2.3 has the following discretized counterpart.

Theorem 2.9 (Reliability)**.** *Let $\underline{\sigma_a}$ be as defined above and take $\underline{\sigma}^\triangle = \sum_{a\in\mathcal{V}} \underline{\sigma_a}$, then*

$$\begin{aligned}|\!|\!| u - U |\!|\!|_\Omega^2 &\leqslant \sum_{K\in\mathcal{T}} \left[\|\underline{\sigma}^\triangle\|_K + \frac{h_K}{\pi} \|(I - \Pi_p)(f)\|_K \right]^2 \\ &\lesssim \sum_{a\in\mathcal{V}} \left[\|\underline{\sigma_a}\|_{\omega_a}^2 + \sum_{K\subset\omega_a} h_K^2 \|(I - \Pi_{p-1})(f)\|_K^2 \right].\end{aligned}$$

Proof. Decompose both r_a and $\tilde{r}_a$ in triangle and edge terms. Since $\Pi_p(\psi_a \Delta U) = \psi_a \Delta U$ one easily sees that most terms in the difference $r_a - \tilde{r}_a$ cancel:

$$\langle r_a - \tilde{r}_a, v\rangle = \sum_{K\subset\omega_a} \langle (I - \Pi_p)(\psi_a f), v\rangle_K = \langle (I - \Pi_p^a)(\psi_a f), v\rangle_{\omega_a} \quad \forall v \in H^1(\omega_a) \cap H_0^1(\Omega).$$

For $v \in H_0^1(\Omega)$, the residual r in terms of the localized residuals r_a and $\tilde{r}_a$ is given by

$$\begin{aligned}\langle r, v\rangle &= \sum_{a\in\mathcal{V}} \langle r_a, v\rangle = \sum_{a\in\mathcal{V}} \langle \tilde{r}_a, v\rangle + \langle r_a - \tilde{r}_a, v\rangle \\ &= \sum_{a\in\mathcal{V}} -\langle \underline{\sigma_a}, \nabla v\rangle_{\omega_a} + \sum_{a\in\mathcal{V}} \sum_{K\subset\omega_a} \langle (I - \Pi_p)(\psi_a f), v\rangle_K \\ &= \sum_{K\in\mathcal{T}} \left[-\langle \underline{\sigma}^\triangle, \nabla v\rangle_K + \langle (I - \Pi_p)(f), v\rangle_K \right].\end{aligned}$$

The last equality holds since ψ_a forms a partition of unity on each element K when summed over its vertices $a \in \mathcal{V}_K$. The projector $(I - \Pi_p)$ is orthogonal to constant functions, so we may replace v by its mean zero variant $v - v_K$. Doing so and invoking Cauchy-Schwarz yields

$$\begin{aligned}
\langle r, v\rangle &\leqslant \sum_{K\in\mathcal{T}} \Big[\|\underline{\sigma}^{\triangle}\|_K \|\nabla v\|_K + \|(I-\Pi_p)(f)\|_K \|v - v_K\|_K\Big] \\
&\leqslant \sum_{K\in\mathcal{T}} \left[\|\underline{\sigma}^{\triangle}\|_K + \frac{h_K}{\pi}\|(I-\Pi_p)(f)\|_K\right] \|\nabla v\|_K \\
&\leqslant \sqrt{\sum_{K\in\mathcal{T}} \|\nabla v\|_K^2} \sqrt{\sum_{K\in\mathcal{T}} \left[\|\underline{\sigma}^{\triangle}\|_K + \frac{h_K}{\pi}\|(I-\Pi_p)(f)\|_K\right]^2}.
\end{aligned}$$

The second inequality follows from the Poincaré inequality for the convex domain K with Poincaré constant $C_{P,K} = \pi^{-1}$ — see §B.2. The first inequality in the theorem follows from the above in combination with $|||u - U|||_\Omega = \sup_{\{v\in H^1_0(\Omega)\,:\,\|\nabla v\|_\Omega = 1\}} \langle r, v\rangle$.

For the second inequality, we reintroduce the partition of unity into the above result, similar to what we did in the proof of Theorem 2.6:

$$\begin{aligned}
|||u - U|||_\Omega^2 &\lesssim \sum_{K\in\mathcal{T}} \Big[\|\underline{\sigma}^{\triangle}\|_K^2 + h_K^2 \|(I-\Pi_p)(f)\|_K^2\Big] \\
&\lesssim \sum_{K\in\mathcal{T}} \sum_{a\in\mathcal{V}_K} \Big[\|\underline{\sigma_a}\|_K^2 + h_K^2 \|(I-\Pi_p)(\psi_a f)\|_K^2\Big] \\
&\leqslant \sum_{a\in\mathcal{V}} \|\underline{\sigma_a}\|_{\omega_a}^2 + \sum_{a\in\mathcal{V}} h_{\omega_a}^2 \|(I-\Pi_p^a)(\psi_a f)\|_{\omega_a}^2.
\end{aligned}$$

The oscillation term $\|(I-\Pi_p^a)(\psi_a f)\|_{\omega_a}$ is a bit inconvenient to work with due to the presence of the hat function ψ_a. This can be circumvented by noting that

$$\begin{aligned}
\|(I-\Pi_p^a)(\psi_a f)\|_{\omega_a} &= \inf_{f_p\in\mathcal{P}_p^{-1}(\omega_a)} \|\psi_a f - f_p\|_{\omega_a} \\
&\leqslant \inf_{f_{p-1}\in\mathcal{P}_{p-1}^{-1}(\omega_a)} \|\psi_a f - \psi_a f_{p-1}\|_{\omega_a} \\
&\leqslant \|\psi_a\|_{\infty,\omega_a} \inf_{f_{p-1}\in\mathcal{P}_{p-1}^{-1}(\omega_a)} \|f - f_{p-1}\|_{\omega_a} \\
&= \|(I-\Pi_{p-1}^a)(f)\|_{\omega_a}.
\end{aligned}$$

From the equivalence $h_{\omega_a} \eqsim h_K$ for elements K in ω_a we infer that

$$h_{\omega_a}^2 \|(I-\Pi_p^a)(\psi_a f)\|_{\omega_a}^2 \leqslant h_{\omega_a}^2 \|(I-\Pi_{p-1}^a)(f)\|_{\omega_a}^2 \eqsim \sum_{K\subset\omega_a} h_K^2 \|(I-\Pi_{p-1})(f)\|_K^2.$$

Combining everything yields the asserted reliability bounds. □

Notice that the oscillation in the first reliability bound is of one order higher than the second. For notational ease, we will define h_a as the piecewise constant function on ω_a

with $h_a|_K := h_K$ for $K \subset \omega_a$. This shortens the notation of data oscillation, i.e.

$$\sum_{K\subset\omega_a} h_K^2 \|(I-\Pi_{p-1})(f)\|_K^2 = \|h_a(I-\Pi_{p-1}^a)(f)\|_{\omega_a}.$$

The discretized residual will also introduce an oscillation term in the discretized analogues of the efficiency results from §2.3, see the following theorem.

Theorem 2.10 (Efficiency)**.** *For the equilibrated flux $\underline{\sigma_a}$ we have patch-wise efficiency with*

$$\|\underline{\sigma_a}\|_{\omega_a} \lesssim \|\tilde{r}_a\|_{H^1_\star(\omega_a)'} \lesssim |\!|\!| u-U |\!|\!|_{\omega_a} + \|h_a(I-\Pi_{p-1}^a)(f)\|_{\omega_a},$$

and element-wise efficiency with

$$\|\underline{\sigma}^{\triangle}\|_K \lesssim \sum_{a\in\mathcal{V}_K} \left[|\!|\!| u-U |\!|\!|_{\omega_a} + \|h_a(I-\Pi_{p-1}^a)(f)\|_{\omega_a} \right].$$

Global efficiency is then given by

$$\sum_{K\in\mathcal{T}} \|\underline{\sigma}^{\triangle}\|_K^2 \lesssim \sum_{a\in\mathcal{V}} \|\underline{\sigma_a}\|_{\omega_a}^2 \lesssim |\!|\!| u-U |\!|\!|_\Omega^2 + \sum_{a\in\mathcal{V}} \|h_a(I-\Pi_{p-1}^a)(f)\|_{\omega_a}^2.$$

The constants only depend on the shape regularity of the triangulation (and are thus independent of the polynomial-degree p used in $\mathbb{V}$).

Proof. The functionals r_a and $\tilde{r}_a$ are both defined on the space $H^1_\star(\omega_a)$. An estimate for the dual norm of $r_a - \tilde{r}_a$ is given by

$$\begin{aligned}
\|r_a-\tilde{r}_a\|_{H^1_\star(\omega_a)'} &= \sup_{\{v\in H^1_\star(\omega_a): v\neq 0\}} \frac{\langle r_a, v\rangle - \langle \tilde{r}_a, v\rangle}{\|\nabla v\|_{\omega_a}} \\
&= \sup_{\{v\in H^1_\star(\omega_a): v\neq 0\}} \frac{\langle (I-\Pi_p^a)(\psi_a f), v\rangle_{\omega_a}}{\|\nabla v\|_{\omega_a}} \\
&\leqslant \sup_{\{v\in H^1_\star(\omega_a): v\neq 0\}} \frac{\|v\|_{\omega_a}\,\|(I-\Pi_p^a)(\psi_a f)\|_{\omega_a}}{\|\nabla v\|_{\omega_a}} \\
&\lesssim h_{\omega_a}\|(I-\Pi_p^a)(\psi_a f)\|_{\omega_a} \lesssim \|h_a(I-\Pi_{p-1}^a)(f)\|_{\omega_a},
\end{aligned}$$

where the second last inequality follows (again) from Poincaré-Friedrichs inequality. From Lemma 2.4 and by the triangle inequality we now find

$$\|\tilde{r}_a\|_{H^1_\star(\omega_a)'} \leqslant \|r_a\|_{H^1_\star(\omega_a)'} + \|r_a-\tilde{r}_a\|_{H^1_\star(\omega_a)'} \lesssim |\!|\!| u-U |\!|\!|_{\omega_a} + \|h_a(I-\Pi_{p-1}^a)(f)\|_{\omega_a}.$$

The asserted patch-wise efficiency follows by invoking the (powerful) theorem from Braess et. al [9] which shows that $\|\underline{\sigma_a}\|_{\omega_a} \lesssim \|\tilde{r}_a\|_{H^1_\star(\omega_a)'}$, for a constant independent of the polynomial degree p.

The element-wise efficiency follows directly from the above result and the triangle inequality, i.e.

$$\|\underline{\sigma}^{\triangle}\|_K \leqslant \sum_{a\in\mathcal{V}_K} \|\underline{\sigma_a}\|_K \leqslant \sum_{a\in\mathcal{V}_K} \|\underline{\sigma_a}\|_{\omega_a}.$$

Global efficiency follows easily from this local result:

$$\begin{aligned}\sum_{K\in\mathcal{T}} \|\underline{\sigma}^{\triangle}\|_K^2 &\lesssim \sum_{a\in\mathcal{V}} \|\underline{\sigma_a}\|_{\omega_a}^2 \\ &\lesssim \sum_{a\in\mathcal{V}} |||u-U|||_{\omega_a}^2 + \|h_a(I-\Pi_{p-1}^a)(f)\|_{\omega_a}^2 \\ &= 3\,|||u-U|||_{\Omega}^2 + \sum_{a\in\mathcal{V}} \|h_a(I-\Pi_{p-1}^a)(f)\|_{\omega_a}^2.\end{aligned}$$

The last equality holds since every triangle is contained in exactly three patches. □

These last two theorems prove that the discretized estimators $\sum_{K\in\mathcal{T}} \|\underline{\sigma}^{\triangle}\|_K^2$ and $\sum_{a\in\mathcal{V}} \|\underline{\sigma_a}\|_{\omega_a}^2$ are proportional to the squared approximation error, up to data oscillation. Of course, an oscillation term will also appear in the local equivalence with the standard residual estimator. The effects are summarized in the following lemma. The proof follows easily by mimicking the proof of Lemma 2.4, using the results from the previous theorem.

Lemma 2.11. *The estimator $\|\underline{\sigma_a}\|_{\omega_a}$ is locally equivalent to the standard residual estimator up to oscillation terms. To be precise,*

$$\begin{aligned}\|\underline{\sigma_a}\| + \|h_a(I-\Pi_{p-1}^a)(f)\|_{\omega_a} &\gtrsim h_{\omega_a}\|f+\Delta U\|_{\omega_a} + h_{\omega_a}^{1/2}\|[\![\nabla U]\!]\|_{\gamma_a}, \\ \|\underline{\sigma_a}\| &\lesssim \|h_a(I-\Pi_{p-1}^a)(f)\|_{\omega_a} + h_{\omega_a}\|f+\Delta U\|_{\omega_a} + h_{\omega_a}^{1/2}\|[\![\nabla U]\!]\|_{\gamma_a}.\end{aligned}$$

2.5.2. Lower order discretization

The above method uses the p-th order Raviart-Thomas space $\mathcal{RT}_{p,0}^{-1}(\omega_a)$ for construction of the equilibrated fluxes $\underline{\sigma_a}$. However, for implementational ease one might want to work with a lower order Raviart-Thomas space. A reduction to the Raviart-Thomas space of order $p-1$ can be accomplished by using another local residual $\hat{r}_a$, that is defined by

$$\langle \hat{r}_a, v\rangle := \sum_{K\subset\omega_a} \langle \Pi_{p-1}\left(\psi_a[f+\Delta U]\right), v\rangle_K + \sum_{e\subset\gamma_a} \langle \Pi_{p-1}\left(\psi_a[\![\nabla U]\!]\right), v\rangle_e,$$

where the second projector Π_{p-1} is the L^2-orthogonal projector on the polynomial edge space $\mathcal{P}_{p-1}(e)$. This local residual also vanishes on constant functions. The corresponding equilibrated flux $\underline{\hat{\sigma}_a}$ is the minimal L^2-norm flux in $\mathcal{RT}_{p-1,0}^{-1}(\omega_a)$ that satisfies

$$-\langle \hat{r}_a, v\rangle = -\langle \underline{\hat{\sigma}_a}, \nabla v\rangle_{\omega_a} \quad \forall v\in H^1(\omega_a)\cap H_0^1(\Omega).$$

This flux $\underline{\hat{\sigma}_a}$ provides a reliability bound, similar to one in Theorem 2.9. Indeed, decomposing the residual r in terms of r_a and $\hat{r}_a$ yields

$$\begin{aligned}\langle r, v\rangle &= \sum_{a\in\mathcal{V}} \langle \hat{r}_a, v\rangle + \sum_{a\in\mathcal{V}} \langle r_a - \hat{r}_a, v\rangle \\ &= \sum_{a\in\mathcal{V}} -\langle \underline{\hat{\sigma}_a}, \nabla v\rangle_{\omega_a} + \sum_{a\in\mathcal{V}} \langle (I-\Pi_{p-1}^a)(\psi_a f), v\rangle_{\omega_a} + \sum_{a\in\mathcal{V}}\sum_{e\subset\gamma_a} \langle (I-\Pi_{p-1})(\psi_a[\![\nabla U]\!]), v\rangle_e,\end{aligned}$$

where we used that $\psi_a \Delta U$ is a polynomial of degree $p-1$, and thus in the range of Π^a_{p-1}. Using that ψ_a forms a partition of unity on every edge e, and that $[\![\nabla U]\!]$ is a polynomial of degree $p-1$, allow us to write

$$\sum_{a\in\mathcal{V}}\sum_{e\subset\gamma_a}\langle (I-\Pi_{p-1})(\psi_a[\![\nabla U]\!]),v\rangle_e = \sum_{e\in\mathcal{E}^{int}}\langle (I-\Pi_{p-1})(\sum_{a\in\mathcal{V}}\psi_a[\![\nabla U]\!]),v\rangle_e = 0.$$

The edge terms vanish, thus following the proof of Theorem 2.9 reveals[2]

$$\begin{aligned}|\!|\!| u-U|\!|\!|^2_\Omega &\leqslant \sum_{K\in\mathcal{T}}\left[\|\hat{\underline{\sigma}}^{\triangle}\|_K + \frac{h_K}{\pi}\|(I-\Pi_{p-1})(f)\|_K\right]^2 \qquad \text{for } \hat{\underline{\sigma}}^{\triangle} = \sum_{a\in\mathcal{V}}\hat{\underline{\sigma}}_a \\ &\lesssim \sum_{a\in\mathcal{V}}\|\hat{\underline{\sigma}}_a\|^2_{\omega_a} + \sum_{a\in\mathcal{V}}\|h_a(I-\Pi^a_{p-2})(f)\|^2_{\omega_a}.\end{aligned}$$

The oscillation term appearing in these reliability bounds are of one order lower than the ones we had in Theorem 2.9. This is the cost of using a lower order estimator.

Similar techniques can be used to derive an efficiency bound for $\hat{\underline{\sigma}}_a$: following the proof of Theorem 2.10 and using that $\|v\|_{\gamma_a} \lesssim h_{\omega_a}\|\nabla v\|_{\omega_a}$, gives

$$\begin{aligned}\|\hat{\underline{\sigma}}_a\|_{\omega_a} &\lesssim \|r_a\|_{H^1_\star(\omega_a)'} + \|r_a - \hat{r}_a\|_{H^1_\star(\omega_a)'} \\ &\lesssim |\!|\!| u-U|\!|\!|_{\omega_a} + h_{\omega_a}\|(I-\Pi^a_{p-1})(\psi_a f)\|_{\omega_a} + h^{1/2}_{\omega_a}\|(I-\Pi^{a,e}_{p-1})(\psi_a[\![\nabla U]\!])\|_{\gamma_a} \\ &\leqslant |\!|\!| u-U|\!|\!|_{\omega_a} + \|h_a(I-\Pi^a_{p-2})(f)\|_{\omega_a} + h^{1/2}_{\omega_a}\|[\![\nabla U]\!]\|_{\gamma_a}.\end{aligned}$$

With $\Pi^{a,e}_{p-1}$ the L^2-orthogonal projector on the broken polynomial space $\mathcal{P}^{-1}_{p-1}(\gamma_a)$. For the jump term we can invoke local efficiency of the classical edge estimator (cf. [39, p. 8]):

$$\begin{aligned}h_{\omega_a}\|[\![\nabla U]\!]\|^2_{\gamma_a} &\eqsim \sum_{e\subset\gamma_a} h_e\|[\![\nabla U]\!]\|^2_e \\ &\lesssim \sum_{e\subset\gamma_a}\sum_{K\in\mathcal{T}_e}\left[|\!|\!| u-U|\!|\!|^2_K + h^2_K\|(I-\Pi_{p-1})(f)\|^2_K\right] \\ &\eqsim |\!|\!| u-U|\!|\!|^2_{\omega_a} + \|h_a(I-\Pi_{p-1})(f)\|^2_{\omega_a},\end{aligned}$$

with $\mathcal{T}_e$ the two triangles $K\in\mathcal{T}$ adjacent to the (interior) edge $e\subset\gamma_a$. This oscillation term is of higher degree than we already had, local efficiency of $\|\hat{\underline{\sigma}}_a\|_{\omega_a}$ therefore becomes

$$\|\hat{\underline{\sigma}}_a\|_{\omega_a} \lesssim |\!|\!| u-U|\!|\!|_{\omega_a} + \|h_a(I-\Pi^a_{p-2})(f)\|_{\omega_a}.$$

Remark. This method provides reliability and efficiency with the gain of being able to calculate the flux in the lower order $\mathcal{RT}^{-1}_{p-1,0}(\omega_a)$. We could repeat this approach to calculate estimators in even lower order Raviart-Thomas spaces. Unfortunately, this will yield bounds for which the decay rate of data oscillation is no longer of higher order. To see this, suppose that $|\!|\!| u-U|\!|\!|_\Omega$ converges as h^p. The data oscillation term in the tightest upper bound for $\hat{\underline{\sigma}}_a$ is given by $\frac{h_K}{\pi}\|f-\Pi_{p-1}f\|_K$. For f piecewise smooth, this term converges as h^{p+1}: one order faster than the finite element solution. If instead one calculates flux estimators in the Raviart-Thomas space $p-2$, then the data oscillation term will converge as h^p and is no longer of higher order.

[2]This last estimate does not hold for $p=1$, one then obtains the oscillation term $\left\|(I-\Pi^a_0)(\psi_a f)\right\|_{\omega_a}$.

3. Optimality of the adaptive finite element method

We will now focus on an adaptive finite element method that utilizes the equilibrated flux estimator. Recall from 1.5 that AFEM can be described by the following loop,

$$\texttt{SOLVE} \to \texttt{ESTIMATE} \to \texttt{MARK} \to \texttt{REFINE}.$$

In words: calculate the Galerkin solution for a triangulation, estimate the error using an a posteriori error estimator, mark elements that need to be refined, refine the triangulation, and start over.

After specifying the details of these four modules, we will be able to show that the produced sequence of discrete solutions converges with an optimal rate. That is, the produced approximations converge at the same rate as the discrete solutions found on the best possible triangulations.

This optimality result is similar to the classical (or standard) one in §1.6, except that we will now use the equilibrated flux estimator. We will provide a simplified form of the optimality proof given by Cascón and Nochetto [12]. A different optimality proof is given by Kreuzer and Siebert [22]; therein slightly different versions of `MARK` and `REFINE` are used.

3.1. Design of the adaptive finite element method

First we need some additional notation. Given a triangulation $\mathcal{T}$, we denote $\mathcal{W}$ for the set of star patches, i.e.

$$\mathcal{W} := \{\omega_a : a \in \mathcal{V}\}.$$

For $a \in \mathcal{V}$, let $\underline{\sigma_a}$ be the equilibrated local flux corresponding to the discrete solution $U \in \mathbb{V}(\mathcal{T})$. We denote the error estimator, oscillation, and total error resp. by:

$$\begin{aligned}
\eta_{\mathcal{T}}^2(U,\omega_a) &:= \|\underline{\sigma_a}\|_{\omega_a}^2, \\
\operatorname{osc}_{\mathcal{T}}^2(f,\omega_a) &:= \|h_a(I-\Pi_{p-1}^a)(f)\|_{\omega_a}^2 = \sum_{K\subset\omega_a} h_K^2\|(I-\Pi_{p-1})(f)\|_K^2, \\
\vartheta_{\mathcal{T}}^2(U,\omega_a) &:= \eta_{\mathcal{T}}^2(U,\omega_a) + \operatorname{osc}_{\mathcal{T}}^2(f,\omega_a).
\end{aligned}$$

The above quantities extend to sets of patches in the usual way — just take the sum over the patches in the set. We stress that any patch $\omega \in \mathcal{W}$ is implicitly associated to a vertex $a \in \mathcal{V}$.

Remark. In contrast to the *element-wise* classical residual estimator, we use the *patch-wise* equilibrated flux estimator $\|\underline{\sigma_a}\|_{\omega_a}$ here. This is a natural choice because of the localized patch-wise properties of the equilibrated flux estimator. Moreover, we require the patch-wise equivalence with the classical residual estimator for the optimality proof of AFEM.

On the other hand, from §2.5.1 we know that $\|\underline{\sigma}^{\triangle}\|_K$ is a reliable and efficient element-wise estimator — without an unknown constant in the reliability bound. This element-wise variant therefore is therefore more convenient in practice. We will come back to this issue later, after we have proven optimality of the patch-wise version.

SOLVE Given a triangulation $\mathcal{T}$, this method computes

$$U = \texttt{SOLVE}(\mathcal{T}),$$

with $U \in \mathbb{V}(\mathcal{T})$ the Ritz-Galerkin solution in *exact* arithmetic. The effects of using an inexact solver have been studied in the literature as well, see for example [33].

ESTIMATE Given a triangulation $\mathcal{T}$ with vertices $\mathcal{V}$ and its discrete solution $U \in \mathbb{V}(\mathcal{T})$, this module calculates the total error indicators — the equilibrated flux estimator plus the oscillation. We have

$$\left\{\vartheta_{\mathcal{T}}(U,\omega_a)\right\}_{a\in\mathcal{V}} = \texttt{ESTIMATE}(U,\mathcal{T}).$$

MARK A *Dörfler Marking* strategy [17] is used for marking. For a parameter $\theta \in (0,1]$, we calculate

$$\mathcal{M} = \texttt{MARK}(\left\{\vartheta_{\mathcal{T}}(U,\omega_a)\right\}_{a\in\mathcal{V}},\mathcal{T}),$$

where $\mathcal{M} \subset \mathcal{W}$ is the *minimal* cardinality set that satisfies

$$\vartheta_{\mathcal{T}}(U,\mathcal{M}) \geqslant \theta\vartheta_{\mathcal{T}}(U,\mathcal{W}).$$

Notice that marking is done using the total error estimator, in contrast to the standard residual case in §1.5, where we marked using only the residual estimator.

REFINE This module refines the marked elements in $\mathcal{T}$. It calculates

$$\mathcal{T}_\star = \texttt{REFINE}(\mathcal{T},\mathcal{M}),$$

with $\mathcal{T}_\star$ the smallest conforming refinement of $\mathcal{T}$ such that the triangles of all patches in the marked set $\mathcal{M}$ are bisected at least 3 times. To provide conformity, one generally also needs to refine extra elements that were not marked.

This three times refinement rule ensures the so-called *interior node* property, which is also used by Mekchay and Nochetto [24] and Morin et al. in [25]. The interior node property states that all the marked triangles, as well as their sides, contain a node of $\mathcal{T}_\star$ in their interior. This condition is needed for proving discrete efficiency. Cascón and Nochetto [12] avoid refining every element three times by spreading the refinements across consecutive steps; they ensure that the interior node property is satisfied after a fixed number of refinement steps.

We can now formulate the AFEM algorithm and its produced sequences. For this we use the iteration counter k as a subscript to differentiate among the sets. Let an initial triangulation $\mathcal{T}_0$ be given, set $k = 0$ and iterate the following steps:

1. $U_k = \texttt{SOLVE}(\mathcal{T}_k)$;
2. $\{\vartheta_k(U_k, \omega_a)\}_{a \in \mathcal{V}_k} = \texttt{ESTIMATE}(U_k, \mathcal{T}_k)$;
3. $\mathcal{M}_k = \texttt{MARK}(\{\vartheta_k(U_k, \omega_a)\}_{a \in \mathcal{V}_k}, \mathcal{T}_k)$;
4. $\mathcal{T}_{k+1} = \texttt{REFINE}(\mathcal{T}_k, \mathcal{M}_k)$;
5. $k := k + 1$.

3.2. Optimality conditions

Cascón and Nochetto provide a general list of assumptions in [12, §4], under which the above AFEM algorithm produces an optimal sequence of solutions $(U_k, \mathcal{T}_k)_{k \geqslant 0}$. They do not prove that the equilibrium flux estimator actually satisfies these assumptions. In Theorem 3.1 we will prove that these assumptions indeed hold for the equilibrium flux estimator. In AFEM driven by the standard residual estimator, we had a refined set containing the triangles that were refined. In the current case, however, we need to define similar sets, but then in terms of refined patches.

First, let $\mathcal{R}^0_{\mathcal{T} \to \mathcal{T}_\star}$ denote the set of all patches in $\mathcal{T}$ that contain at least *one* triangle that is refined when going to $\mathcal{T}_\star$, i.e.

$$\mathcal{R}^0_{\mathcal{T} \to \mathcal{T}_\star} := \mathcal{W} \setminus \mathcal{W}_\star = \{\omega_a \in \mathcal{W} : K \notin \mathcal{T}_\star \quad \text{for some } K \subset \omega_a\}.$$

For $j \geqslant 1$, the set $\mathcal{R}^j_{\mathcal{T} \to \mathcal{T}_\star}$ consists of all patches ω_a such that every triangle in ω_a is bisected at least j times when going from $\mathcal{T}$ to $\mathcal{T}_\star$. Formally,

$$\mathcal{R}^j_{\mathcal{T} \to \mathcal{T}_\star} := \Big\{\omega_a \in \mathcal{W} : g(K') - g(K) \geqslant j,\ \forall K' \in \mathcal{T}_\star \text{ with } K' \subset K,\ \forall K \in \mathcal{T} \text{ with } K \subset \omega_a\Big\},$$

where $g(K)$ is the generation of K — the number of bisections needed to create K from the initial triangulation $\mathcal{T}_0$. For $j = 1$, this set contains all patches that are *entirely* refined, for $j = 3$ it contains all patches that satisfy the interior node property. Notice that $\mathcal{R}^0_{\mathcal{T} \to \mathcal{T}_\star} \subset \mathcal{R}^1_{\mathcal{T} \to \mathcal{T}_\star} \subset \dots$. We let $\Omega^j_{\mathcal{R}} \subset \Omega$ denote the domain spanned by the patches $\omega_a \in \mathcal{R}^j_{\mathcal{T} \to \mathcal{T}_\star}$. With this extra notation we are able to state and prove the assumptions for optimality given in [12].

Theorem 3.1. *Let a triangulation $\mathcal{T}$ and a refinement $\mathcal{T}_\star \geqslant \mathcal{T}$ be given, with $U \in \mathbb{V}(\mathcal{T})$ and $U_\star \in \mathbb{V}(\mathcal{T})$ the appropriate discrete solutions. There exists constants C_1, C_2, C_3 such that:*

1. *The estimator is reliable; the approximation error can be bound using the total error estimator:*

$$|||u - U|||^2_\Omega \leqslant C_1 \vartheta^2_{\mathcal{T}}(U, \mathcal{W}) = C_1 \left[\eta^2_{\mathcal{T}}(U, \mathcal{W}) + \operatorname{osc}^2_{\mathcal{T}}(f, \mathcal{W})\right] \tag{3.1}$$

2. *The estimator is efficient; the estimator provides a lower bound for the error (up to oscillation):*
$$C_2\eta_{\mathcal{T}}^2(U,\mathcal{W}) \leqslant |||u-U|||_\Omega^2 + \operatorname{osc}_{\mathcal{T}}^2(f,\mathcal{W}). \tag{3.2}$$

3. *The estimator provides discrete reliability; the error between U and $U_\star$ can be bound using the refined set $\mathcal{R}^0_{\mathcal{T}\to\mathcal{T}_\star}$:*
$$|||U_\star - U|||_\Omega^2 \leqslant C_1\vartheta_{\mathcal{T}}^2(U,\mathcal{R}^0_{\mathcal{T}\to\mathcal{T}_\star}). \tag{3.3}$$

4. *The estimator provides discrete efficiency, i.e. the reverse of the above*
$$C_3\eta_{\mathcal{T}}^2(U,\mathcal{R}^3_{\mathcal{T}\to\mathcal{T}_\star}) \leqslant |||U_\star - U|||^2_{\Omega_{\mathcal{R}^3}} + \operatorname{osc}_{\mathcal{T}}^2(U,\mathcal{R}^3_{\mathcal{T}\to\mathcal{T}_\star}). \tag{3.4}$$

Proof.

Reliability and Efficiency. Reliability follows directly from Theorem 2.9:

$$|||u-U|||_\Omega^2 \lesssim \sum_{a\in\mathcal{V}} \|\underline{\sigma_a}\|_{\omega_a}^2 + \|h_a(I-\Pi^a_{p-1})(f)\|_{\omega_a}^2 = \sum_{a\in\mathcal{V}} \eta_{\mathcal{T}}^2(U,\omega_a) + \operatorname{osc}_{\mathcal{T}}^2(f,\omega_a) = \vartheta_{\mathcal{T}}^2(U,\mathcal{W}).$$

Efficiency follows from Theorem 2.10. ■

Discrete reliability. Discrete reliability can be proven using equivalence with the classical residual estimator. We opt not to follow this road, and instead give the following direct proof, inspired by [12, §3.2]. Write $E_\star = U_\star - U$ for the discrete error and consider a $V \in \mathbb{V}(\mathcal{T})$ to be specified later. Notice that $\mathbb{V}(\mathcal{T}) \subset \mathbb{V}(\mathcal{T}_\star)$, and thus from Galerkin orthogonality we deduce

$$|||E_\star|||_\Omega^2 = a(E_\star, E_\star - V) = a(u-U, E_\star - V) = \langle r, E_\star - V\rangle = \sum_{a\in\mathcal{V}} \langle r_a, E_\star - V\rangle \tag{3.5}$$

The second equality holds since $U_\star$ is the Galerkin approximation on $\mathbb{V}(\mathcal{T}_\star) \ni E_\star - V$. From orthogonality of r_a on constants for $a \in \mathcal{V}^{int}$, we infer that

$$\sum_{a\in\mathcal{V}} \langle r_a, E_\star - V\rangle \leqslant \sum_{a\in\mathcal{V}} \|r_a\|_{H^1_\star(\omega_a)'} \left\|\nabla\left(E_\star - V\right)\right\|_{\omega_a}. \tag{3.6}$$

We will construct V using the Scott-Zhang interpolant [32] of $E_\star$.

Consider $\mathcal{R} := \mathcal{R}^0_{\mathcal{T}\to\mathcal{T}_\star}$, the set of patches in $\mathcal{T}$ such that at least *one* triangle is refined when going from $\mathcal{T}$ to $\mathcal{T}_\star$. The union of patches in $\mathcal{R}$ make up a domain $\Omega_{\mathcal{R}} \subset \Omega$. Let Ω_k be one of the connected components of the interior of $\Omega_{\mathcal{R}}$, and consider $\mathbb{V}_k(\mathcal{T})$ the restriction of $\mathbb{V}(\mathcal{T})$ to Ω_k. We use the Scott-Zhang interpolation operator onto the space $\mathbb{V}_k(\mathcal{T})$, i.e. $I_k : H^1(\Omega_k) \to \mathbb{V}_k(\mathcal{T})$ as defined in [32, p. 2.13]. An important property of I_k is that it preserves conforming boundary values: for $v \in H^1(\Omega_k)$ with $v = \tilde{v}$ on $\partial\Omega_k$ for some $\tilde{v} \in \mathbb{V}_k(\mathcal{T})$, we have $I_k v = v$ on $\partial\Omega_k$. Now define V by

$$V := I_k E_\star \quad \text{in } \Omega_k, \quad V := E_\star \quad \text{elsewhere.}$$

Since $E_\star$ has conforming boundary values on $\partial\Omega_k$, we infer that V is continuous. Furthermore, as $\Omega_\mathcal{R}$ is the union all the patches that were refined, we conclude that $V \in \mathbb{V}(\mathcal{T})$.

Plug this specific V into equation (3.6) and use that $V = E_\star$ outside $\Omega_\mathcal{R}$ to discover

$$\begin{aligned}\sum_{a\in\mathcal{V}} \|r_a\|_{H^1_\star(\omega_a)'} \|\nabla(E_\star - V)\|_{\omega_a} &= \sum_{\omega_a\in\mathcal{R}} \|r_a\|_{H^1_\star(\omega_a)'} \|\nabla(E_\star - V)\|_{\omega_a} \\ &\leqslant \sqrt{\sum_{\omega_a\in\mathcal{R}} \|r_a\|^2_{H^1_\star(\omega_a)'}} \sqrt{\sum_{\omega_a\in\mathcal{R}} \|\nabla(E_\star - V)\|^2_{\omega_a}}.\end{aligned} \tag{3.7}$$

Focus on this last factor. We have $E_\star - V = (I - I_k)(E_\star)$ on every connected component Ω_k. From the Scott-Zhang interpolation theory (cf. [32, Thm 4.1]) we know that $\|(I - I_k)(E_\star)\|_{H^1(\Omega_k)} \lesssim \|E_\star\|_{H^1(\Omega_k)}$, for a constant only depending on shape regularity. Since every triangle in $\Omega_\mathcal{R}$ is contained in a maximum of three patches we deduce

$$\begin{aligned}\sum_{\omega_a\in\mathcal{R}} \|\nabla(E_\star - V)\|^2_{\omega_a} &\leqslant \sum_{\omega_a\in\mathcal{R}} \|E_\star - V\|^2_{H^1(\omega_a)} \\ &\eqsim \sum_{\{\Omega_k\subset\Omega_\mathcal{R}\}} \|E_\star - V\|^2_{H^1(\Omega_k)} \\ &\lesssim \sum_{\{\Omega_k\subset\Omega_\mathcal{R}\}} \|E_\star\|^2_{H^1(\Omega_k)} \\ &\leqslant \|E_\star\|^2_{H^1(\Omega)} \lesssim |||E_\star|||^2_\Omega .\end{aligned}$$

The last inequality follows since $U, U_\star \in H^1_0(\Omega)$ and thus that $E_\star \in H^1_0(\Omega)$, which allows us to invoke Poincare-Friedrichs inequality. Insert this derivation into (3.7), and chain its result with equations (3.5) and (3.6), to obtain

$$|||E_\star|||^2_\Omega \leqslant \sum_{a\in\mathcal{V}} \|r_a\|_{H^1_\star(\omega_a)'} \|\nabla(E_\star - V)\|_{\omega_a} \lesssim \sqrt{\sum_{\omega_a\in\mathcal{R}} \|r_a\|^2_{H^1_\star(\omega_a)'}}\, |||E_\star|||_\Omega .$$

For the first factor in this equality we can use that $\|r_a\|_{H^1_\star(\omega_a)'} \leqslant \|\underline{\sigma_a}\|_{\omega_a} + \operatorname{osc}_\mathcal{T}(f, \omega_a)$, as shown in the proof of Lemma 2.8. Standard calculations then prove discrete reliability:

$$|||U_\star - U|||^2_\Omega \lesssim \sum_{w_a\in\mathcal{R}} \|r_a\|^2_{H^1_\star(\omega_a)'} \lesssim \sum_{\omega_a\in\mathcal{R}} \|\underline{\sigma_a}\|^2_{\omega_a} + \operatorname{osc}^2_\mathcal{T}(f, \omega_a) = \vartheta^2_\mathcal{T}(U, \mathcal{R}). \qquad \blacksquare$$

Discrete efficiency. Discrete efficiency will be proven using equivalence with the classical residual estimator. Write $\mathcal{R}^j := \mathcal{R}^j_{\mathcal{T}\to\mathcal{T}_\star}$, and invoke Lemma 2.11 for equivalence with the standard residual estimator:

$$\eta^2_\mathcal{T}(U, \mathcal{R}^3) = \sum_{\omega_a\in\mathcal{R}^3} \|\underline{\sigma_a}\|^2_{\omega_a} \lesssim \sum_{\omega_a\in\mathcal{R}^3} \left[\operatorname{osc}^2_\mathcal{T}(f, \omega_a) + h^2_{\omega_a}\|f + \Delta U\|^2_{\omega_a} + h_{\omega_a}\|[\![\nabla U]\!]\|^2_{\gamma_a}\right].$$

Focus on the last two terms of this expression. Every patch $\omega_a \in \mathcal{R}^3$ consists entirely of triangles that satisfy the interior node property. This implies that both the triangular neighbours of an edge $e \subset \gamma_a$ satisfy the interior node property. For such an edge e, there

exists a discrete lower bound using the classical residual *edge* estimator ([33, Thm 4.3] or [26, Lem 4.2]):

$$\sum_{K\in\mathcal{T}_e}\left[h_K^2\|f+\Delta U\|_K^2\right]+h_e\|[\![U]\!]\|_e^2\lesssim\sum_{K\in\mathcal{T}_e}\left[|||U_\star-U|||_K^2+h_K^2\|(I-\Pi_{p-1})(f)\|_K^2\right],$$

where $\mathcal{T}_e$ consists of the two triangles $K\in\mathcal{T}$ adjacent to the (interior) edge $e\subset\gamma_a$.

Using that $h_K\eqsim h_{\omega_a}\eqsim h_e$ for every triangle $K\subset\omega_a$ and edge $e\subset\gamma_a$, allow us to relate the classical residual edge estimator to its patch-wise version:

$$h_{\omega_a}^2\|f+\Delta U\|_{\omega_a}^2+h_{\omega_a}\|[\![\nabla U]\!]\|_{\gamma_a}^2\eqsim\sum_{e\subset\gamma_a}\left[\sum_{K\in\mathcal{T}_e}h_K^2\|f+\Delta U\|_K^2+h_e\|[\![\nabla U]\!]\|_e^2\right].$$

After chaining the above results, we obtain discrete efficiency:

$$\begin{aligned}
\eta_{\mathcal{T}}^2(U,\mathcal{R}^3)&\lesssim\sum_{\omega_a\in\mathcal{R}^3}\left[\operatorname{osc}_{\mathcal{T}}^2(f,\omega_a)+h_{\omega_a}^2\|f+\Delta U\|_{\omega_a}^2+h_{\omega_a}\|[\![\nabla U]\!]\|_{\gamma_a}^2\right]\\
&\eqsim\operatorname{osc}_{\mathcal{T}}^2(f,\mathcal{R}^3)+\sum_{\omega_a\in\mathcal{R}^3}\sum_{e\subset\gamma_a}\left[\sum_{K\in\mathcal{T}_e}h_K^2\|f+\Delta U\|_K^2+h_e\|[\![\nabla U]\!]\|_e^2\right]\\
&\lesssim\operatorname{osc}_{\mathcal{T}}^2(f,\mathcal{R}^3)+\sum_{\omega_a\in\mathcal{R}^3}\sum_{e\subset\gamma_a}\sum_{K\in\mathcal{T}_e}\left[|||U_\star-U|||_K^2+h_K^2\|(I-\Pi_{p-1})(f)\|_K^2\right]\\
&\eqsim\operatorname{osc}_{\mathcal{T}}^2(f,\mathcal{R}^3)+\sum_{\omega_a\in\mathcal{R}^3}\left[|||U_\star-U|||_{\omega_a}^2+\|h_a(I-\Pi_{p-1})(f)\|_{\omega_a}^2\right]\\
&\eqsim\operatorname{osc}_{\mathcal{T}}^2(f,\mathcal{R}^3)+|||U_\star-U|||_{\Omega_{\mathcal{R}}^3}^2.
\end{aligned}$$

□

3.3. Contraction property

It is possible that the equilibrated flux estimator does not strictly decrease upon a refinement. The oscillation term, on the other hand, does satisfy such a reduction property.

Lemma 3.2. *The oscillation satisfies the oscillation reduction property. That is, there exists a constant* $0<\lambda<1$*, such that for any refinement* $\mathcal{T}_\star\geqslant\mathcal{T}$ *one has*

$$\operatorname{osc}_{\mathcal{T}_\star}^2(f,\mathcal{W}_{\mathcal{T}_\star})\leqslant\operatorname{osc}_{\mathcal{T}}^2(f,\mathcal{W}_{\mathcal{T}})-\lambda\operatorname{osc}_{\mathcal{T}}^2(f,\mathcal{R}^3_{\mathcal{T}\to\mathcal{T}_\star}).$$

Proof. Denote $\mathcal{W}=\mathcal{W}_{\mathcal{T}}$ and $\mathcal{W}_\star=\mathcal{W}_{\mathcal{T}_\star}$. Rewriting the left hand side yields,

$$\begin{aligned}
\operatorname{osc}_{\mathcal{T}_\star}^2(f,\mathcal{W}_\star)&=\operatorname{osc}_{\mathcal{T}_\star}^2(f,\mathcal{W}_\star\cap\mathcal{W})+\operatorname{osc}_{\mathcal{T}_\star}^2(f,\mathcal{W}_\star\setminus\mathcal{W})\\
&=\operatorname{osc}_{\mathcal{T}}^2(f,\mathcal{W}_\star\cap\mathcal{W})+\operatorname{osc}_{\mathcal{T}_\star}^2(f,\mathcal{W}_\star\setminus\mathcal{W})\\
&=\operatorname{osc}_{\mathcal{T}}^2(f,\mathcal{W})-\operatorname{osc}_{\mathcal{T}}^2(f,\mathcal{W}\setminus\mathcal{W}_\star)+\operatorname{osc}_{\mathcal{T}_\star}^2(f,\mathcal{W}_\star\setminus\mathcal{W}).
\end{aligned}$$

We see that the result holds if there exists $0<\lambda<1$ such that

$$\operatorname{osc}_{\mathcal{T}_\star}^2(f,\mathcal{W}_\star\setminus\mathcal{W})\leqslant\operatorname{osc}_{\mathcal{T}}^2(f,\mathcal{W}\setminus\mathcal{W}_\star)-\lambda\operatorname{osc}_{\mathcal{T}}^2(f,\mathcal{R}^3_{\mathcal{T}\to\mathcal{T}_\star}).\tag{3.8}$$

The set $\mathcal{W}_\star \setminus \mathcal{W}$ contains all newly introduced patches in $\mathcal{T}_\star$, whereas $\mathcal{W} \setminus \mathcal{W}_\star$ contains all patches that have at least one triangle refined.

The use of patches instead of elements makes things a little confusing. Therefore we start with a triangle K that is contained in some patch from $\mathcal{R}^3_{\mathcal{T}\to\mathcal{T}_\star}$. This triangle is bisected three times, so the diameter of any subtriangle $K_\star \in \mathcal{T}_\star$ is reduced by a factor $\frac{1}{2}$, i.e. $h_{K_\star} \leqslant \frac{1}{2} h_K$. Together with properties of the orthogonal projector Π_{p-1}, we infer

$$\begin{aligned}\sum_{\{K_\star\in\mathcal{T}_\star : K_\star\subset K\}} h^2_{K_\star}\|(I-\Pi_{p-1})(f)\|^2_{K_\star} &\leqslant \frac{1}{4}h^2_K \sum_{\{K_\star\in\mathcal{T}_\star : K_\star\subset K\}} \|(I-\Pi_{p-1})(f)\|^2_{K_\star} \\ &\leqslant \frac{1}{4}h^2_K\|(I-\Pi_{p-1})(f)\|^2_K\,.\end{aligned} \tag{3.9}$$

That is, oscillation on K is strictly reduced after three refinements. Enlarging this to a single patch $\omega_a \in \mathcal{R}^3_{\mathcal{T}\to\mathcal{T}_\star}$ gives

$$\sum_{\{K_\star\in\mathcal{T}_\star : K_\star\subset \omega_a\}} h^2_{K_\star}\|(I-\Pi_{p-1})(f)\|^2_{K_\star} \leqslant \frac{1}{4}\sum_{K\subset\omega_a} h^2_K\|(I-\Pi_{p-1})(f)\|^2_K = \frac{1}{4}\operatorname{osc}^2_{\mathcal{T}}(f,\omega_a). \tag{3.10}$$

Unfortunately, this left hand side cannot be written as the sum over patches in $\mathcal{T}_\star$. The difficulty is that refinement of ω_a introduces new patches in $\mathcal{T}_\star$, for which some only partially overlap with ω_a.

For this reason, we will decompose the patch oscillations from (3.8) over its triangles:

$$\begin{aligned}\operatorname{osc}^2_{\mathcal{T}}(f,\mathcal{W}\setminus\mathcal{W}_\star) &= \sum_{\omega_a\in\mathcal{W}\setminus\mathcal{W}_\star}\sum_{K\subset\omega_a} h^2_K\|(I-\Pi_{p-1})(f)\|^2_K\,,\\ \operatorname{osc}^2_{\mathcal{T}_\star}(f,\mathcal{W}_\star\setminus\mathcal{W}) &= \sum_{\omega_a\in\mathcal{W}_\star\setminus\mathcal{W}}\sum_{K_\star\subset\omega_a} h^2_{K_\star}\|(I-\Pi_{p-1})(f)\|^2_{K_\star}\,.\end{aligned}$$

Pick some triangle $K \in \mathcal{T}$ from the upper sum, and consider an arbitrary subtriangle $K_\star \in \mathcal{T}_\star$ of K. This triangle K is contained in either one, two or three patches ω_a from $\mathcal{W}\setminus\mathcal{W}_\star$. The crucial observation is that $K_\star$ is contained in the same number of patches from $\mathcal{W}_\star\setminus\mathcal{W}$. In other words, every triangle K appearing in the upper sum induces a set of subtriangles $\{K_\star \in \mathcal{T}_\star : K_\star \subset K\}$ in the lower sum. Use this relation and apply the strict oscillation reduction on patches $\omega_a \in \mathcal{R}^3_{\mathcal{T}\to\mathcal{T}_\star}$ from (3.10) to obtain

$$\begin{aligned}\operatorname{osc}^2_{\mathcal{T}_\star}(f,\mathcal{W}_\star\setminus\mathcal{W}) &\leqslant \operatorname{osc}^2_{\mathcal{T}}\left(f,(\mathcal{W}\setminus\mathcal{W}_\star)\setminus\mathcal{R}^3_{\mathcal{T}\to\mathcal{T}_\star}\right) + \frac{1}{4}\operatorname{osc}^2_{\mathcal{T}}\left(f,\mathcal{R}^3_{\mathcal{T}\to\mathcal{T}_\star}\right)\\ &= \operatorname{osc}^2_{\mathcal{T}}\left(f,\mathcal{W}\setminus\mathcal{W}_\star\right) - \frac{3}{4}\operatorname{osc}^2_{\mathcal{T}}\left(f,\mathcal{R}^3_{\mathcal{T}\to\mathcal{T}_\star}\right).\end{aligned}$$ □

We can now prove the so-called contraction property; this shows convergence of a weighted average of the approximation - and the oscillation error produced the AFEM method described above.

Theorem 3.3. *There exists a constant $0<\alpha<1$ and $\gamma>0$, depending on the shape regularity of $\mathcal{T}_0$, and $\{\lambda,\theta\}$ such that*

$$|||u-U_{k+1}|||^2 + \gamma\operatorname{osc}^2_{k+1}(f,\mathcal{W}_{k+1}) \leqslant \alpha^2\left(|||u-U_k|||^2 + \gamma\operatorname{osc}^2_k(f,\mathcal{W}_k)\right)$$

Proof. Adapt the notation used in [12], i.e. for $j = 0, 1, 2, \dots$ we write

$$\begin{aligned} e_j^2 &:= \left|\!\left|\!\left| u - U_j \right|\!\right|\!\right|_\Omega^2, & E_j^2 &:= \left|\!\left|\!\left| U_{j+1} - U_j \right|\!\right|\!\right|_\Omega^2, \\ \operatorname{osc}_j^2 &:= \operatorname{osc}_{\mathcal{T}_j}^2(f, \mathcal{W}_{\mathcal{T}_j}), & \operatorname{osc}_j^2(\mathcal{M}_j) &:= \operatorname{osc}_{\mathcal{T}_j}^2(f, \mathcal{M}_j), \\ \eta_j^2 &:= \eta_{\mathcal{T}_j}^2(U_j, \mathcal{W}_{\mathcal{T}_j}), & \eta_j^2(\mathcal{M}_j) &:= \eta_{\mathcal{T}_j}^2(U, \mathcal{M}_j), \\ \mathcal{R}_j^i &:= \mathcal{R}_{\mathcal{T}_j \to \mathcal{T}_{j+1}}^i. \end{aligned}$$

A piecewise polynomial on $\mathcal{T}_k$ is also a piecewise polynomial on $\mathcal{T}_{k+1} \geqslant \mathcal{T}_k$. Therefore, we have $U_{k+1} - U_k \in \mathbb{V}(\mathcal{T}_{k+1})$, and thus by Galerkin orthogonality

$$\left|\!\left|\!\left| u - U_k \right|\!\right|\!\right|_\Omega^2 = \left|\!\left|\!\left| u - U_{k+1} \right|\!\right|\!\right|_\Omega^2 + \left|\!\left|\!\left| U_{k+1} - U_k \right|\!\right|\!\right|_\Omega^2 \implies e_{k+1}^2 = e_k^2 - E_k^2.$$

Introduce some constants $\gamma > 0$ and $0 < \beta < 1$ that will be selected later:

$$e_{k+1}^2 + \gamma \operatorname{osc}_{k+1}^2 = e_k^2 - E_k^2 + \gamma \operatorname{osc}_{k+1}^2 \leqslant e_k^2 - \beta E_k^2 + \gamma \operatorname{osc}_{k+1}^2 .$$

Application of the oscillation reduction property from Lemma 3.2 results in

$$e_{k+1}^2 + \gamma \operatorname{osc}_{k+1}^2 \leqslant e_k^2 - \beta E_k^2 + \gamma \operatorname{osc}_k^2 - \gamma\lambda \operatorname{osc}_k^2(\mathcal{R}_k^3). \tag{3.11}$$

Discrete efficiency (3.4) from Theorem 3.1 reads as

$$E_k^2 = \left|\!\left|\!\left| U_{k+1} - U_k \right|\!\right|\!\right|_\Omega^2 \geqslant C_3 \eta_k^2(\mathcal{R}_k^3) - \operatorname{osc}_k^2(\mathcal{R}_k^3) \geqslant C_3 \eta_k^2(\mathcal{M}_k) - \operatorname{osc}_k^2(\mathcal{R}_k^3),$$

where the last inequality follows since the all marked patches satisfy the interior node property when going from $\mathcal{T}_k$ to $\mathcal{T}_{k+1}$, and thus $\mathcal{M}_k \subset \mathcal{R}_k^3$. By inserting this into inequality (3.11) we obtain

$$\begin{aligned} e_{k+1}^2 + \gamma \operatorname{osc}_{k+1}^2 &\leqslant e_k^2 - \beta C_3 \eta_k^2(\mathcal{M}_k) + \beta \operatorname{osc}_k^2(\mathcal{R}_k^3) + \gamma \operatorname{osc}_k^2 - \gamma\lambda \operatorname{osc}_k^2(\mathcal{R}_k^3) \\ &= e_k^2 + \gamma \operatorname{osc}_k^2 - \beta C_3 \eta_k^2(\mathcal{M}_k) - [\lambda\gamma - \beta] \operatorname{osc}_k^2(\mathcal{R}_k^3). \end{aligned}$$

Under the assumption that $\lambda\gamma \geqslant \beta$, we may replace $\operatorname{osc}_k^2(\mathcal{R}_k^3)$ by the smaller $\operatorname{osc}_k^2(\mathcal{M}_k)$:

$$e_{k+1}^2 + \gamma \operatorname{osc}_{k+1}^2 \leqslant e_k^2 + \gamma \operatorname{osc}_k^2 - C_3 \beta \eta_k^2(\mathcal{M}_k) - [\lambda\gamma - \beta] \operatorname{osc}_k^2(\mathcal{M}_k).$$

Now pick β such that the coefficients of $\operatorname{osc}_k^2(\mathcal{M}_k)$ and $\eta_k^2(\mathcal{M}_k)$ match, i.e.

$$\beta C_3 = \lambda\gamma - \beta \implies \beta := \frac{1}{1 + C_3} \lambda\gamma,$$

which ensures that $\lambda\gamma \geqslant \beta$. Substitute β and use the definition $\vartheta_k^2 = \eta_k^2 + \operatorname{osc}_k^2$ to attain

$$\begin{aligned} e_{k+1}^2 + \gamma \operatorname{osc}_{k+1}^2 &\leqslant e_k^2 + \gamma \operatorname{osc}_k^2 - \frac{C_3}{1 + C_3} \lambda\gamma \vartheta_k^2(\mathcal{M}_k) \\ &\leqslant e_k^2 + \gamma \operatorname{osc}_k^2 - \frac{C_3}{1 + C_3} \lambda\gamma\theta^2 \vartheta_k^2, \end{aligned}$$

with the last step following from the Dörfler marking property, i.e. $\vartheta_k^2(\mathcal{M}_k) \geqslant \theta^2 \vartheta_k^2$. Since the total error dominates oscillation, $\vartheta_k^2 \geqslant \operatorname{osc}_k^2$, we infer that

$$e_{k+1}^2 + \gamma \operatorname{osc}_{k+1}^2 \leqslant e_k^2 + \gamma \operatorname{osc}_k^2 - \frac{C_3}{2(1+C_3)} \lambda\gamma\theta^2 \left(\vartheta_k^2 + \operatorname{osc}_k^2\right).$$

Rewriting the (global) reliability bound (3.1) shows $\vartheta_k^2 \geqslant C_1^{-1} e_k^2$, and thus

$$\begin{aligned} e_{k+1}^2 + \gamma \operatorname{osc}_{k+1}^2 &\leqslant \left(1 - \frac{C_3\lambda\theta^2}{2C_1(1+C_3)}\gamma\right) e_k^2 + \gamma \left(1 - \frac{C_3}{2(1+C_3)}\lambda\theta^2\right) \operatorname{osc}_k^2 \\ &:= \alpha_1^2(\gamma) e_k^2 + \gamma \alpha_2^2 \operatorname{osc}_k^2 . \end{aligned}$$

Since λ and θ^2 are both contained in the interval $(0,1)$ we have $0 < \alpha_2^2 < 1$. We are left to pick γ such that $0 < \alpha_1^2(\gamma) < 1$, which translates into

$$0 < \gamma < \frac{2C_1(1+C_3)}{C_3\lambda\theta^2}.$$

Any γ satisfying the above completes the proof, since $\alpha^2 = \max\left\{\alpha_1^2(\gamma), \alpha_2^2\right\} < 1$ with

$$e_{k+1}^2 + \gamma \operatorname{osc}_{k+1}^2 \leqslant \alpha_1^2(\gamma) e_k^2 + \gamma\alpha_2^2 \operatorname{osc}_k^2 \leqslant \alpha^2 \left(e_k^2 + \gamma \operatorname{osc}_k^2\right).$$

□

3.4. Optimality of AFEM

We will show that the discrete solutions U_k convergence with the best possible rate. More precisely, we will prove that the total error $|||u - U_k|||_\Omega + \operatorname{osc}_{\mathcal{T}}^2(f, \mathcal{W}_k)$ on the adaptively generated meshes $\mathcal{T}_k$ decays at the same rate as the total error of the discrete solution U on the best possible triangulation $\mathcal{T}$ with $\#\mathcal{T}_k$ triangles. To formalise this, we require some additional definitions.

From the reliability and efficiency of the estimator we know that

$$|||u - U|||_\Omega^2 + \operatorname{osc}_{\mathcal{T}}^2(f, \mathcal{W}) \eqsim \eta_{\mathcal{T}}^2(U, \mathcal{W}) + \operatorname{osc}_{\mathcal{T}}^2(f, \mathcal{W}) = \vartheta_{\mathcal{T}}^2(U, \mathcal{W}).$$

The total error estimator is equivalent to the approximation error (up to an oscillation term). Since we cannot get rid of the oscillation term (cf. [13, Rem 5.1]), it makes sense to incorporate this into the definition of an approximation class.

Denote $\mathbb{T}$ for the set of all conforming refinements of $\mathcal{T}_0$ found by applying bisection. Moreover, let $\mathbb{T}_N \subset \mathbb{T}$ be the restriction of triangulations with at most N triangles more than $\mathcal{T}_0$, i.e. $\mathcal{T} \in \mathbb{T}_N$ if $\mathcal{T} \leqslant \#T_0 + N$. We define the optimal approximation class in terms of $s > 0$. For $v \in H_0^1(\Omega)$ with $\Delta v \in L^2(\Omega)$, let $V_{\mathcal{T}} \in \mathbb{V}(\mathcal{T})$ denote the Ritz-Galerkin approximation of the Poisson problem associated with v. A measure for the optimal convergence rate of v is given by

$$|v|_{\mathcal{A}_s} := \sup_{N>0} (N+1)^s \inf_{\mathcal{T} \in \mathbb{T}_N} \left(|||v - V_{\mathcal{T}}|||_\Omega^2 + \operatorname{osc}_{\mathcal{T}}^2(f, \mathcal{W})\right)^{1/2}.$$

As approximation class we take all the functions for which this measure is finite, i.e.

$$\mathcal{A}_s := \left\{ v \in H^1_0(\Omega) : \Delta v \in L^2(\Omega), \quad |v|_{\mathcal{A}_s} < \infty \right\}.$$

So if the Poisson solution u satisfies $u \in \mathcal{A}_s$, then the total error $e(N)$ of the discrete solution $U_N \in \mathbb{V}(\mathcal{T})$ on the *best* partition $\mathcal{T}_N \in \mathbb{T}_N$ satisfies

$$e(N) := \left(|||u - U_N|||^2_\Omega + \operatorname{osc}^2_{\mathcal{T}_N}(f, \mathcal{W}) \right)^{1/2} \leqslant (N+1)^{-s} |u|_{\mathcal{A}_s}. \tag{3.12}$$

Our goal is to prove that the sequence $(U_k)_{k \geqslant 0}$ satisfies a similar convergence rate.

Notice that this optimality class explicitly depends on our definition of oscillation, and thus differs from the definition for the classical estimator given in §1.6. These definitions are equivalent, as one can see from

$$\begin{aligned} \operatorname{osc}^2_{\mathcal{T}}(f, \mathcal{W}) &= \sum_{\omega_a \in \mathcal{W}} \sum_{K \subset \omega_a} h^2_K \|(I - \Pi_{p-1})(f)\|^2_K \\ &= 3 \sum_{K \in \mathcal{T}} h^2_K \|(I - \Pi_{p-1})(f)\|^2_K = 3 \hat{\operatorname{osc}}^2_{\mathcal{T}}(f, \mathcal{T}). \end{aligned}$$

The proof of optimality for AFEM driven by the equilibrated flux estimator is very similar to the standard case in §1.6. We follow the steps given by Cascón and Nochetto in [12]. Define the maximum Dörfler marking parameter by

$$\theta^2_* := \frac{C_2}{(1 + C_1)(1 + C_2)}. \tag{3.13}$$

The following lemma establishes a relation between the total error reduction and the Dörfler marking.

Lemma 3.4. *Let $\mathcal{T} \in \mathbb{T}$ a triangulation with discrete solution $U \in \mathbb{V}(\mathcal{T})$, a marking parameter $\theta \in (0, \theta_*)$ and a refinement $\mathcal{T}_\star \geqslant \mathcal{T} \in \mathbb{T}$ with discrete solution $U_\star \in \mathbb{V}(\mathcal{T}_\star)$ that satisfies*

$$|||u - U_\star|||^2_\Omega + \operatorname{osc}^2_{\mathcal{T}_\star}(f, \mathcal{W}_\star) \leqslant \mu(|||u - U|||^2_\Omega + \operatorname{osc}^2_{\mathcal{T}}(f, \mathcal{W})),$$

for $\mu := 1 - \frac{\theta^2}{\theta^2_*}$.

Then the refined set $\mathcal{R}^0 = \mathcal{R}^0_{\mathcal{T} \to \mathcal{T}_\star}$ satisfies the Dörfler property

$$\vartheta_{\mathcal{T}}(U, \mathcal{R}^0) \geqslant \theta \vartheta_{\mathcal{T}}(U, \mathcal{W}).$$

Proof. From the efficiency bound (3.2) in Theorem 3.1 we deduce

$$C_2 \vartheta^2_{\mathcal{T}}(U, \mathcal{W}) \leqslant |||u - U|||^2_\Omega + (1 + C_2) \operatorname{osc}^2_{\mathcal{T}}(f, \mathcal{W}) \leqslant (1 + C_2) \left(|||u - U|||_\Omega + \operatorname{osc}^2_{\mathcal{T}}(f, \mathcal{W}) \right),$$

and thus

$$\frac{C_2}{1 + C_2} \vartheta^2_{\mathcal{T}}(U, \mathcal{W}) \leqslant |||u - U|||^2_\Omega + \operatorname{osc}^2_{\mathcal{T}}(f, \mathcal{W}).$$

Together with the assumption this becomes

$$(1-\mu)\frac{C_2}{1+C_2}\vartheta^2_{\mathcal{T}}(U,\mathcal{W}) \leqslant (1-\mu)\left[|||u-U|||^2_\Omega + \operatorname{osc}^2_{\mathcal{T}}(f,\mathcal{W})\right] \\ \leqslant |||u-U|||^2_\Omega + \operatorname{osc}^2_{\mathcal{T}}(f,\mathcal{W}) - |||u-U_\star|||^2_\Omega - \operatorname{osc}^2_{\mathcal{T}_\star}(f,\mathcal{W}_\star). \tag{3.14}$$

Using the Galerkin orthogonality and $\mathcal{R}^0 = \mathcal{W}\setminus\mathcal{W}_\star$, shows

$$|||U_\star - U|||^2_\Omega = |||u-U|||^2_\Omega - |||u-U_\star|||^2_\Omega\,, \\ \operatorname{osc}^2_{\mathcal{T}}(f,\mathcal{W}) = \operatorname{osc}^2_{\mathcal{T}_\star}(f,\mathcal{W}\cap\mathcal{W}_\star) + \operatorname{osc}^2_{\mathcal{T}_\star}(f,\mathcal{R}^0) \leqslant \operatorname{osc}^2_{\mathcal{T}_\star}(f,\mathcal{W}_\star) + \operatorname{osc}^2_{\mathcal{T}}(f,\mathcal{R}^0).$$

After inserting these relations into (3.14) we infer that

$$\begin{aligned}(1-\mu)\frac{C_2}{1+C_2}\vartheta^2_{\mathcal{T}}(U,\mathcal{W}) &\leqslant |||U_\star - U|||^2_\Omega + \operatorname{osc}^2_{\mathcal{T}}(f,\mathcal{R}^0)\\ &\leqslant |||U_\star - U|||^2_\Omega + \vartheta_{\mathcal{T}}(U,\mathcal{R}^0)\\ &\leqslant (1+C_1)\vartheta^2_{\mathcal{T}}(U,\mathcal{R}^0),\end{aligned}$$

where the last step follows from discrete reliability (3.3). In conclusion, by definition of θ^2_* and μ we see that

$$\vartheta^2_{\mathcal{T}}(U,\mathcal{R}^0) \geqslant (1-\mu)\frac{C_2}{(1+C_1)(1+C_2)}\vartheta^2_{\mathcal{T}}(U,\mathcal{W}) \geqslant \theta^2\vartheta^2_{\mathcal{T}}(U,\mathcal{W}).$$

□

The following corollary relates the triangulations produced by AFEM with the optimal triangulations. This hinges on the fact that AFEM selects a *minimal* cardinality set $\mathcal{M}_k$. This key idea was first used by Stevenson in [33].

Corollary. *Let $\{\mathcal{T}_k,\mathbb{V}_k,U_k\}_{k\geqslant 0}$ be the sequence of results produced by AFEM for a marking parameter $\theta\in(0,\theta_*)$. If $u\in\mathcal{A}_s$, then the minimal Dörfler set $\mathcal{M}_k$ satisfies*

$$\#\mathcal{M}_k \lesssim \left(1-\frac{\theta^2}{\theta^2_*}\right)^{-1/2s} |u|^{1/s}_{\mathcal{A}_s}\left(|||u-U_k|||^2_\Omega + \operatorname{osc}^2_{\mathcal{T}_k}(f,\mathcal{W}_k)\right)^{-1/2s}.$$

Proof. Let μ be as in the previous Lemma and set $\epsilon^2 := \mu\left(|||u-U_k|||^2_\Omega + \operatorname{osc}^2_{\mathcal{T}_k}(f,\mathcal{W}_k)\right)$. Since $u\in\mathcal{A}_s$, by definition of the approximation class there exists a triangulation $\mathcal{T}_\epsilon\in\mathbb{T}$ with discrete solution U_ϵ such that[1]

$$\#\mathcal{T}_\epsilon - \#\mathcal{T}_0 \leqslant |u|^{1/s}_{\mathcal{A}_s}\,\epsilon^{-1/s}, \quad \text{and } |||u-U_\epsilon|||^2_\Omega + \operatorname{osc}^2_{\mathcal{T}_\epsilon}(f,\mathcal{W}_\epsilon) \leqslant \epsilon^2.$$

[1] To see this, let $e(N)$ denote the smallest error on a triangulation from $\mathbb{T}_N$. This function is decreasing, since $\mathbb{T}_N\subset\mathbb{T}_{N+1}$, and we have $e(N)\to 0$ as $N\to\infty$. Therefore, there exists a N such that $e(N)\leqslant\epsilon\leqslant e(N-1)$. From the definition we also find $N^s e(N-1)\leqslant|u|_{\mathcal{A}_s}$, and thus $N\leqslant e(N-1)^{-1/s}|u|^{1/s}_{\mathcal{A}_s}\leqslant\epsilon^{-1/s}|u|^{1/s}_{\mathcal{A}_s}$.

For $\mathcal{T}_\star = \mathcal{T}_k \oplus \mathcal{T}_\epsilon$ — the smallest common conforming refinement of $\mathcal{T}_k$ and $\mathcal{T}_\epsilon$ — we find from the Galerkin orthogonality and the oscillation reduction property that

$$|||u - U_\star|||_\Omega^2 + \operatorname{osc}_{\mathcal{T}_\star}^2(f, \mathcal{W}_\star) \leqslant |||u - U_\epsilon|||_\Omega^2 + \operatorname{osc}_{\mathcal{T}_\epsilon}^2(f, \mathcal{W}_\epsilon) \leqslant \epsilon^2.$$

By the choice of ϵ it follows that $u - U_\star$ satisfies the conditions of the previous lemma, and thus that $\mathcal{R}^0_{\mathcal{T}_k \to \mathcal{T}_\star}$ satisfies the Dörfler property. Since $\mathcal{M}_k$ is the *minimal* cardinality set that satisfies the Dörfler property we may conclude that $\#\mathcal{M}_k \leqslant \#\mathcal{R}^0_{\mathcal{T}_K \to \mathcal{T}_\star}$.

Now $\mathcal{R}^0_{\mathcal{T}_k \to \mathcal{T}_\star}$ consists of all those patches containing a refined triangle going from $\mathcal{T}_k$ to $\mathcal{T}_\star$. Since every patch has a uniformly bounded number of triangles, with a bound only depending on $\mathcal{T}_0$, we conclude

$$\#\mathcal{M}_k \leqslant \#\mathcal{R}^0_{\mathcal{T}_k \to \mathcal{T}_\star} \lesssim \#\mathcal{T}_\star - \#\mathcal{T}_k \leqslant \#\mathcal{T}_\epsilon - \#\mathcal{T}_0 \leqslant |u|_{\mathcal{A}_s}^{1/s}\, \epsilon^{-1/s}.$$

The third inequality is a property of the smallest common refinement [13, Lem 3.7]. □

We are almost ready to prove the main optimality result. For this we need to bound the number of refined triangles. This number can inflate beyond $\#\mathcal{M}_k$ in one iteration because of the conformity constraint. Fortunately, the cumulative sum behaves correctly as summarized in the following theorem.

Theorem 3.5. *For a constant only depending on shape regularity, we have*

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \lesssim \sum_{i=0}^{k-1} \#\mathcal{M}_i. \tag{3.15}$$

Proof. A proof for the classical (element-wise) residual estimator is given in [35]. In this standard case, the marked sets $\hat{\mathcal{M}}_i$ consist of *triangles* that are bisected once. We, however, consider marked sets $\mathcal{M}_i$ consisting of *patches* that are to be bisected three times. The proof in [35] is every technical and therefore omitted here. We only remark that the main arguments are also valid for our case, hence providing (3.15). □

We are finally ready to prove the main optimality theorem.

Theorem 3.6. *Suppose that the marking parameter θ satisfies $\theta \in (0, \theta_*)$ (see (3.13)), and suppose that $u \in \mathcal{A}_s$ for some $s > 0$.*

Then the produced sequence $\{U_k, \mathbb{V}_k, \mathcal{T}_k\}_{k \geqslant 0}$ of AFEM solutions satisfies

$$\left(|||u - U_k|||_\Omega^2 + \operatorname{osc}_k^2(f, \mathcal{W}_k)\right)^{1/2} \lesssim |u|_{\mathcal{A}_s} \left(\#\mathcal{T}_k - \#\mathcal{T}_0\right)^{-s},$$

for a constant depending on $\mathcal{T}_0$, but independent of u.

Proof. Combine the contraction property, the previous corollary and inequality (3.15):

$$\begin{aligned}
\#\mathcal{T}_k - \#\mathcal{T}_0 &\lesssim \sum_{i=0}^{k-1} \#\mathcal{M}_i \\
&\lesssim \left(1 - \frac{\theta^2}{\theta_*^2}\right)^{-1/2s} |u|_{\mathcal{A}_s}^{1/s} \sum_{i=0}^{k-1} \Big(|\!|\!| u - U_i |\!|\!|_\Omega^2 + \operatorname{osc}_{\mathcal{T}_i}^2(f, \mathcal{W}_i) \Big)^{-1/2s} \\
&\eqsim \left(1 - \frac{\theta^2}{\theta_*^2}\right)^{-1/2s} |u|_{\mathcal{A}_s}^{1/s} \sum_{i=0}^{k-1} \Big(|\!|\!| u - U_i |\!|\!|_\Omega^2 + \gamma \operatorname{osc}_{\mathcal{T}_i}^2(f, \mathcal{W}_i) \Big)^{-1/2s} \\
&\leqslant \left(1 - \frac{\theta^2}{\theta_*^2}\right)^{-1/2s} |u|_{\mathcal{A}_s}^{1/s} \left(\sum_{i=1}^{k} \alpha^{\frac{i}{s}} \right) \Big(|\!|\!| u - U_k |\!|\!|_\Omega^2 + \gamma \operatorname{osc}_{\mathcal{T}_k}^2(f, \mathcal{W}_k) \Big)^{-1/2s} \\
&\lesssim |u|_{\mathcal{A}_s}^{1/s} \Big(|\!|\!| u - U_k |\!|\!|_\Omega^2 + \operatorname{osc}_{\mathcal{T}_k}^2(f, \mathcal{W}_k) \Big)^{-1/2s}.
\end{aligned}$$

Here we used that $\alpha < 1$, so that it forms a geometric series. Standard algebra calculations on this inequality provide us with the desired result. □

Compare this result to the definition of the approximation class $\mathcal{A}_s$. The best possible error $e(N)$ on $\mathbb{T}_N$ satisfies $e(N) \leqslant N^{-s} |u|_{\mathcal{A}_s}$, see (3.12). By the previous theorem, there exists a constant C_{opt} such that the error made by AFEM is bound by

$$\Big(|\!|\!| u - U_k |\!|\!|_\Omega^2 + \operatorname{osc}_k^2(f, \mathcal{W}_k) \Big)^{1/2} \leqslant C_{opt} \left(\#\mathcal{T}_k - \#\mathcal{T}_0 \right)^{-s} |u|_{\mathcal{A}_s}.$$

The total error for the discrete solutions U_k constructed by AFEM decays at the same rate as the total error of the best possible solution. Hence AFEM driven by the equilibrated flux estimator is optimal.

3.5. Discussion

A few remarks can be added to the above optimality proof.

3.5.1. Interior node property

In the current proof we need the interior node property which requires `REFINE` to bisect the marked patches at least three times. In practice one often wants to avoid this many refinements; one would like to just bisect every marked patch once. However, by letting `REFINE` apply single bisections we cannot guarantee that the discrete efficiency holds, which is needed to prove the contraction property of AFEM solutions.

Cascón and Nochetto [12] reduce the number of bisections by proposing a variant of the `REFINE` method that ensures the interior node property after a fixed amount of steps. Although this requires less bisections in one step, it does require a non-standard `REFINE` implementation.

Kreuzer and Siebert [22] circumvent the interior node property altogether by using equivalence with the standard residual estimator and by requiring that the error estimator dominates the oscillation, i.e. $\eta_{\mathcal{T}}(U,\omega_a) \geqslant \operatorname{osc}_{\mathcal{T}}(f,\omega_a)$. In the equilibrated flux estimator this does not hold for a general f. However, if one considers a Poisson problem without oscillation (or with a very smooth f), then this does hold. For such problems one can therefore obtain an optimal convergence rate using a `REFINE` method that applies only a single bisection to the marked patches.

3.5.2. Element-wise estimator

We have proven optimality of AFEM driven by the patch-wise estimators $\|\underline{\sigma_a}\|_{\omega_a}$. As noted before, we ideally would use the element-wise estimators $\|\underline{\sigma}^{\triangle}\|_K$. We have proven that this element-wise variant $\|\underline{\sigma}^{\triangle}\|_K$ is reliable and efficient in Theorems 2.9 and 2.10. We are left to prove discrete reliability and discrete efficiency from Theorem 3.1.

Discrete reliability can proven using the Scott-Zhang local interpolant. For simplicity, assume that we do not have data oscillation. With $E_\star = U_\star - U$ and $V \in \mathbb{V}(\mathcal{T})$, we deduce from Galerkin orthogonality that

$$|||E_\star|||_\Omega^2 = a(E_\star, E_\star - V) = a(u-U, E_\star - V) = \langle r, E_\star - V\rangle = \langle \underline{\sigma}^{\triangle}, \nabla V - \nabla E_\star\rangle_\Omega.$$

The second equality holds since $U_\star$ is the Galerkin approximation on $\mathbb{V}(\mathcal{T}_\star) \ni E_\star - V$. The last equality follows from the equilibrated characteristic, i.e. $\langle \underline{\sigma}^{\triangle}, \nabla v\rangle = -\langle r, v\rangle$. Let V be the Scott-Zhang [32] interpolator of $E_\star$, so that $V - E_\star = 0$ on $\mathcal{T} \cap \mathcal{T}_\star$. From the Scott-Zhang approximation theory [32], and Poincaré-Friedrichs inequality for $E_\star \in H^1_0(\Omega)$ we discover that

$$\begin{aligned}
\langle \underline{\sigma}^{\triangle}, \nabla V - \nabla E_\star\rangle_\Omega = \sum_{K\in\mathcal{T}\setminus\mathcal{T}_\star} \langle \underline{\sigma}^{\triangle}, \nabla V - \nabla E_\star\rangle_K &\leqslant \sum_{K\in\mathcal{T}\setminus\mathcal{T}_\star} \|\underline{\sigma}^{\triangle}\|_K \|\nabla V - \nabla E_\star\|_K \\
&\lesssim \sum_{K\in\mathcal{T}\setminus\mathcal{T}_\star} \|\underline{\sigma}^{\triangle}\|_K \|E_\star\|_{H^1(\omega_K)} \\
&\lesssim \|E_\star\|_{H^1(\Omega)} \sqrt{\sum_{K\in\mathcal{T}\setminus\mathcal{T}_\star} \|\underline{\sigma}^{\triangle}\|_K^2} \\
&\lesssim |||E_\star|||_\Omega \sqrt{\sum_{K\in\mathcal{T}\setminus\mathcal{T}_\star} \|\underline{\sigma}^{\triangle}\|_K^2}.
\end{aligned}$$

This gives the desired discrete upper bound

$$|||U_\star - U|||_\Omega^2 \lesssim \sum_{K\in\mathcal{T}\setminus\mathcal{T}_\star} \|\underline{\sigma}^{\triangle}\|_K^2.$$

For the discrete efficiency no such direct proof comes to mind. In Theorem 3.1 we proved discrete efficiency using an equivalence with the standard residual estimator. It is not directly clear if such an equivalence also holds on elements for $\|\underline{\sigma}^{\triangle}\|_K$. However, from the triangle inequality we deduce

$$\|\underline{\sigma}^{\triangle}\|_K \leqslant \sum_{a\in\mathcal{V}_K} \|\underline{\sigma_a}\|_{\omega_a}.$$

For the latter terms we may invoke the patch-wise equivalence with the residual estimator. Because of this similarities, we *conjecture* that discrete efficiency also holds for the element-wise estimator $\|\underline{\sigma}^{\triangle}\|_K$. Together with discrete reliability, this would imply optimal convergence of AFEM driven by the element-wise equilibrated flux estimator.

3.5.3. Lower order estimator

In §2.5.2 we introduced an equilibrated flux estimator $\underline{\hat{\sigma}_a}$ in the lower order Raviart-Thomas $\mathcal{RT}^{-1}_{p-1,0}(\omega_a)$. The element-wise estimator $\|\underline{\hat{\sigma}}\|_K$ provides a reliability bound for which the oscillation is still of higher order. As with the above, we conjecture that AFEM driven by this estimator is optimal as well. This claim is supported by the numerical results in Chapter 5.

4. Practical aspects

The (theoretical) results from the previous chapters show the potency of the equilibrated flux estimator. Most results are upper- or lower bounds and some of these still depend on unknown constants. To address these uncertainties we would like to measure the performance in practice. Construction of $\underline{\sigma_a}$ from Definition 2.1 requires one to solve a constrained optimization problem over the broken Raviart-Thomas space $\mathcal{RT}_{p,0}^{-1}(\omega_a)$. An equivalent construction using the smaller Raviart-Thomas space $\mathcal{RT}_{p,0}(\omega_a)$ is given by Ern and Vohralík in [19]. In this chapter we will prove this equivalence and give some other implementational details.

4.1. Ern and Vohralík's construction

The equilibration method presented in the previous chapter is based on constructing the difference flux $\underline{\sigma}^{\triangle} = \nabla U - \underline{\sigma}$, with $\underline{\sigma}^{\triangle}$ from the broken Raviart-Thomas space $\mathcal{RT}_p^{-1}(\mathcal{T})$. Ern and Vohralík [19] propose constructing the actual flux $\underline{\sigma} \in \mathcal{RT}_p(\mathcal{T})$ instead of the difference.

We adapt the notation of [19], i.e. we locally define $\underline{\zeta_a} := \underline{\sigma_a} - \psi_a \nabla U$ and globally write $\underline{\zeta} := \sum_{a \in \mathcal{V}} \underline{\zeta_a}$. By the partition of unity we have $\underline{\zeta} = \underline{\sigma}^{\triangle} - \nabla U$, so that one retrieves a reliable and efficient estimator after replacing $\underline{\sigma}^{\triangle}$ with $\underline{\zeta} + \nabla U$ in Theorems 2.9 and 2.10. Rewriting the characteristic property (2.9) of a locally equilibrated flux $\underline{\sigma_a}$ reveals

$$\begin{aligned} -\langle \underline{\sigma_a}, \nabla v \rangle_{\omega_a} = \langle \tilde{r}_a, v \rangle &= \langle \Pi_{p-1}^a(\psi_a f), v \rangle_{\omega_a} - \langle \nabla U, \nabla (\psi_a v) \rangle_{\omega_a} \\ &= \langle \Pi_{p-1}^a(\psi_a f) - \nabla \psi_a \cdot \nabla U, v \rangle_{\omega_a} - \langle \psi_a \nabla U, \nabla v \rangle_{\omega_a} \quad \forall v \in H_\star^1(\omega_a). \end{aligned}$$

This holds if and only if for all $v \in H_\star^1(\omega_a)$ we have

$$\langle \Pi_{p-1}^a(\psi_a f) - \nabla \psi_a \cdot \nabla U, v \rangle_{\omega_a} = -\langle \underline{\sigma_a} - \psi_a \nabla U, \nabla v \rangle_{\omega_a} = -\langle \underline{\zeta_a}, \nabla v \rangle_{\omega_a}. \tag{4.1}$$

Notice[1] that the weak divergence of $\underline{\zeta_a}$ apparently satisfies $\operatorname{div} \underline{\zeta_a} = \Pi_{p-1}^a(\psi_a f) - \nabla \psi_a \cdot \nabla U$. The latter is a $L^2(\omega_a)$ function, and since $\underline{\sigma_a}, \psi_a \nabla U \in \mathcal{RT}_{p,0}^{-1}(\omega_a)$ we conclude

$$\underline{\sigma_a} - \psi_a \nabla U = \underline{\zeta_a} \in H(\operatorname{div}; \omega_a) \cap \mathcal{RT}_{p,0}^{-1}(\omega_a) = \mathcal{RT}_{p,0}(\omega_a).$$

Constructing $\underline{\sigma_a} - \psi_a \nabla U$ is therefore equivalent to finding $\underline{\zeta_a} \in \mathcal{RT}_p(\omega_a)$ that satisfies $\operatorname{div} \underline{\zeta_a} = \Pi_{p-1}^a(\psi_a f) - \nabla \psi_a \cdot \nabla U$ with minimal $\|\underline{\zeta_a} + \psi_a \nabla U\|_{\omega_a}$. This latter problem can be turned into a system of equations.

[1] For $a \in \mathcal{V}^{bdr}$ it is clear that $C_c^\infty(\omega_a) \subset H_\star^1(\omega_a)$. For $a \in \mathcal{V}^{int}$, take $v \in C_c^\infty(\omega_a)$ and consider $v - v_{\omega_a} \in H_\star^1(\omega_a)$. Plugging $v - v_{\omega_a}$ into the equation (4.1) and using orthogonality relations shows that actually $-\langle \underline{\zeta_a}, \nabla v \rangle = \langle \Pi_{p-1}^a(\psi_a f) - \nabla \psi_a \cdot \nabla U, v \rangle_{\omega_a}$ for $v \in C_c^\infty(\omega_a)$.

Theorem 4.1. *For vertex $a \in \mathcal{V}$ the flux $\underline{\zeta_a} \in \mathcal{RT}_{p,0}(\omega_a)$ can be found by solving*

$$\langle \underline{\zeta_a}, \underline{\tau} \rangle_{\omega_a} - \langle \operatorname{div} \underline{\tau}, \lambda_a \rangle_{\omega_a} = -\langle \psi_a \nabla U, \underline{\tau} \rangle_{\omega_a} \qquad \forall \underline{\tau} \in \mathcal{RT}_{p,0}(\omega_a), \tag{4.2a}$$

$$\langle \operatorname{div} \underline{\zeta_a}, q_a \rangle_{\omega_a} = \langle \psi_a f - \nabla \psi_a \cdot \nabla U, q_a \rangle_{\omega_a} \quad \forall q_a \in \mathcal{Q}_p(\omega_a), \tag{4.2b}$$

for the pair $(\underline{\zeta_a}, \lambda_a) \in \mathcal{RT}_{p,0}(\omega_a) \times \mathcal{Q}_p(\omega_a)$, with $\mathcal{Q}_p(\omega_a)$ a polynomial subspace given by

$$\mathcal{Q}_p(\omega_a) := \begin{bmatrix} \left\{ q \in \mathcal{P}_p^{-1}(\omega_a) : q_{\omega_a} = 0 \right\} & a \in \mathcal{V}^{int}, \\ \left\{ q \in \mathcal{P}_p^{-1}(\omega_a) \right\} & a \in \mathcal{V}^{bdr}. \end{bmatrix}$$

Proof. We will start by showing that the characteristic property (4.1) of $\underline{\zeta_a}$ is equivalent to (4.2b). Here we will use that Π_{p-1}^a is the L_2-orthogonal projector on the broken polynomial space $\mathcal{P}_p^{-1}(\omega_a)$. Fix $v \in H_\star^1(\omega_a)$, since $\underline{\zeta_a} \in \mathcal{RT}_{p,0}(\omega_a)$ we know that $\operatorname{div} \underline{\zeta_a} \in \mathcal{P}_p^{-1}(\omega_a)$, and thus that $\Pi_{p-1}^a(\operatorname{div} \underline{\zeta_a}) = \operatorname{div} \underline{\zeta_a}$. Combined with orthogonality of the projector, the right hand side of (4.1) reads as

$$-\langle \underline{\zeta_a}, \nabla v \rangle_{\omega_a} = \langle \operatorname{div} \underline{\zeta_a}, v \rangle_{\omega_a} = \langle \Pi_{p-1}^a(\operatorname{div} \underline{\zeta_a}), v \rangle_{\omega_a} = \langle \operatorname{div} \underline{\zeta_a}, \Pi_{p-1}^a(v) \rangle_{\omega_a}.$$

Something similar for the left hand side of (4.1) holds. The function $\nabla \psi_a \cdot \nabla U$ is contained in $\mathcal{P}_p^{-1}(\omega_a)$, since $\psi_a \in \mathcal{P}_1(\omega_a)$ and $U \in \mathcal{P}_p(\omega_a)$. Similar steps as before give,

$$\begin{aligned} \langle \Pi_{p-1}^a(\psi_a f) - \nabla \psi_a \cdot \nabla U, v \rangle_{\omega_a} &= \langle \Pi_{p-1}^a(\psi_a f) - \Pi_{p-1}^a \left(\nabla \psi_a \cdot \nabla U \right), v \rangle_{\omega_a} \\ &= \langle \Pi_{p-1}^a \left(\psi_a f - \nabla \psi_a \cdot \nabla U \right), v \rangle_{\omega_a} \\ &= \langle \psi_a f - \nabla \psi_a \cdot \nabla U, \Pi_{p-1}^a(v) \rangle_{\omega_a}. \end{aligned}$$

From this we conclude that (4.1) is equivalent to

$$\langle \operatorname{div} \underline{\zeta_a}, q_a \rangle_{\omega_a} = \langle \psi_a f - \nabla \psi_a \cdot \nabla U, q_a \rangle \quad \forall q \in \Pi_{p-1}^a \left(H_\star^1(\omega_a) \right). \tag{4.3}$$

A density argument can be used to show that $\Pi_{p-1}^a \left(H_\star^1(\omega_a) \right) = \mathcal{Q}_p(\omega_a)$, for both interior and boundary vertices a. This completes the equivalence of (4.1) and (4.2b).

Next, we tackle rewriting the minimizing property. By the polarization identity

$$\|\underline{\zeta_a} + \psi_a \nabla U\|_{\omega_a}^2 = \|\underline{\zeta_a}\|_{\omega_a}^2 + \|\psi_a \nabla U\|_{\omega_a}^2 + 2 \langle \underline{\zeta_a}, \psi_a \nabla U \rangle_{\omega_a}.$$

Now $\psi_a \nabla U$ is fixed, so minimizing $\|\underline{\zeta_a} + \psi_a \nabla U\|_{\omega_a}$ for $\underline{\zeta_a}$ is equivalent to minimizing

$$\frac{1}{2} \|\underline{\zeta_a}\|_{\omega_a}^2 + \langle \underline{\zeta_a}, \psi_a \nabla U \rangle_{\omega_a}.$$

Together with the previous paragraph we conclude that $\underline{\zeta_a}$ can be found as a solution of (4.3) with minimal $\frac{1}{2} \|\underline{\zeta_a}\|_{\omega_a}^2 + \langle \underline{\zeta_a}, \psi_a \nabla U \rangle_{\omega_a}$.

The optimality conditions of the above minimisation problem are given by the system (4.2). To see this, we require some evolved optimisation theory. Finding the pair $(\underline{\zeta_a}, \lambda_a) \in \mathcal{RT}_{p,0}(\omega_a) \times \mathcal{Q}_p(\omega_a)$ that solves the system (4.2) is actually a *saddle point*

problem (cf. [7, 8]). That is, it fits in the general framework of finding $(u, \lambda) \in X \times M$ for Hilbert spaces X and M with

$$\begin{aligned} a(u, v) + b(v, \lambda) &= \langle f, v\rangle \quad \forall v \in X, \\ b(u, \mu) &= \langle g, \mu\rangle \quad \forall \mu \in M, \end{aligned}$$

for $f \in X', g \in M'$ and *continuous* bilinear forms $a : X \times X \to \mathbb{R}, b : X \times M \to \mathbb{R}$. The main result about saddle point problems [7, Thm 4.2.1] provides sufficient conditions under which there is a unique solution (u, λ) to the saddle point problem such that u is also the unique minimizer of $\frac{1}{2}a(u, u) - \langle f, u\rangle$. Define the operator $B : X \to M' : v \mapsto b(v, \cdot)$. The first condition is that the bilinear form $a(\cdot, \cdot)$ is symmetric and coercive on the set of functions $V := \{v \in X : Bv = 0\}$. Secondly, the operator B should be surjective.

Apply this general framework to our system (4.2); we have $X = \mathcal{RT}_{p,0}(\omega_a)$ and $M = \mathcal{Q}_p(\omega_a)$ with bilinear forms

$$a(\underline{\zeta_a}, \underline{\tau}) = \langle \underline{\zeta_a}, \underline{\tau}\rangle_{\omega_a} \quad \text{and} \quad b(\underline{\zeta_a}, q_a) = \langle \operatorname{div} \underline{\zeta_a}, q_a\rangle_{\omega_a} \quad \text{for } \underline{\zeta_a}, \underline{\tau} \in \mathcal{RT}_{p,0}(\omega_a), q_a \in \mathcal{Q}_p(\omega_a).$$

The flux space $\mathcal{RT}_{p,0}(\omega_a)$ is equipped with the $H^1(\operatorname{div}; \omega_a)$-norm: $\|\underline{\tau}\|^2_{H^1(\operatorname{div};\omega_a)} = \|\underline{\tau}\|^2_{\omega_a} + \|\operatorname{div} \underline{\tau}\|^2_{\omega_a}$, cf. Definition B.1. One easily sees that the above bilinear forms are continuous. Let $\underline{\tau} \in V$, then by definition $\langle \operatorname{div} \underline{\tau}, q_a\rangle_{\omega_a} = 0$ for all $q_a \in \mathcal{Q}_p(\omega_a)$, which implies that $\operatorname{div} \underline{\tau} = 0$, since $\operatorname{div} \underline{\tau} \in \mathcal{Q}_p(\omega_a)$. Coercivity of $a(\cdot, \cdot)$ on V therefore follows easily:

$$a(\underline{\tau}, \underline{\tau}) = \langle \underline{\tau}, \underline{\tau}\rangle_{\omega_a} = \|\underline{\tau}\|^2_{\omega_a} = \|\underline{\tau}\|^2_{\omega_a} + \|\operatorname{div} \underline{\tau}\|^2_{\omega_a} = \|\underline{\tau}\|^2_{H(\operatorname{div};\omega_a)}.$$

Given $q_a \in \mathcal{Q}_p(\omega_a)$, there exists a $\underline{\tau_a} \in \mathcal{RT}_{p,0}(\omega_a)$ such that $\operatorname{div} \underline{\tau_a} = q_a$ (cf. proof of Theorem 2.2). By duality of Q, this turns B into a surjective map. All in all, we satisfy the necessary conditions and hence we may conclude from [7, Thm 4.2.1] that there is a unique pair $(\underline{\zeta_a}, \lambda) \in \mathcal{RT}_{p,0}(\omega_a) \times \mathcal{Q}_p(\omega_a)$ that solves the saddle point system (4.2). From [7, Rem 4.2.1] we deduce that $\underline{\zeta_a}$ is the unique minimizer of $\frac{1}{2}\|\underline{\zeta_a}\|^2_{\omega_a} + \langle \underline{\zeta_a}, \psi_a \nabla U\rangle_{\omega_a}$ subject to equation (4.2b). □

Since both function spaces in the system (4.2) are finite dimensional, the system can be reduced to a finite set of equations over the bases of $\mathcal{RT}_{p,0}(\omega_a)$ and $\mathcal{Q}_p(\omega_a)$. This provides a straightforward algorithm for constructing $\underline{\zeta_a}$. Using the correspondence with $\underline{\sigma_a}$, we directly find an upper bound (cf. Theorem 2.9). Moreover, for the data oscillation we will show that $\|(I - \Pi_p)(f)\|_K = \|f - \operatorname{div} \underline{\zeta}\|_K$.

Theorem 4.2. *Let $\underline{\zeta_a}$ be found by solving* (4.2)*, then for $\underline{\zeta} = \sum_{a \in \mathcal{V}} \underline{\zeta_a}$ we have*

$$\begin{aligned} |||u - U|||^2_\Omega &\leqslant \sum_{K \in \mathcal{T}} \left[\|\underline{\zeta} + \nabla U\|_K + \frac{h_K}{\pi} \|(I - \Pi_p)(f)\|_K \right]^2, \\ &= \sum_{K \in \mathcal{T}} \left[\|\underline{\zeta} + \nabla U\|_K + \frac{h_K}{\pi} \|f - \operatorname{div} \underline{\zeta}\|_K \right]^2. \end{aligned}$$

Proof. The upper bound follows from Theorem 2.9, since $\underline{\sigma}^{\triangle} = \underline{\zeta_a} + \nabla U$.

We will show that $(f - \operatorname{div} \underline{\zeta})|_K$ is perpendicular to $\mathcal{P}_p(K)$ on every element K. For an interior vertex $a \in \mathcal{V}^{int}$, we find from the divergence theorem that

$$\langle \operatorname{div} \underline{\zeta_a}, 1\!\!1 \rangle_{\omega_a} = \langle \underline{\zeta_a} \cdot n, 1\!\!1 \rangle_{\partial\omega_a} - \langle \underline{\zeta_a}, \nabla 1\!\!1 \rangle_{\omega_a} = 0,$$

since $\underline{\zeta_a} \in \mathcal{RT}_{p,0}(\omega_a)$ ensures that $\underline{\zeta_a} \cdot n = 0$ on $\partial\omega_a$. Now pick $q \in \mathcal{P}_p^{-1}(\omega_a)$, then $q - q_{\omega_a} \in \mathcal{Q}_p(\omega_a)$ and thus from the above and (4.2b) we deduce

$$\begin{aligned} \langle \operatorname{div} \underline{\zeta_a}, q \rangle_{\omega_a} &= \langle \operatorname{div} \underline{\zeta_a}, q - q_{\omega_a} \rangle_{\omega_a} \\ &= \langle \psi_a f - \nabla \psi_a \cdot \nabla U, q - q_{\omega_a} \rangle_{\omega_a} = \langle \psi_a f - \nabla \psi_a \cdot \nabla U, q \rangle_{\omega_a}, \end{aligned}$$

with the last equality following from the Galerkin property, since $\psi_a \in \mathbb{V}(\mathcal{T})$.

This shows that (4.2b) holds with the test space $\mathcal{P}_p^{-1}(\omega_a)$ for all vertices $a \in \mathcal{V}$. Now pick a polynomial $q \in \mathcal{P}_p(K)$, which we implicitly extend to zero outside K. Since $\mathcal{P}_p^{-1}(\omega_a)$ consists of broken polynomials, we see that $q \in \mathcal{P}_p^{-1}(\omega_a)$ for all patches ω_a. Calculating the inner product with $f - \operatorname{div} \underline{\zeta}$ yields

$$\begin{aligned} \langle f - \operatorname{div} \underline{\zeta}, q \rangle_K &= \sum_{a \in \mathcal{V}_K} \langle \psi_a f - \operatorname{div} \underline{\zeta_a}, q \rangle_K = \sum_{a \in \mathcal{V}_K} \langle \psi_a f - \operatorname{div} \underline{\zeta_a}, q \rangle_{\omega_a} \\ &= \sum_{a \in \mathcal{V}_K} \langle \nabla \psi_a \cdot \nabla U, q \rangle_{\omega_a} \\ &= \langle \nabla 1\!\!1 \cdot \nabla U, q \rangle_K = 0, \end{aligned}$$

where the third inequality follows from (4.2b). We conclude that $(f - \operatorname{div} \underline{\zeta})|_K \perp \mathcal{P}_p(K)$, and because $\operatorname{div} \underline{\zeta}|_K \in \mathcal{P}_p(K)$, we infer that $\operatorname{div} \underline{\zeta} = \Pi_p f$. □

4.2. Raviart-Thomas space

The system (4.2) provides us with concrete steps for an actual implementation of the equilibrated flux estimator. In particular, we must find bases for $\mathcal{RT}_{p,0}(\omega_a)$ and $\mathcal{Q}_p(\omega_a)$. We will give a short intermezzo deriving some theory for this first space.

Consider an arbitrary domain $\Omega \subset \mathbb{R}^2$ with triangulation $\mathcal{T}$, vertices $\mathcal{V}$, et cetera. We will start by examining the Raviart-Thomas space $\mathcal{RT}_p(\mathcal{T})$, without the boundary conditions, generated by the Raviart-Thomas element.

4.2.1. Raviart-Thomas element

Recall from §1.2 that finite elements are formally represented by a triplet $(K, \mathcal{P}, \mathcal{N})$. For the construction of the Raviart-Thomas space $\mathcal{RT}_p(\mathcal{T})$ we obviously have $\mathcal{P} := \mathcal{RT}_p(K)$. We are left with the task of finding suitable nodal variables $\mathcal{N}$ that ensure $H(\operatorname{div}; \Omega)$-conformity of the entire space. For this latter constraint we use the following characterization of $\mathcal{RT}_p(\mathcal{T})$.

Theorem 4.3. *The Raviart-Thomas space $\mathcal{RT}_p(\mathcal{T})$ can be characterized by*

$$\mathcal{RT}_p(\mathcal{T}) := H(\mathrm{div};\Omega) \cap \mathcal{RT}_p^{-1}(\mathcal{T}) = \left\{ \underline{\sigma} \in \mathcal{RT}_p^{-1}(\mathcal{T}) : [\![\underline{\sigma}]\!] = 0 \ \text{ in } L^2(e) \quad \forall e \in \mathcal{E}^{int} \right\}.$$

Proof. We show the inclusion "$\supset$", that is, let $\underline{\sigma} \in \mathcal{RT}_p^{-1}(\mathcal{T})$ be given such that $[\![\underline{\sigma}]\!]$ vanishes on all interior edges. We have to show that $\underline{\sigma}$ has a weak divergence in $L^2(\Omega)$. Let $\varphi \in D(\Omega)$ be given, from the divergence theorem for $\underline{\sigma}|_K \in H^1(\mathrm{div}; K)$ we infer that

$$\begin{aligned} -\int_\Omega \nabla\varphi \cdot \underline{\sigma} &= -\sum_{K\in\mathcal{T}} \int_K \nabla\varphi \cdot \underline{\sigma} \\ &= \sum_{K\in\mathcal{T}} \left[\int_K \varphi \cdot \mathrm{div}\, \underline{\sigma} - \int_{\partial K} \varphi \underline{\sigma} \cdot n \right] \\ &= \int_\Omega \varphi \cdot \mathrm{div}\, \underline{\sigma} - \sum_{e\in\mathcal{E}^{int}} \int_e \varphi [\![\underline{\sigma}]\!] = \int_\Omega \varphi \cdot \mathrm{div}\, \underline{\sigma}. \end{aligned}$$

Here we used that φ vanishes on the boundary, whilst $[\![\underline{\sigma}]\!]$ vanishes on interior edges. The result follows since $\mathrm{div}\, \underline{\sigma} \in \mathcal{P}_p^{-1}(\Omega) \subset L^2(\Omega)$, hence we may conclude that $\underline{\sigma} \in H(\mathrm{div};\Omega)$.

The other inclusion follows similarly (cf. [20, Thm 3.2]). □

For a triangle $K \in \mathcal{T}$ we have $\mathcal{RT}_p(K) := [\mathcal{P}_p(K)]^2 + \mathcal{P}_p(K)\underline{x}$. Let $\widetilde{\mathcal{P}}_p(K)$ denote polynomials of *exactly* degree p, then one easily proves that $\mathcal{RT}_p(K) = [\mathcal{P}_p(K)]^2 \oplus \widetilde{\mathcal{P}}_p(K)\underline{x}$. From this decomposition it follows that

$$\dim \mathcal{RT}_p(K) = 2\dim \mathcal{P}_p(K) + \dim \widetilde{\mathcal{P}}_p(K) = 2\binom{2+p}{p} + \binom{2+p-1}{p} = (p+1)(p+3).$$

The Raviart-Thomas space $\mathcal{RT}_p(K)$ has the following (unisolvent) characterisation.

Theorem 4.4. *Let $K \in \mathcal{T}$ and $\underline{\sigma} \in \mathcal{RT}_p(K)$ be given such that*

$$\begin{aligned} \langle \underline{\sigma} \cdot n_e, q \rangle_e = 0 \quad & \forall q \in \mathcal{P}_p(e), && \forall \text{edge } e \subset \partial K, \\ \langle \underline{\sigma}, \underline{\tau} \rangle_K = 0 \quad & \forall \underline{\tau} \in [\mathcal{P}_{p-1}(K)]^2, && p \geqslant 1. \end{aligned}$$

Then $\underline{\sigma} \equiv 0$ in K.

Proof. Examination of the dimensions of the function spaces reveals that the number of equations equals the dimension of $\mathcal{RT}_p(K)$. Since $\underline{\sigma} \cdot n_e \in \mathcal{P}_p(e)$, the first equation implies that $\underline{\sigma} \cdot n_e$ vanishes on all edges of K.

Let $p \geqslant 1$, then for any $q \in \mathcal{P}_p(K)$ we have $\nabla q \in [\mathcal{P}_{p-1}(K)]^2$, and by the divergence theorem this gives that $\langle \mathrm{div}\, \underline{\sigma}, q \rangle_K = -\langle \underline{\sigma}, \nabla q \rangle_K + \langle \underline{\sigma} \cdot n, q \rangle_{\partial K} = 0$. Since $\mathrm{div}\, \underline{\sigma} \in P_p(K)$, we may conclude that $\mathrm{div}\, \underline{\sigma} = 0$. Decompose $\underline{\sigma} = \underline{q} + q_0 \underline{x}$ for $\underline{q} = [q_1, q_2]^\top$ with $q_1, q_2 \in \mathcal{P}_p(K)$ and $q_0 \in \widetilde{\mathcal{P}}_p(K)$. From the fact that $\mathrm{div}\, \underline{\sigma} = 0$, we infer that

$$0 = \mathrm{div}\, \underline{\sigma} = \mathrm{div}\, \underline{q} + \mathrm{div}\, (q_0 \underline{x}) = \mathrm{div}\, \underline{q} + (2+p) q_0 \implies (2+p) q_0 = -\mathrm{div}\, \underline{q} \in \mathcal{P}_{p-1}(K).$$

Since $q_0 \in \widetilde{\mathcal{P}}_p(K)$ we must have $q_0 \equiv 0$ and thus $\underline{\sigma} = \underline{q} \in [\mathcal{P}_p(K)]^2$.

Notice that $\underline{\sigma} \cdot n_e = \underline{q} \cdot n_e$ is a polynomial of degree p on the whole of K. For any polynomial $g \in \mathcal{P}_{p-1}(K)$, we deduce from the second assumption that

$$\langle \underline{q} \cdot n_e, g \rangle_K = \int_K g \left(\underline{q} \cdot n_e \right) = \int_K \underline{q} \cdot \underline{\tilde{g}} = \langle \underline{\sigma}, \underline{\tilde{g}} \rangle_K = 0 \quad \text{for } \underline{\tilde{g}} = [g n_e^1, g n_e^2]^\top \in \mathcal{P}_{p-1}(K).$$

On the other hand, for some edge $e \in \partial K$ we know that $\underline{q} \cdot n_e$ vanishes. Therefore, we can decompose $\underline{q} \cdot n_e = Lg$ in a polynomial $g \in \mathcal{P}_{p-1}$ and L the hyperplane of e [11, Lem 3.1.10]. Using this polynomial g in the above identity then shows

$$0 = \langle \underline{q} \cdot n_e, g \rangle_K = \langle Lg, g \rangle_K = \langle L, g^2 \rangle_K.$$

Since L does not vanish on K, we conclude that $g = 0$, and thus that $\underline{q} \cdot n_e$ vanishes on the entire element K. This holds for all edge normals n_e, hence we infer that $\underline{\sigma} = \underline{q} = 0$. □

The number of equations in the previous theorem equals $\dim \mathcal{RT}_p(K)$, and therefore provides us with a set of nodal variables. That is, for each edge e let $\{q_{e,1}, \dots, q_{e,m}\}$ be a basis of $\mathcal{P}_p(e)$, and let $\{\underline{\tau}_1, \dots, \underline{\tau}_r\}$ be a basis for $\left[\mathcal{P}_{p-1}(K)\right]^2$. Define linear functionals on $\mathcal{RT}_p(K)$ associated to edges and elements by

$$\begin{aligned} N_{e,i}(\underline{\sigma}) &= \langle \underline{\sigma} \cdot n_e, q_{e,i} \rangle_e \quad && 1 \leqslant i \leqslant m, e \subset \partial K \\ N_{K,j}(\underline{\sigma}) &= \langle \underline{\sigma}, \underline{\tau_j} \rangle_K && 1 \leqslant j \leqslant r. \end{aligned} \tag{4.4}$$

The set $\mathcal{N}$ containing all of the above functionals forms a basis for $\mathcal{RT}_p(K)'$, and thus turns the triplet $(K, \mathcal{RT}_p(K), \mathcal{N})$ into a valid finite element.

4.2.2. Raviart-Thomas basis

We have formally introduced an Raviart-Thomas element $(K, \mathcal{RT}_p(K), \mathcal{N})$. How can we glue these elements together to form a basis for the Raviart-Thomas space $\mathcal{RT}_p(\mathcal{T})$? In particular we must ensure $H(\mathrm{div}; \Omega)$-conformity.

For simplicity, we start with two neighbouring Raviart-Thomas elements $K_1, K_2 \in \mathcal{T}$ having a common edge $e = K_1 \cap K_2$. From the characterization in Theorem 4.3 we know that

$$\mathcal{RT}_p(K_1 \cup K_2) = \left\{ \underline{\sigma} \in \mathcal{RT}_p^{-1}(K_1 \cup K_2) : [\![\underline{\sigma}]\!] = 0 \text{ in } L_2(e) \right\}.$$

Pick some $\underline{\sigma} \in \mathcal{RT}_p^{-1}(K_1 \cup K_2)$ and let n_e be the unit outward normal on e for element K_1, then the conformity condition can be rewritten as

$$0 = [\![\underline{\sigma}]\!] = \underline{\sigma}|_{K_2} \cdot n_e - \underline{\sigma}|_{K_1} \cdot n_e.$$

Since $\underline{\sigma}|_{K_1} \cdot n_e$ and $\underline{\sigma}|_{K_2} \cdot n_e$ are both elements in $\mathcal{P}_p(e)$, we see that the above condition is equivalent to requiring

$$\langle \underline{\sigma}|_{K_2} \cdot n_e, q \rangle = \langle \underline{\sigma}|_{K_1} \cdot n_e, q \rangle \quad \forall q \in \mathcal{P}_p(e).$$

This requirement is satisfied if and only if the edge-related nodal variables of K_1 and K_2 coincide for $\underline{\sigma}$.

The above observation is easily extended to the entire Raviart-Thomas space $\mathcal{RT}_p(\mathcal{T})$: the space $\mathcal{RT}_p(\mathcal{T})$ consists of functions from the broken space $\mathcal{RT}_p^{-1}(\mathcal{T})$ for which all the edge-related nodal variables of two adjoint elements coincide. For each element $K \in \mathcal{T}$, we let $\phi_{e,i}^K$ and $\phi_{K,j}^K$ denote the nodal basis with respect to (4.4). A global nodal basis is then found by glueing together edge related functions. To be precise, the nodal basis for $\mathcal{RT}_p(\mathcal{T})$ is given by

$$\{\phi_{K,j}^K : K \in \mathcal{T}, 1 \leqslant j \leqslant r\} \cup \{\phi_{e,i}^{K^1} + \phi_{e,i}^{K^2} : K_1, K_2 \in \mathcal{T}, e = K_1 \cap K_2, 1 \leqslant i \leqslant m\}, \quad (4.5)$$

with all basis functions naturally extended from K to Ω.

For construction of the estimator from (4.2) we actually require the Raviart-Thomas subspace $\mathcal{RT}_{p,0}(\mathcal{T})$ as defined in (2.7). Functions in this subspace have some additional Neumann boundary conditions. Such boundary conditions are easily incorporated into the basis (4.5): one can just remove all the basis functions associated with vanishing Neumann edges.

4.3. Explicit lowest order basis

Armed with this theory about the Raviart-Thomas space, we will again look at the system of equations (4.2) that define $\underline{\zeta_a}$. For simplicity we start by investigating this system for the lowest order estimator ($p = 0$). We will derive explicit (nodal) basis functions for both the spaces used in (4.2), i.e. $\mathcal{RT}_{0,0}(\omega_a)$ and $\mathcal{Q}_0(\omega_a)$.

4.3.1. Raviart-Thomas space

The lowest order Raviart-Thomas space contains only edge related nodal variables, see (4.4). Let K be a triangle with vertices $\{v_1, v_2, v_3\}$ and label the opposite edges by $\{e_1, e_2, e_3\}$. A basis for $\mathcal{P}_0(e_i)$ is simply the constant function $\mathbb{1}$. Focus on the basis function $\underline{\phi_1} \in \mathcal{RT}_0(K)$ associated with e_1. Per definition this is the solution of

$$\int_{e_1} \underline{\phi_1} \cdot n_{e_1} = 1 \quad \text{and} \int_{e_2} \underline{\phi_1} \cdot n_{e_2} = 0 = \int_{e_3} \underline{\phi_1} \cdot n_{e_3}. \quad (4.6)$$

This system can be solved using a geometrical observations: vertex v_1 lies on both edges e_2 and e_3, and thus for every constant α we have $\alpha(\underline{x} - v_1) \cdot n_{e_2} = 0$ for $\underline{x} \in e_2$ and $\alpha(\underline{x} - v_1) \cdot n_{e_3} = 0$ for $\underline{x} \in e_3$. We see that $\alpha(\underline{x} - v_1)$ solves both vanishing conditions of (4.6). For the first condition, we note that $(\underline{x} - v_1) \cdot n_{e_3}$ with $\underline{x} \in e_1$ is constant and equal to the *height* h_1 of e_1. From the well known formula $\frac{1}{2} h_1 \operatorname{vol}(e_1) = \operatorname{vol}(K)$ we observe that $\alpha(\underline{x} - v_1)$ with $\alpha = \frac{\operatorname{vol}(e_1)}{2 \operatorname{vol}(K)}$ solves (4.6). The nodal basis for the Raviart-Thomas element $\mathcal{RT}_0(K)$ is therefore given by

$$\underline{\phi_i} = \frac{1}{2} \frac{\operatorname{vol}(e_1)}{\operatorname{vol}(K)} (\underline{x} - v_i) \quad \text{for} \quad i = 1, 2, 3.$$

Next we consider the entire Raviart-Thomas space $\mathcal{RT}_0(\mathcal{T})$. For this space all the degrees of freedom are associated to edges. For every shared edge e we must fix a *global*

orientation of the normal vector n_e. After fixing such an orientation we can (uniquely) identify the adjacent elements by K_+ and K_-; with K_+ the element for which n_e is an outward pointing normal vector. Similarly write v_+, v_- for the vertex opposite of e in K_+ resp K_-. The global basis function associated to e is given by

$$\underline{\phi_e}(\underline{x}) = \begin{cases} \frac{1}{2}\frac{\operatorname{vol}(e)}{\operatorname{vol}(K_+)}(\underline{x} - v_+) & \underline{x} \in K_+ \\ -\frac{1}{2}\frac{\operatorname{vol}(e)}{\operatorname{vol}(K_-)}(\underline{x} - v_-) & \underline{x} \in K_-. \end{cases} \tag{4.7}$$

The extra Neumann boundary constraints needed for $\mathcal{RT}_{0,0}$ are easily incorporated, one simple removes the basis functions associated to the vanishing Neumann edges.

4.3.2. Polynomial space $\mathcal{Q}_0(\omega_a)$

The system (4.2) defines the polynomial space $\mathcal{Q}_0(\omega_a) \subset \mathcal{P}_0^{-1}(\omega_a)$. For boundary vertices we can simply take the constant functions $\mathbb{1}$ on each of the elements as basis. For an interior vertex $a \in \mathcal{V}^{int}$, we have the additional mean zero constraint, i.e.

$$\mathcal{Q}_0(\omega_a) = \left\{ q \in \mathcal{P}_0^{-1}(\omega_a) : \langle q, 1\rangle_{\omega_a} = 0 \right\}.$$

Write $K_1, \ldots, K_n$ for the elements in the patch ω_a. A function $p \in \mathcal{P}_0^{-1}(\omega_a)$ is constant with value p_i on K_i. The mean zero condition translates into

$$0 = \langle p, \mathbb{1}\rangle_{\omega_a} = \sum_{i=1}^{n} \langle p_i, \mathbb{1}\rangle_{K_i} = \sum_{i=1}^{n} p_i \operatorname{vol}(K_i).$$

A basis for $\mathcal{Q}_0(\omega_a)$ is therefore given by $q_1, \ldots, q_{n-1}$ with

$$q_i(x) = \begin{cases} \operatorname{vol}(K_n) & x \in K_i \\ -\operatorname{vol}(K_i) & x \in K_n. \end{cases}$$

Unfortunately, this is not an orthogonal basis, but for our test purposes it suffices.

4.4. Implementation of the lowest order estimator

Lets focus on actually solving $\underline{\zeta_a}$ from the system (4.2). For simplicity we consider the lowest order estimator ($p = 0$) with f piecewise constant and we let U be the *linear* Lagrange finite element solution for some triangulation $\mathcal{T}$. Notice that U is a polynomial of degree 1 and thus of one order higher than the order of the estimator. In §2.5.2 we saw that such an estimator still provides reliability and efficiency.

Implementation of the estimator $\underline{\zeta_a}$ for $p = 0$ now follows quite easily. Starting with an interior vertex $a \in \mathcal{V}$ we must solve $(\underline{\zeta_a}, \lambda_a) \in \mathcal{RT}_{0,0}(\omega_a) \times \mathcal{Q}_0(\omega_a)$ from the system

$$\begin{aligned} \langle \underline{\zeta_a}, \underline{\tau}\rangle_{\omega_a} - \langle \operatorname{div} \underline{\tau}, \lambda_a\rangle_{\omega_a} &= -\langle \psi_a \nabla U, \underline{\tau}\rangle_{\omega_a} && \forall \underline{\tau} \in \mathcal{RT}_{0,0}(\omega_a), \\ \langle \operatorname{div} \underline{\zeta_a}, q_a\rangle_{\omega_a} &= \langle \psi_a f - \nabla\psi_a \cdot \nabla U, q_a\rangle_{\omega_a} && \forall q_a \in \mathcal{Q}_0(\omega_a). \end{aligned} \tag{4.8}$$

Let n be the number of triangles in ω_a. For the space $\mathcal{RT}_{0,0}(\omega_a)$ we have one degree of freedom associated to every interior edge, and thus n basis functions $\underline{\tau_i}$. The space $\mathcal{Q}_0(\omega_a)$ has $n-1$ basis functions q_k. To solve the above system we must calculate four type of basis interactions:

$$\langle \underline{\tau_i}, \underline{\tau_j} \rangle_{\omega_a},\ \langle \operatorname{div} \underline{\tau_i}, q_k \rangle_{\omega_a},\ \langle \psi_a \nabla U, \underline{\tau_i} \rangle_{\omega_a},\ \text{and}\ \langle \psi_a f - \nabla \psi_a \cdot \nabla U, q_k \rangle_{\omega_a}. \tag{4.9}$$

The standard finite element approach is to decompose these inner products over the elements K, and then use an affine transformation to reduce the calculations to interactions on one reference triangle $\hat{K}$. Unfortunately, these transformations do no preserve normal components; the result of an affine transformation does not necessarily preserve $H(\operatorname{div};\Omega)$-conformity. This can be fixed by using the Piola transformation [7, §2.1.3].

We opt not to follow this approach and instead solve the inner products by quadrature. Since f is piecewise constant, all of the terms in (4.9) are actually polynomials of degree less or equal 2. The edge-midpoint quadrature rule is exact for quadratic polynomials, and hence provides us with an easy method to *exactly* calculate the integrals. Denote the edge-midpoints of a triangle K by m_i, then the edge-quadrature is given by

$$\int_K p = \frac{\operatorname{vol}(K)}{3} \sum_{k=1}^{3} p(m_k) \quad \forall p \in \mathcal{P}_2(K).$$

Applying this to the first basis interaction in (4.9) gives

$$\langle \underline{\tau_i}, \underline{\tau_j} \rangle_{\omega_a} = \sum_{K \subset \omega_a} \frac{\operatorname{vol}(K)}{3} \sum_{k=1}^{3} \underline{\tau_i}(m_k) \cdot \underline{\tau_j}(m_k).$$

The formula of the functions $\underline{\tau_i}$ is given by (4.7). Recall that this formula requires a fixed global orientation of the edge normal vectors. The other terms in (4.9) have similar decompositions.

In the next step we collect the computed values from (4.9) into the matrix analogue of the system (4.8). Solving this matrix equation then provides us with the basis coefficients of $\underline{\zeta_a}$. Using a local-to-global mapping of edge coefficients allows for an iterative construction of $\underline{\zeta} = \sum_{a \in \mathcal{V}} \underline{\zeta_a}$.

The construction of $\underline{\zeta_a}$ for a boundary vertex $a \in \mathcal{V}^{bdr}$ is similar; the interactions are just calculated for a slightly different set of functions. Since f is piecewise constant we do not have data oscillation; the upper bounds in Theorem 4.2 are therefore completely given in terms of $\underline{\zeta}$ and $\underline{\zeta_a}$. The last step thus consists of calculating the actual estimators: $\|\underline{\zeta} + \nabla U\|_K$ or $\|\underline{\zeta_a} + \psi_a \nabla U\|_{\omega_a}$. Since these terms are linear polynomials, we again use edge-midpoint quadrature to calculate them exactly.

4.5. Higher order Raviart-Thomas basis

In the previous sections we discussed an implementation of the lowest order Raviart-Thomas space $\mathcal{RT}_0(\mathcal{T})$, using explicit formulas for the nodal basis functions. Unfortunately, an implementation of $\mathcal{RT}_p(\mathcal{T})$ for arbitrary order p is much more involved, due

to the absence of simple $H(\mathrm{div};\Omega)$-conforming basis functions. There is little literature available on implementing the Raviart-Thomas space beyond the lowest order; many textbooks, e.g. [7, 8, 20], about the mixed finite elements seem to ignore this topic as well.

In this section we will summarize a few techniques that can be used for implementing the Raviart-Thomas space for higher order p. The numerical results in the next Chapter are all gathered using the lowest order space. A reader can therefore freely skip this section.

4.5.1. Nodal basis

The nodal basis (4.4) of $\mathcal{RT}_p(K)$ has two pleasant features: it naturally extends to a $H(\mathrm{div};\Omega)$-conforming basis of $\mathcal{RT}_p(\mathcal{T})$, and Neumann boundary conditions are easily incorporated. Instead of deriving explicit formulas for the nodal basis functions, one can (numerically) express the nodal basis in terms of another known basis of $\mathcal{RT}_p(K)$. This relatively simple approach is implemented in `FIAT` [21], a module that calculates finite element basis functions, which is used in the well-known `FEniCS` project [4].

We will describe this approach from [21] in a bit more detail. Let $n := \dim \mathcal{RT}_p(K)$ and denote $\{N_i, \phi_i\}$ for the nodal variables resp. nodal basis of the Raviart-Thomas element $\mathcal{RT}_p(K)$. Suppose that we have some basis τ_i of the space $\mathcal{RT}_p(K)$, then there exist coefficients $a_{i,j}$ such that $\phi_j = \sum_{k=1}^{n} \tau_k a_{k,j}$ holds. From the nodal requirement — $n_j(\phi_j) = \delta_{i,j}$ — it is clear that the coefficients $a_{i,j}$ can be solved from

$$VA = I \quad \text{where } V_{i,j} = n_i(\tau_j) \text{ and } A_{i,j} = a_{i,j}.$$

The matrix V is possibly ill-conditioned. `FIAT` therefore uses an orthonormal Dubiner basis [18] for the polynomial spaces, because this adds some stability. Write D_k for the set of Dubiner polynomials of degree k or less, then the basis for $\mathcal{RT}_p(K)$ is chosen as[2]

$$\{\tau_1, \ldots, \tau_n\} = \{(p,0)\}_{p\in D_k} \cup \{(0,p)\}_{p\in D_k} \cup \{(xp, yp)\}_{p\in D_k \setminus D_{k-1}}.$$

The entries $n_i(\tau_j)$ are calculated using higher order quadrature. After solving $VA = I$, one is able to express the nodal basis in terms of τ_j.

The above construction is sensitive to numerical instabilities. The advantage of this method, however, is that we directly find an $H(\mathrm{div};\Omega)$-conforming basis of $\mathcal{RT}_p(\mathcal{T})$. Furthermore, it is trivial to incorporate homogeneous Neumann conditions, and thus to provide a basis for the subspace $\mathcal{RT}_{p,0}(\omega_a)$ that is used in calculation of the equilibrated flux estimator.

4.5.2. Hierarchical basis

Instead of seeking the nodal basis, one can also directly define another $H(\mathrm{div};\Omega)$-conforming basis. Quite some research has been devoted to the construction of a *hierarchical basis*. The latter basis type is useful in FEM methods where the polynomial

[2] Actually, a few post-processing steps are performed in `FIAT` to turn this into an orthonormal basis for $\mathcal{RT}_p(K)$.

degree p varies between elements. Such a varying degree makes it harder to ensure conformity of the finite element space; the nodal basis often does not work here.

Construction of hierarchical bases is an entirely different subject in itself. Most of the (early) relevant research has been devoted to finding a $H(\mathrm{curl})$-conforming hierarchical basis for three dimensional elements. Because of similarities, this can be extended to two-dimensional $H(\mathrm{div};\Omega)$-conforming elements. The general approach is to associate polynomial spaces with edges and with the interior of elements, for which one then constructs a basis using some recipe. Influential research has been done by Webb [40], Ainsworth and Coyle [3] and Schöberl and Zaglmayr [31]. The methods used to find these basis functions make extensive use of Rham Complexes and other exterior calculus devices.

Here we will not provide the theoretical details, because of its complexity. For illustration purposes, however, we simply state the hierarchical basis from [31] for the two-dimensional Raviart-Thomas element $\mathcal{RT}_p(K)$. In this basis, one chooses the polynomial order p_K for the interior shape functions, and similarly picks a polynomial order p_e for the normal component of the shape functions on each edge $e \subset \partial K$. The method from [31] then provides basis functions matching these orders p_e and p_K. For the Raviart-Thomas element we simply have $p_e = p_K = p$.

Schöberl and Zaglmary use (integrated) scaled *Legendre polynomials* to span the polynomial spaces. Denote $(\ell_i)_{0\leqslant i\leqslant p}$ for the Legendre polynomials up to order p, then these scaled types are given by

$$\ell_k^S(s,t) := t^k \ell_k(s/t) \quad \text{and} \quad L_k^S(s,t) := t^k \int_{-1}^{s/t} \ell_{k-1}(x)\ \mathrm{d}x.$$

They also use the two-dimensional curl-operator, which is simply a 90 degrees rotation of the gradient operator.

The basis functions from [31] are as follows: for every edge $e \subset \partial K$, denote its vertices by $\{a, b\}$, then the edge-based basis functions are

$$\mathrm{curl}(\psi_a)\psi_b - \psi_a \,\mathrm{curl}(\psi_b) \quad \text{and} \quad \mathrm{curl}(L_{i+2}^S(\psi_a - \psi_b, \psi_a + \psi_b)) \quad 0 \leqslant i \leqslant p_e - 1,$$

with ψ_a the hat function at vertex a. Let $\{a, b, c\}$ be the vertices of K, then the element-based basis functions are given by:

$$\begin{aligned} \mathrm{curl}(u_i)v_j + u_i\,\mathrm{curl}(v_j) \quad & \text{for } 0 \leqslant i + j \leqslant p_K - 2, \\ \mathrm{curl}(u_i)v_j - u_i\,\mathrm{curl}(v_j) \quad & \text{for } 0 \leqslant i + j \leqslant p_K - 2, \\ \left(\mathrm{curl}(\psi_a)\psi_b - \psi_a\,\mathrm{curl}(\psi_b)\right) v_j \quad & \text{for } 0 \leqslant j \leqslant p_K - 2, \end{aligned}$$

where the polynomials u_i and v_j are defined as

$$u_i := L_{i+2}^S(\psi_b - \psi_a, \psi_a + \psi_b), \quad v_j := \psi_c \ell_j^S(2\psi_c - 1).$$

The crucial insight for implementation is that the Legendre polynomials and their derivatives can be computed using short recurrence relations. These basis functions provide a stable and efficient basis for the Raviart-Thomas element $\mathcal{RT}_p(K)$.

The element-based functions have a vanishing normal components and the set of edge-based basis functions match for two neighbouring triangles. The functions are therefore $H(\mathrm{div};\Omega)$-conforming by construction and can be used to span the entire Raviart-Thomas space $\mathcal{RT}_p(\mathcal{T})$. A basis for $\mathcal{RT}_{p,0}(\mathcal{T})$ is found by removing all edge-based functions associated to the Neumann boundary conditions. The above basis is implemented and used in `NGsolve` [30], a finite element library developed by Schöberl.

4.5.3. Bernstein-Bézier Basis

The hierarchal basis in the above is flexible with respect to the polynomial degrees, and has other beneficial properties [31]. Unfortunately, implementation of this basis for the Raviart-Thomas space is still relatively complex. Another basis construction is advocated by Ainsworth et al. [1]. They propose constructing a basis for $\mathcal{RT}_p(K)$ using the Bernstein-Bézier polynomials, for which there exist efficient computational procedures. This latter property should — according to the authors of [1] — allow for a simple, yet stable, implementation of the Raviart-Thomas space.

The basis derivation by Answorth et al. [1] is pretty elegant compared to the hierarchical basis construction of Schöberl and Zaglmayr [31]. This follows from the fact that the Bernstein-Bézier polynomials satisfy a lot of useful properties. We refer an interested reader to [1, 2] for the theoretical details.

The Raviart-Thomas basis as presented in [1] consists of three function types: the lowest order edge-based Raviart-Thomas functions from (4.7), the curl of higher order edge-based Bernstein-Bézier polynomials and element-based Bernstein-Bézier polynomials. For an implementation of the equilibrated flux estimator, one must calculate the four type of basis interactions from (4.9). Integrals containing the general function f are to be calculated using quadrature. In [2], an algorithm is presented for efficient evaluation of such quadratures containing Bézier-Bernstein basis functions. Algorithms for (exact) evaluation of the other type of basis interactions are also given in [2].

The Bernstein-Bézier basis functions from [1] provide a $H(\mathrm{div};\Omega)$-conforming basis of the Raviart-Thomas space $\mathcal{RT}_p(\mathcal{T})$. Unfortunately, an implementation of the equilibrated flux estimator that uses this basis is still quite a lot of work: for each of the three Bernstein-Bézier basis types one has to compute the interactions in (4.9).

5. Numerical results

The content of the previous chapter allow us to easily implement the equilibrated flux estimator, and present some (interesting) results. We will analyze the unknown constants present in the error bounds using this implementation. In particular, we seek out to answer questions like:

- How tight is the error bound given by the equilibrated flux estimator?
- What is the practical convergence rate of AFEM driven by the equilibrated flux estimator?
- How do the equilibrated flux estimator and classical estimator — both optimal AFEM drivers — compare to each other?

`MATLAB` [23] is used for the actual implementation. We make extensive use of `iFEM` [14], a `MATLAB` framework for (adaptive) finite element methods. All of the results are gathered for the *linear* Lagrange finite element space $\mathbb{V}(\mathcal{T})$. We let `iFEM` calculate the linear discrete solution U for a given partition $\mathcal{T}$. The package ships with an implementation of the *newest vertex bisection* refinement algorithm. This refinement method bisects marked triangles and ensures conformity as well as shape regularity of the resulting triangulation. We will first assess some classical (non-adaptive) finite element solutions. After this, we will investigate the performance of the adaptive finite element method driven by the (lowest order) equilibrated flux estimator. The results will be compared to the classical residual estimator. The implementation solves $\underline{\zeta}$ from (4.2) and uses the element-wise estimator from Theorem 4.2.

5.1. Exact error

For meaningful results, we ideally want to compare the error bounds from Theorem 4.2 with the *exact* error $|||u - U|||_\Omega$. Given the exact u, we can directly use the `iFEM` package to calculate $|||u - U|||_\Omega$. This is, unfortunately, not possible in most situations. If one knows the exact solution u of an interesting problem, then the right hand side f of the associated Poisson problem is almost never constant, requiring approximation of the calculations with f. On the other hand, the exact solution u of an interesting problem with a constant right hand side f is almost never known. We have two ways to overcome this problem.

The first is to treat the f as if it was piecewise constant. That is, we approximate $\langle \psi_a f - \nabla\psi_a \nabla U, q_k \rangle_K$ by the edge-midpoint quadrature and ignore this approximation

error. Similarly we simply replace the data oscillation term $\|f - \operatorname{div} \underline{\zeta}\|_K$ by its edge-midpoint quadrature. Provided that f is sufficiently smooth, the effects of these extra errors are of higher order h, and thus should be insignificant for small diameter.

The other approach is to use a constant f, without knowing the exact solution u. We propose to approximate the true error $|||u - U|||_\Omega$ by replacing u with another finite element solution $U_\star$ — quite dazzling if one thinks about it. To make this work we let $U_\star$ be a solution on a finer triangulation $\mathcal{T}_\star \geqslant \mathcal{T}$, such that every element in $\mathcal{T}_\star$ is at least *one* bisection away from its corresponding element in $\mathcal{T}$. We have

$$|||u - U|||_\Omega \leqslant |||u - U_\star|||_\Omega + |||U - U_\star|||_\Omega,$$

where we hope — which is most likely the case if u possesses some smoothness — that $\|u - U_\star\|_\Omega \ll \|U - U_\star\|_\Omega$. A straightforward implementation of this approximation follows from noting that U lies in the finite element spaces $\mathbb{V}(\mathcal{T}_\star)$ for refined triangulations $\mathcal{T}_\star \geqslant \mathcal{T}$. `iFEM` conveniently provides a function that calculates the coefficients of U with respect to the linear basis of $\mathcal{T}_\star$.

5.2. Error estimators

In the numerical results, we will compare the real error $|||u - U|||_\Omega$ with various estimators. A reliable and efficient estimator η — without data oscillation terms — satisfies

$$C_{\text{eff}}\eta \leqslant |||u - U|||_\Omega \leqslant C_{\text{rel}}\eta.$$

We[1] will calculate the efficiency index C_{eff} of an estimator η as $\frac{|||u-U_k|||_\Omega}{\eta}$. The equilibrated flux estimator satisfies $C_{\text{rel}} = 1$. Therefore, one would like its efficiency index to be close to one, as this shows tightness of the estimator. In the residual case, we do not know the value of C_{rel}, and thus tightness of the estimator cannot be deduced from the efficiency index alone.

The `iFEM` package provides an implementation of the standard residual estimator $\eta_{\text{res}}(U, K)$. We compare this estimator against the lowest order equilibrated flux estimator $\eta_{\text{eq}}(U, K)$ from Theorem 4.2, i.e. for $\underline{\zeta} \in \mathcal{RT}_0(\mathcal{T})$ the solution of (4.2) so that

$$|||u - U|||_\Omega \leqslant \sqrt{\sum_{K\in\mathcal{T}} \eta_{\text{eq}}^2(U, K)} \quad \text{with} \quad \eta_{\text{eq}}^2(U, K) := \left[\|\underline{\zeta} + \nabla U\|_K + \frac{h_K}{\pi}\|f - \operatorname{div} \underline{\zeta}\|_K\right]^2.$$

5.3. Uniform refinements

We start with classical finite element solutions: we let `iFEM` calculate the linear discrete solutions $(U_k)_{k\geqslant 0}$ for a sequence of uniformly bisected triangulations $(\mathcal{T}_k)_{k\geqslant 0}$, i.e. every triangle is bisected into four subtriangles.

[1] In literature, the efficiency index is often plotted as $\frac{\eta}{|||u-U_k|||_\Omega}$. We chose the inverse definition, because this is in line with our placement of the efficiency constant in Theorem 3.1.

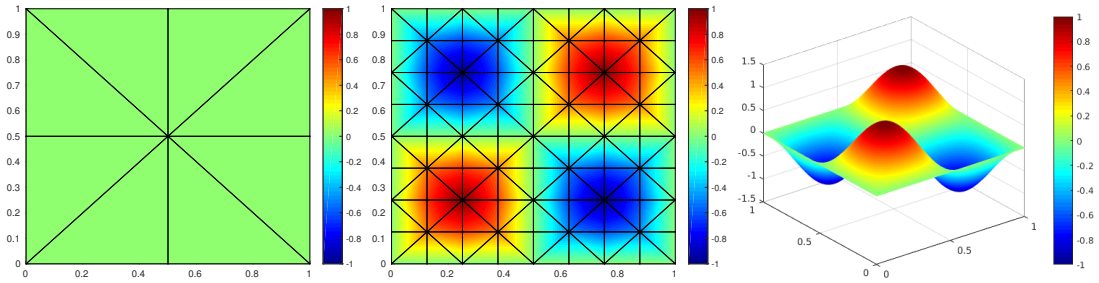


Figure 5.1.: The finite element solutions U_k for Example 5.1 with uniformly bisected triangulations, from left to right we have diameters $h = 2^{-1}, 2^{-3}, 2^{-8}$.

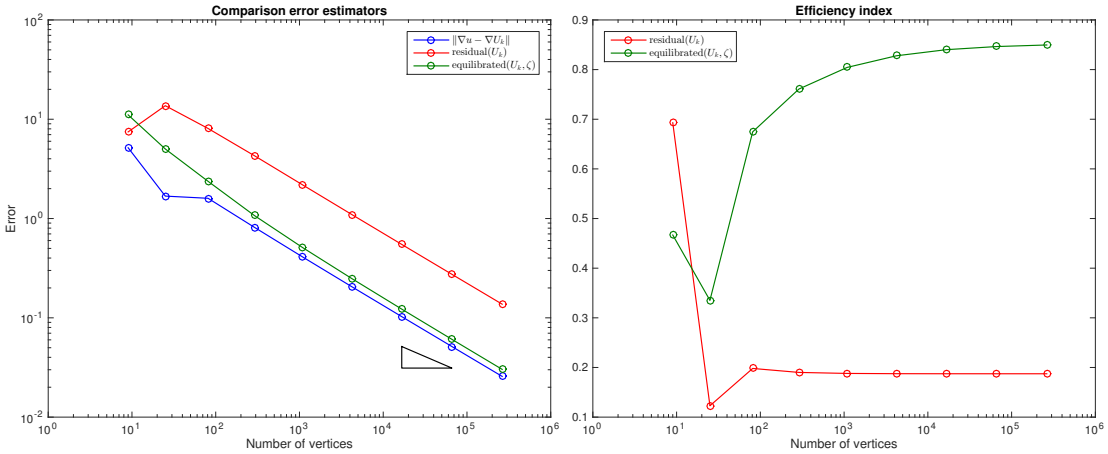


Figure 5.2.: Results for discrete solutions U_k of the sinus unit square (Example 5.1) using a sequence of uniformly refined triangulations. The left compares the exact error in the energy norm with the standard residual estimator and the equilibrated flux estimator. The triangle has a slope of $-1/2$. The right figure plots $|||u - U_k||| / \eta$; an estimation of the efficiency index.

Example 5.1 (Unit square)**.** Take the unit square domain $\Omega = (0,1) \times (0,1)$, with exact solution $u(x,y) = \sin(2\pi x)\sin(2\pi y)$. The corresponding Poisson problem is given by

$$\begin{aligned} -\Delta u &= 8\pi^2 \sin(2\pi x)\sin(2\pi y) && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega. \end{aligned}$$

The exact solution u has four peaks and is very smooth. `iFEM` uses third order quadrature to approximate the integrals containing f. The produced discrete solutions U_k are visualized in Figure 5.1. The images clearly illustrate some convergence of the FEM solution. As $u \in H^2(\Omega)$, we expect that $|||u - U_k|||_\Omega \lesssim h|u|_{H^2(\Omega)}$ from (1.5) . For uniform refinements, this latter is equivalent to $|||u - U_k|||_\Omega \lesssim N^{-1/2}|u|_{H^2(\Omega)}$, with N the total number of vertices.

Results for the actual error (estimators) are given in Figure 5.2. The experimental convergence rate is in line with the theoretical convergence rate, as can see using the

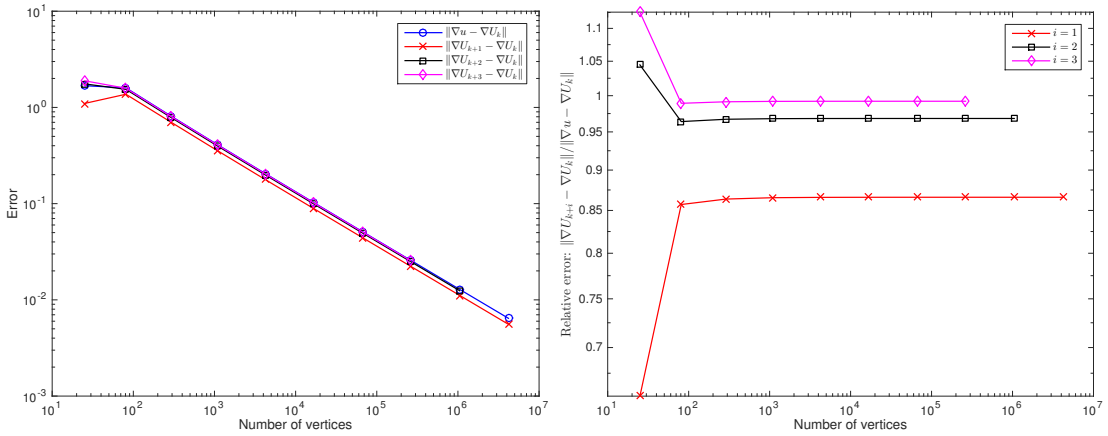


Figure 5.3.: A comparison of the exact error $|||u - U_k|||_\Omega$ with the approximation $|||U_{k+i} - U_k|||_\Omega$ $(i = 1, 2, 3)$ for Example 5.1. The left image plots all the norms, which results in a cluttered image as expected. The right image displays the relative terms, i.e. $|||U_{k+i} - U_k|||_\Omega \,/\, |||u - U_k|||_\Omega$.

slope triangle. The efficiency indices of η_{eq} and η_{res} differ greatly. The equilibrated flux estimator is about four times closer to the real error than the residual estimator. Moreover, the efficiency index of the equilibrated flux estimator seems to increase, which could be explained by a reduction of quadrature errors for smaller diameters.

We know the exact solution u for this system, and thus we can experimentally verify the claim that $|||u - U_k|||_\Omega$ is approximately equal to $|||U_{k+i} - U_k|||_\Omega$ for $i \geqslant 1$. A comparison for $i = 1, 2, 3$ is given in Figure 5.3. The quality of the error approximation increases with i as expected.

Example 5.2. Another Poisson problem for the unit square domain is given by

$$u(x, y) = 2^{40} x^{10} (1 - x)^{10} y^{10} (1 - y)^{10}.$$

This problem has homogeneous Dirichlet boundary conditions.

The corresponding right hand side f is a high order polynomial; we let `MATLAB` derivate it using symbolic calculations. The solution u has a peak of value 1 centered around $(\frac{1}{2}, \frac{1}{2})$ and is again very smooth. In Figure 5.4 the efficiency indices are given, alongside the relative errors generated by $|||U_{k+i} - U_k|||_\Omega$ for this new problem. The efficiency of the equilibrated flux estimator is smaller for this problem, which is most likely due to the large oscillation factor, but it still outperforms the classical residual estimator.

The behaviour of $|||U_{k+i} - U_k|||_\Omega$ is very similar to that of the previous problem; it appears to be a good estimation of the exact error $|||u - U_k|||_\Omega$. In the following problems we will treat this approximation as the 'true error', since we no longer have a closed form of the exact solution u. For accuracy reasons we pick $i = 3$, which should be good enough for a true error indicator. We will write $U_{\star,k}$ for the discrete solution on $\mathcal{T}_{\star,k}$, the triangulation found by uniformly refining every triangle in $\mathcal{T}_k$ three times.

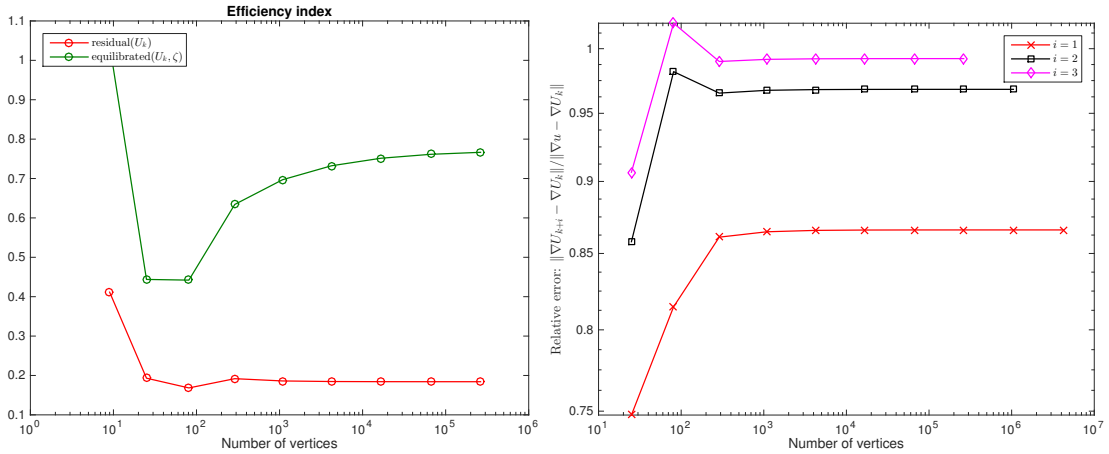


Figure 5.4.: Results for the square peak (Example 5.2). Left image compares the efficiency indices of the standard residual estimator with the equilibrated flux estimator. Right image shows the exact error approximation quality.

Example 5.3 (Re-entrant corner)**.** Here we consider the domain $\Omega = (-1,1)^2 \setminus [-1,0]^2$, which is also known as the L-shaped domain. The Poisson problem is given by

$$-\Delta u = \mathbb{1} \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega.$$

The solution u of this problem has a singularity at the origin, so that $u \notin H^2(\Omega)$.

We calculate discrete solutions U_k for uniform refinements. The exact solution u is unknown, and thus we will use $|||U_{\star,k} - U_k|||_\Omega$ as a true error indicator. The results are given in Figure 5.5. The experimental convergence of the FEM solution appears to be of order $h^{2/3}$. The equilibrated flux estimator has a better efficiency index than the classical residual estimator. There is a drop in terms of efficiency compared with the previous problems, because of the singularity u has at the origin. Interestingly, the efficiency of the equilibrated estimator seems to decrease, whereas the efficiency of the standard estimator is somewhat increasing in N. This might be related to the approximation quality of $|||U_{\star,k} - U_k|||_\Omega$.

Example 5.4 (Crack domain)**.** We solve $f = \mathbb{1}$ and $u = 0$ on $\partial\Omega$, for the crack domain

$$\Omega = \{(x,y) \in \mathbb{R}^2 : |x| + |y| < 1\} \setminus ([0,1] \times \{0\}).$$

The solution u has a line singularity.

This domain is implemented by adding the vertex $(1,0)$ twice. The results are given in Figure 5.6. The experimental convergence rate on uniformly refined triangulations is of order $h^{1/2}$. This is even lower than the rate from the previous example, because we now have an entire line singularity. There is also a drop in the efficiency index compared to the re-entrant corner; this could be explained by the fact that u is less regular. The overall behaviour of the efficiency is similar to the previous example: the efficiency of the equilibrated flux estimator seems to decrease, while the efficiency of the residual estimator increases.

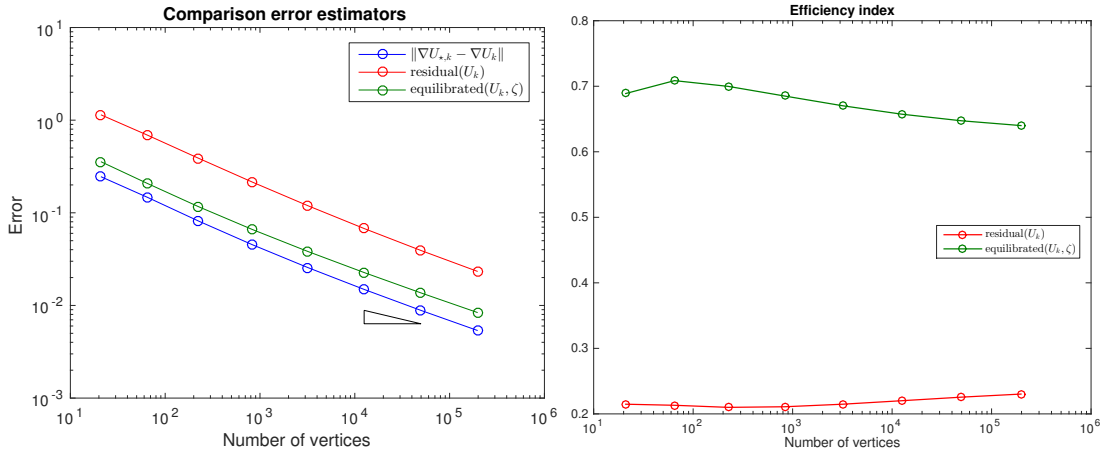


Figure 5.5.: Error estimators and their efficiency for the re-entrant corner domain (Example 5.3). The triangle has a slope of $-1/3$.

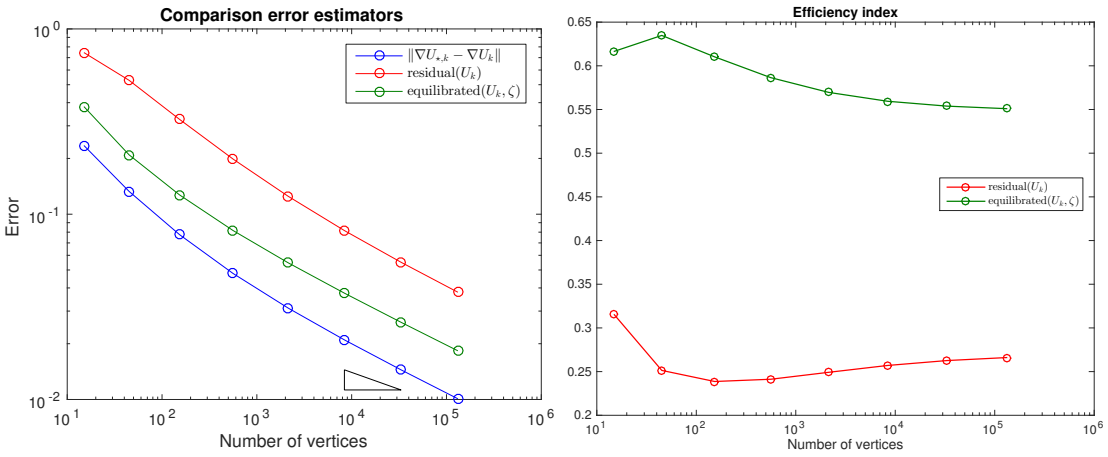


Figure 5.6.: Error estimators and their efficiency for the crack domain (Example 5.4). The triangle has a slope of $-1/4$.

5.4. Adaptive finite element method

The convergence rates obtained for these last two examples are lower than rates found in the first two (smooth) examples. This is a direct consequence of the singularities appearing in these last two examples. These rates can be improved by using the adaptive finite element method. The `iFEM` package contains an implementation of newest vertex bisection refinement and Dörfler marking, making it easy to implement the adaptive method. We will restrict ourself to a constant right hand side f, because this avoid the effects of data oscillation, and thus makes the total error indicator ϑ — necessary in the AFEM optimality proof — equal to the estimator η.

Recall that the equilibrated flux estimator η_{eq} used here differs from the patch-wise estimator, $\|\underline{\zeta_a} + \psi_a \nabla U\|^2_{\omega_a}$, used in the optimality AFEM proof. Since these are closely related, it is reasonable to expect some sort of optimality from the above version as well (see the discussion in §3.5). The advantage of the element-wise estimator η_{eq} is that we can mark elements instead of patches — allowing us to use `iFEM` directly.

The Dörfler marking parameter θ used in AFEM must be small enough to ensure optimality. This Dörfler upper bound is defined in terms of the efficiency constant and the *discrete* reliability constant, see Theorem 3.1 and Lemma 3.4. Unfortunately, we have no estimations of the discrete reliability constant. Earlier experiments, e.g. [13], show that $\theta = 1/2$ is often small enough.

The performance of AFEM driven by the equilibrated flux estimator is compared with AFEM driven by the classical estimator, and classical FEM using uniform refinements. That is, we produce a sequence $(U_k, \mathcal{T}_k)_{k\geqslant 0}$ of (A)FEM solutions for all of these three methods. For each solution U_k we measure its approximation error by $|||U_{\star,k} - U_k|||_\Omega$.

In Figure 5.7, results are given for the crack domain and the L-shaped domain using Dörfler marking parameter $\theta = 1/2$. One directly sees that the uniform convergence rate is improved by adaptivity — as predicted by the theory. The AFEM solutions produced by the residual estimator and the equilibrated flux estimator are of more or less the same quality. This confirms our conjecture that AFEM driven by the (lower order) element-wise equilibrated flux estimator is also optimal. It appears that eventually residual driven AFEM produces solutions with a slightly smaller approximation error. This might be related to the decreasing efficiency of the equilibrated flux estimator, cf. Figures 5.6 and 5.5.

5.5. Mixed finite element solution

Prager and Synge' theorem provided us with the following upper bound:

$$|||u - U|||^2_\Omega \leqslant \|\nabla U - \underline{\sigma}\|^2_\Omega \qquad \text{for } \underline{\sigma} \in H(\operatorname{div};\Omega) \text{ s.t. } \operatorname{div}\underline{\sigma} + f = 0$$

The equilibration method constructs a $\underline{\zeta} \in \mathcal{RT}_p(\Omega)$ such that $-\underline{\zeta}$ satisfies the equilibrium condition, and thus we find an upper bound in terms of $\nabla U + \underline{\zeta}$. As noted before, the best upper bound in the Raviart-Thomas space is found by minimizing $\|\nabla U - \underline{\sigma}\|^2_\Omega$ over all fluxes $\underline{\sigma} \in \mathcal{RT}_p(\Omega)$ that are in equilibrium. The mixed finite element method provides

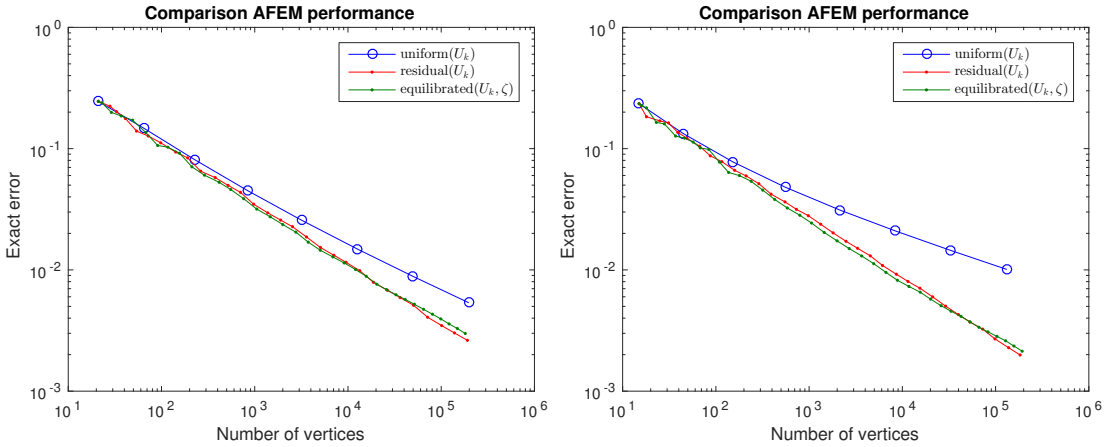


Figure 5.7.: Comparison of (A)FEM methods on the L-shaped domain (left) and the crack domain (right). The three methods are compared by plotting the approximation error $|||U_{\star,k} - U_K|||_\Omega$ against the number of vertices. The adaptive methods used Dörfler marking $\theta = 1/2$.

the flux $\underline{\sigma_m}$ that globally minimizes this norm [7]. Since this method is too expensive for an estimator calculation, the method of minimizing local problems was introduced. This makes it interesting to compare the performance of the equilibrated flux estimator with the mixed flux estimator $\|\nabla U - \underline{\sigma_m}\|_\Omega$.

The `iFEM` package ships with an implementation of the lowest order mixed finite element solution. Figure 5.8 gives the efficiency indices of the mixed- and equilibrated flux estimator for the example problems using uniform refinements. The mixed flux estimator has a higher efficiency index in all examples — as one would expect. More surprising is the behaviour on the unit square problems. The efficiency coefficients of the equilibrated flux estimator seem to converge to the efficiency index of the mixed flux estimator. This suggests that the penalty of applying local instead of global minimization decreases for finer triangulations. For the L-shaped and crack domain we have different behaviour, both the efficiency indices seem to decreases for fine triangulations.

The mixed flux estimator provides an element-wise error estimator by restricting the norm to an element, i.e. $\eta^2_{\text{mix}}(U, K) := \|\nabla U - \underline{\sigma_m}\|^2_K$. We can use this estimator to drive AFEM. Figure 5.9 compares the AFEM performance of the various methods on the crack domain with the smaller Dörfler parameter $\theta = 3/10$. As before, all of the estimators produce discrete solutions of the same approximation quality. This suggests that the earlier marking parameter of $1/2$ was already small enough for optimality.

5.6. Zienkiewicz-Zhu error estimator

We have seen the advantages of the equilibrated flux estimator. Unfortunately, this estimator and many other promising estimators are not often used in practice. This is mainly due to the implementational complexity and cost of such estimators. Instead,

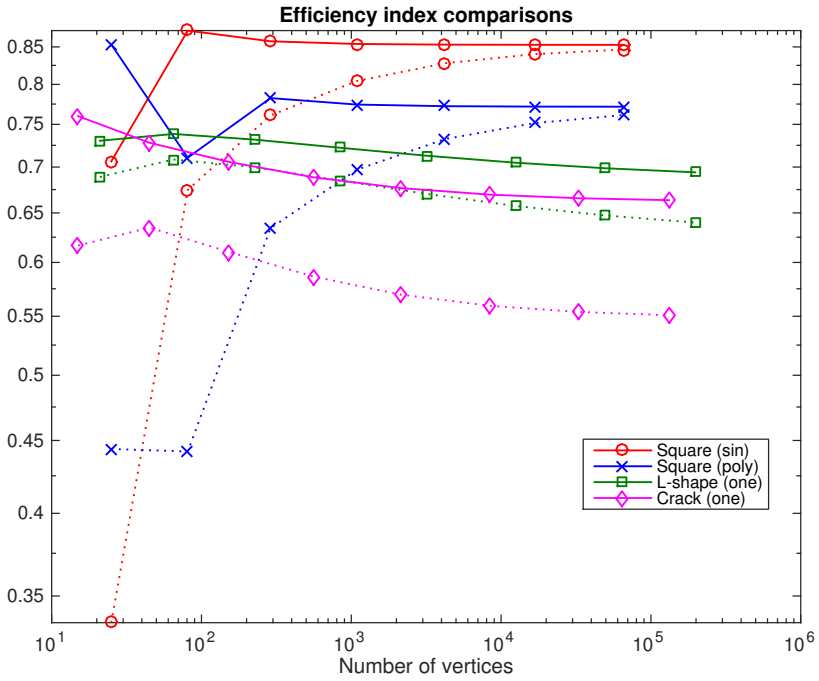


Figure 5.8.: Compares the efficiency index of the equilibrated flux estimator (dashed lines) against the mixed flux estimator (solid lines), for four different domains indicated by color and marker styles.

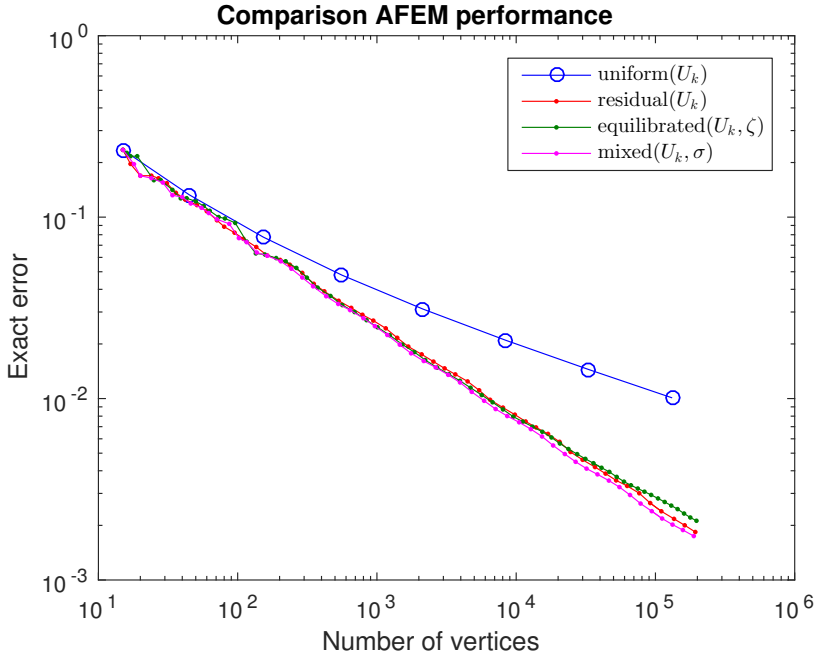


Figure 5.9.: Approximation quality of various (A)FEM solutions for the crack domain. The adaptive solutions are found using $\theta = 3/10$.

engineers tend to use far easier estimators. One well-known example is the *Zienkewicz-Zhu* estimator [42, 43]. It is praised for its simplicity and cost effectiveness.

The general idea is based on *gradient recovery*. That is, one smoothens the discrete gradient ∇U to obtain a recovered gradient $G_{\mathcal{T}}U$ in the continuous space $\left[\mathbb{V}(\mathcal{T})\right]^2$. Here one hopes that $G_{\mathcal{T}}U$ provides a better approximation of the exact gradient ∇u than the discrete gradient ∇U. If this is the case, we would get $\nabla u - \nabla U \approx G_{\mathcal{T}}U - \nabla U$ and thus the latter provides an error estimator. Various gradient recovery strategies are documented in the literature, for an overview see [43].

The initial gradient recovery estimator proposed by Zienkiewicz and Zhu in [42] is based on the orthogonal projection of ∇U into $\left[\mathbb{V}(\mathcal{T})\right]^2$. Write $G^*_{\mathcal{T}}U$ for the orthogonal projector onto $\left[\mathbb{V}(\mathcal{T})\right]^2$, then $G^*_{\mathcal{T}}U$ is given as the solution of

$$\langle G^*_{\mathcal{T}}U, \underline{v}\rangle_\Omega = \langle \nabla U, \underline{v}\rangle_\Omega \quad \forall \underline{v} \in \left[\mathbb{V}(\mathcal{T})\right]^2.$$

A function $\underline{v} \in \left[\mathbb{V}(\mathcal{T})\right]^2$ is determined by its values at the vertices $a \in \mathcal{V}$, because we consider the linear finite element space. Rewrite the above equation using the hat functions ψ_a as basis for $\mathbb{V}(\mathcal{T})$. This reveals that the values of $G^*_{\mathcal{T}}U$ at vertices $a \in \mathcal{V}$ are the solution of

$$\sum_{a\in\mathcal{V}} (G^*_{\mathcal{T}}U)(a)\langle \psi_a, \psi_b\rangle_\Omega = \langle \nabla U, \psi_b\rangle_\Omega = \sum_{K\subset\omega_b} \frac{\mathrm{vol}(K)}{3}\nabla U|_K \quad \forall b \in \mathcal{V}. \tag{5.1}$$

The second equality holds because ∇U is a piecewise constant vector.

Solving the above equation is as expensive as calculating the discrete solution U itself. To reduce this computational cost, Zienkiewcz and Zhu [42] propose to approximate the above integrals using trapezoidal quadrature rule, i.e. $\int_K g \approx \frac{\mathrm{vol}(K)}{3}\sum_{a\in\mathcal{V}_K} g(a)$. The recovered gradient $G_{\mathcal{T}}U$ is the solution of (5.1), with the integrals calculated by this quadrature. All hat functions ψ_b with $b \neq a$ vanish at vertex a, and therefore we obtain the remarkably simple expression for $G_{\mathcal{T}}U$ at a vertex $a \in \mathcal{V}$:

$$(G_{\mathcal{T}}U)(a) = \sum_{K\subset\omega_a} \frac{\mathrm{vol}(K)}{\mathrm{vol}(\omega_a)}\nabla U|_K.$$

In other words, the recovered gradient $G_{\mathcal{T}}U$ at a vertex $a \in \mathcal{V}$ is simply the weighted average of ∇U over elements in the patch ω_a. Notice that ∇U is piecewise constant, whereas the recovered gradient $G_{\mathcal{T}}U$ is a continuous piecewise linear function. The recovered gradient $G_{\mathcal{T}}U$ is therefore possibly a better approximation of ∇u — especially near the discontinuities of ∇U.

The Zienkiewicz-Zhu (ZZ) estimator is given by $\eta_{\mathrm{ZZ}}(U, K) := \|G_{\mathcal{T}}U - \nabla U\|_K$. We will not provide theoretical details about reliability and efficiency of this estimator [29, 43], but only consider the experimental performance. The patch-wise nature of this ZZ estimator provides some resemblance with the equilibrated flux estimator. It is clearly interesting to compare this cheap and simple ZZ estimator with the equilibrated flux estimator.

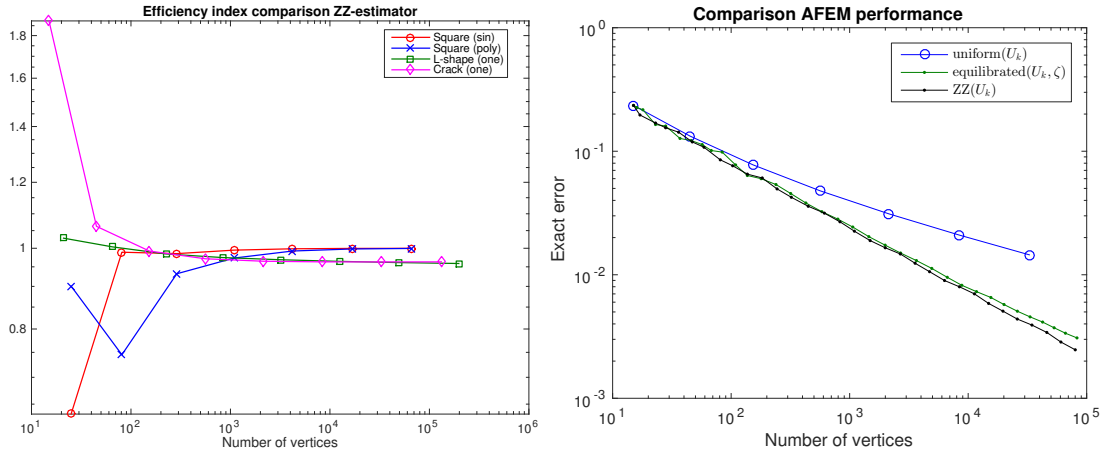


Figure 5.10.: The left image compares the efficiency index of the ZZ estimator for various domains. The right figure shows the quality of three (A)FEM produced sequences for the crack domain (Example 5.4). The adaptive solutions are found using Dörfler marking parameter $\theta = 1/2$.

An implementation of the ZZ estimator is straightforward using the framework provided by `iFEM`. Numerical results for the ZZ estimator are given in Figure 5.10. The left image displays the efficiency index of the ZZ estimator for the various examples using uniform refinements. The results are fascinating! The efficiency index of the ZZ estimator is close to 1 for all examples previously studied. Compare these results to the efficiency indices of the equilibrated- and mixed flux estimator in Figure 5.8. The ZZ estimator outperforms both these (relatively) expensive flux estimators.

This high accuracy is closely connected to the polynomial degree used in the FEM space $\mathbb{V}(\mathcal{T})$. That is, the accuracy is expected to drop if one uses higher order polynomials in the FEM space [5]. There are different gradient recovery methods which are also stable for higher order polynomials [43].

The ZZ estimator can be used to drive AFEM. The right image in Figure 5.10 compares this ZZ driven AFEM method with the AFEM results produced by the equilibrated flux estimator. The results are (again) gathered for the crack domain (Example 5.4). Unsurprisingly, we see that the ZZ driven AFEM solutions are of the same quality as the ones produced by the equilibrated flux estimator.

5.7. Discussion

The numerical results in this chapter support the promising claims about the equilibrated flux estimator. The efficiency index of the equilibrated flux estimator behaves well for all of the examples studied: the lowest efficiency index of somewhere above 0.55 was found for the crack domain. The estimator therefore provides tight error bounds, since its reliability bound is constant-free.

Precision of the equilibrated flux estimator seems to be related to the smoothness of the

exact solution. Indeed, the efficiency index of the estimator decreases when comparing examples 5.1–5.4. Since these examples are also ordered on regularity of u, this suggests that a less smooth solution u negatively effects the efficiency index. The mixed flux estimator η_{mix} also follows this pattern. The efficiency can be improved by using the first order equilibrated flux estimator, but according to Braess et. al [9] this increases the efficiency index only slightly.

Out of the estimators reviewed, the standard residual estimator has the worst efficiency index of approximately between 0.2 and 0.3. Surprisingly, the ZZ-estimator provided the highest quality with an efficiency index near one. Quality of this estimator, however, is expected to deteriorate if one considers higher order finite element spaces. The convergence rate of AFEM driven by the various estimators is similar, as expected.

All results were gathered for the linear FEM space $\mathbb{V}(\mathcal{T})$ using the zeroth order equilibrated flux estimator. Furthermore, we used the element-wise estimator instead of the patch-wise version that was used to prove optimality. The numerical AFEM results seem to provide optimal convergence rates, and therefore numerically confirm our conjecture from §3.5 stating that AFEM driven by the element-wise equilibrated flux estimator is also optimal.

Unfortunately, the current implementation is not capable of calculating higher order equilibrated flux estimators. This limitation leaves open some important research aspects, the most important aspect being p-robustness of the estimator. Further research has to be conducted to determine the behaviour of the equilibrated flux estimator under a variation of the polynomial degree used in the FEM space. This latter aspect could play a crucial role in the development of optimal hp-AFEM algorithms.

Popular summary

Exact solutions of many partial differential equations are hard or impossible to compute. This is a problem in many practical (research) fields — think of an engineer wanting to solve a heat equation for some object. One way to overcome this problem is to use an *approximation* method. That is, instead of seeking the exact solution one constructs a function that is close in norm to the exact solution. For many types of partial differential equations, such approximations are provided by the *finite element method* (FEM).

Consider for example a two-dimensional Poisson boundary value problem on the unit square $\Omega = [0,1] \times [0,1]$: solve u from

$$-\Delta u = 1 \text{ in } \Omega, \qquad u = 0 \text{ on } \partial\Omega.$$

The first step of FEM is to subdivide the domain Ω into smaller pieces, called *elements*. For our example, this means that we subdivide the unit square into triangles. In the next step, we define a function space on each of these elements, and combine the elements to form a *finite* dimensional function space over the domain Ω. An approximation U of the exact solution u is then found in this finite element space by solving a system of equations, where each element imposes a set of equations. The resulting U is the best approximation of u in this finite element space.

In order to improve the quality of U we could subdivide every element and calculate an approximation for the refined subdivision. This classical approach works, i.e. the sequence of FEM approximations converges towards the exact solution. For some problems, a faster convergence rate can be achieved by only subdividing those elements where the current approximation U has the worst quality. This is the so-called *adaptive* finite element method (AFEM).

The immediate question is how to determine the approximation error $\|u - U\|$ on an element. After all, we apply approximation techniques because the exact solution u is unknown. Fascinatingly, there are *estimators* that can be used to estimate the error $\|u - U\|$ on an element — without knowing the exact solution u! One well-known estimator is the residual error estimator. This was the first estimator for which *optimal* convergence of adaptive finite element method was proved.

The bound provided by the residual error estimator contains an unknown constant, making it hard to determine when the approximation error $\|u - U\|$ is small enough. Recently, quite some research interest has been conducted in finding estimators with more favourable properties. In this work, we concentrate on the *equilibrated flux* estimator. This estimator provides a guaranteed upper bound, i.e. without an unknown constant. We give an in-depth overview of its properties and use these properties to show that AFEM driven by this estimator also provides optimal convergence. Lastly, we experimentally examine the performance of the estimator using an actual implementation.

Appendices

A. Notation

Ω	A bounded polyhedral domain in $\mathbb{R}^n$
u	Almost always used to denote the *exact* solution of the boundary-value problem at hand
$a \lesssim b$	There holds $a \leqslant C_{st} b$ for some constant C_{st}
$a \gtrsim b$	There holds $b \lesssim a$
$a \eqsim b$	If $a \lesssim b$ and $a \gtrsim b$
$\underline{v}$	Used to indicate that v is a vector-valued function
$C^m(\Omega)$	The space of m-times continuously differentiable functions
$L^2(\Omega)$	Space of square-integrable functions over Ω
$H^m(\Omega)$	Sobolev space of $L^2(\Omega)$ functions with weak derivatives up to order m
$H_0^m(\Omega)$	Subspace of $H^m(\Omega)$ of all functions with vanishing trace on $\partial\Omega$
$H(\mathrm{div};\Omega)$	Space of vector-valued $\left[L^2(\Omega)\right]^2$ functions with a weak divergence in $L^2(\Omega)$
$\langle f, g\rangle_\Omega$	The regular (vector-valued) $L^2(\Omega)$ inner product
$\|\cdot\|_\Omega$	The regular (vector-valued) $L^2(\Omega)$ norm
$\|\cdot\|_{H^m(\Omega)}$	The Sobolev norm of order m on Ω
$\|\cdot\|_{H^m(\Omega)}$	The Sobolev seminorm of order m on Ω
$\|\!\|\!\|\cdot\|\!\|\!\|_\Omega$	The energy norm, i.e. $\|\!\|\!\|v\|\!\|\!\|_\Omega^2 = a(v,v)$
$\mathcal{T}$	A conforming triangulation of Ω
$\mathcal{V}, \mathcal{E}$	The vertices and edges of a given triangulation $\mathcal{T} \subset \mathbb{R}^2$
ω_a	The patch centered at vertex $a \in \mathcal{V}$: all the triangles that touch a
$\mathcal{P}_p(K)$	The space of polynomials of degree p for an element K
$\mathcal{P}_p^{-1}(\mathcal{T})$	$:= \prod_{K\in\mathcal{T}} \mathcal{P}_p(K)$

$\mathcal{RT}_p(K)$	$:= \left[\mathcal{P}_p(K)\right]^2 + \mathcal{P}_p(K)\underline{x}$
$\mathcal{RT}_p^{-1}(\mathcal{T})$	The broken Raviart-Thomas space, i.e. $\prod_{K\in\mathcal{T}} \mathcal{RT}_p(K)$
$\mathcal{RT}_p(\mathcal{T})$	$:= H(\mathrm{div};\Omega) \cap \mathcal{RT}_p^{-1}(\mathcal{T})$
$[\![\underline{\sigma}]\!]$	$[\![\underline{\sigma}]\!](x) := \lim_{\epsilon\to 0} \underline{\sigma}(x+\epsilon n)\cdot n - \underline{\sigma}(x-\epsilon n)\cdot n$, the jump over an interface in the direction of the unit normal n
v_Ω	The mean of the function v over some domain Ω: $v_\Omega := \int_\Omega v$

B. Definitions and Reference theorems

B.1. Sobolev spaces

A proof of the following theorems can be found in most standard works, e.g. [20].

Theorem B.1. *Let Ω be a bounded domain with Lipschitz boundary, then for all $v, w \in H^1(\Omega)$ there holds*

$$\int_\Omega v \frac{\partial w}{\partial x_i} = -\int_\Omega w \frac{\partial v}{\partial x_i} + \int_{\partial\Omega} vwn_i \quad i \in 1, \ldots, n,$$

so that for $u \in H^2(\Omega)$ we have

$$\int_\Omega v\Delta u = -\int_\Omega \nabla u \cdot \nabla v + \int_{\partial\Omega} v\nabla u \cdot \underline{n}.$$

Here the functions on the boundary are to be interpreted using the trace operator.

Definition B.1. The weak divergent Sobolev space is defined by

$$H(\mathrm{div};\Omega) := \left\{ \underline{\sigma} \in [L^2(\Omega)]^2 : \mathrm{div}\, \underline{\sigma} \in L^2(\Omega) \right\}.$$

Where $\mathrm{div}\, \underline{\sigma} \in L^2(\Omega)$ is interpreted in distributional sense, i.e. there exists a $v \in L^2(\Omega)$ such that

$$-\int_\Omega \nabla\varphi \cdot \underline{\sigma} = \int_\Omega v\varphi \quad \forall \varphi \in D(\Omega).$$

This turns $H(\mathrm{div};\Omega)$ into a Hilbert space with inner product

$$\langle \underline{\sigma}, \underline{\tau} \rangle_{\mathrm{div},\Omega} := \langle \underline{\sigma}, \underline{\tau} \rangle_\Omega + \langle \mathrm{div}\, \underline{\sigma}, \mathrm{div}\, \underline{\tau} \rangle_\Omega \quad \text{for } \underline{\sigma}, \underline{\tau} \in H(\mathrm{div};\Omega).$$

Theorem B.2. *Let Ω be a bounded domain with Lipschitz boundary, then for all $\sigma \in H(\mathrm{div};\Omega)$ and $v \in H^1(\Omega)$ we have*

$$\int_\Omega \underline{\sigma} \cdot \nabla v + \int_\Omega v\, \mathrm{div}\, \underline{\sigma} = \int_{\partial\Omega} v\underline{\sigma} \cdot \underline{n}$$

Actually, the right hand side is an abuse of notation. Formally one has that $\underline{\sigma} \cdot \underline{n}$ is an operator working on the trace of v on $\partial\Omega$.

B.2. Poincaré-Friedrichs inequality

For some H^1 functions it is possible to (universally) bound the L^2-norm by its H^1-seminorm. Both Poincaré and Friedrichs inequality provide such bounds. Because of their similarities, these type of inequalities are named *Poincaré-Friedrichs inequality*.

Theorem B.3 (Friedrichs inequality)**.** *Suppose that $\Omega \subset \mathbb{R}^n$ is a bounded Lipschitz domain, then for h_Ω the diameter of Ω we have*

$$\|v\|_\Omega \leqslant h_\Omega \|\nabla v\|_\Omega \quad \forall v \in H^1_0(\Omega).$$

For $\Gamma \subset \partial\Omega$ with positive $(n-1)$-dimensional measure there exists a constant $C_F(\Omega, \Gamma)$ such that

$$\|v\|_\Omega \leqslant C_F h_\Omega \|\nabla v\|_\Omega \quad \forall v \in H^1(\Omega), v|_\Gamma = 0.$$

Theorem B.4 (Poincaré inequality)**.** *Again suppose that $\Omega \subset \mathbb{R}^n$ is a bounded Lipschitz domain, then there exists a constant $C_P(\Omega)$ such that*

$$\|v - v_\Omega\|_\Omega \leqslant C_P h_\Omega \|\nabla v\|_\Omega \quad \forall v \in H^1(\Omega).$$

Here C_F and C_P — combined C_{FP} in short — are actually taken as the smallest constant such that the bound holds. It is not directly clear how C_{FP} depends on the shape of the domain. For some domain types, explicit upper bounds can be given. All *convex* domains Ω satisfy $C_P(\Omega) \leqslant \frac{1}{\pi}$. Constant bounds for Friedrichs inequality in terms of geometry can be found in literature, e.g. [41, 37]. In particular, for stars ω_a — the elements touching a vertex a in some triangulation — a constant bound can be given that only depends on the shape regularity of the associated triangulation:

$$C_F(\omega_a) \leqslant C(\sup_{K \in \mathcal{T}} h_K/p_K).$$

B.3. Raviart-Thomas elements

Raviart-Thomas functions can be (partially) determined by their divergence and surface normal, as formalised by the folowing lemma.

Lemma B.5. *Consider an element K with Raviart-Thomas space $\mathcal{RT}_p(K)$. There exists a function $\underline{\sigma} \in \mathcal{RT}_p(K)$ that solves*

$$\begin{aligned} \operatorname{div} \underline{\sigma} &= p_K \\ \underline{\sigma} \cdot n &= p_e, \end{aligned} \tag{B.1}$$

for polynomials $p_K \in \mathcal{P}_p(K)$ and $p_e \in \mathcal{P}_p(\partial K)$ satisfying the compatibility constraint

$$\int_K p_K = \int_{\partial K} p_e.$$

Proof. Counting degrees of freedom shows that the system (B.1) is overdetermined by maximally one degree of freedom. Imposing the compatibility constraint takes away this extra freedom. Existence of a solution then follows from showing that the system is consistent.

Another — perhaps more elegant — derivation is as follows. Consider the Neumann problem

$$\Delta u = p_K \quad \text{in } K, \quad \nabla u \cdot n = p_e \quad \text{in } \partial K,$$

or in weak formulation,

$$\int_K \nabla u \cdot \nabla v = \int_{\partial K} v p_e - \int_K v p_K \quad \forall v \in H^1(K).$$

This has a unique solution u under the assumption that the right hand side has mean zero, which is exactly the compatibility constraint. Notice that ∇u satisfies the wanted properties (B.1):

$$\operatorname{div} \nabla u = \Delta = p_K, \quad \nabla u \cdot n = p_e.$$

We then simply set $\underline{\sigma} = \Pi^{RT} u$, for Π^{RT} the interpolation projector on the space $\mathcal{RT}_p(K)$, formally defined in [7, §3.4]. From the properties of this projector [7, Lem 3.7] we deduce that $\underline{\sigma}$ is a function in $\mathcal{RT}_p(K)$ that satisfies (B.1). □

B.4. Auxiliary results

Here the proof of Lemma 2.8 is included here for self-containedness.

Lemma. *The estimator $\|\underline{\sigma_a}\|_{\omega_a}$ is bounded from below by the classical estimator,*

$$\|\underline{\sigma_a}\|_{\omega_a} \gtrsim h_{\omega_a} \|f + \Delta U\|_{\omega_a} + h_{\omega_a}^{1/2} \|[\![\nabla U]\!]\|_{\gamma_a},$$

for a constant depending on the shape regularity and the polynomial-degree p used in $\mathbb{V}$.

Proof. A lower bound for $\|\underline{\sigma_a}\|_{\omega_a}$ can be given in terms or r_a. By the definition of the local equilibrium property (2.2) we have $\langle r_a, v\rangle = -\langle \underline{\sigma_a}, \nabla v\rangle_{\omega_a}$ for $v \in H^1_\star(\omega_a)$, and thus

$$\|r_a\|_{H^1_\star(\omega_a)'} = \sup_{\{v \in H^1_\star(\omega_a) : \|\nabla v\|_{\omega_a} = 1\}} \langle r_a, v\rangle = \sup_{\{v \in H^1_\star(\omega_a) : \|\nabla v\|_{\omega_a} = 1\}} \langle \underline{\sigma_a}, \nabla v\rangle_{\omega_a} \leqslant \|\underline{\sigma_a}\|_{\omega_a}.$$

Fix an interior vertex $a \in \mathcal{V}^{int}$. Applying the above and using orthogonality on constant functions reveals

$$\begin{aligned}
\|\underline{\sigma_a}\|_{\omega_a} &\geqslant \|r_a\|_{H^1_\star(\omega_a)'} \\
&= \sup_{\{v \in H^1_\star(\omega_a) : v \neq 0\}} \frac{\langle r_a, v\rangle}{\|\nabla v\|_{\omega_a}} \\
&= \sup_{\{v \in H^1(\omega_a) : v \neq C_{st}\}} \frac{\langle r_a, v - v_{\omega_a}\rangle}{\|\nabla (v - v_{\omega_a})\|_{\omega_a}} \\
&= \sup_{\{v \in H^1(\omega_a) : v \neq C_{st}\}} \frac{\langle r_a, v\rangle}{\|\nabla v\|_{\omega_a}}.
\end{aligned}$$

Similar to the proof of Lemma 2.7 we decompose r_a in triangle and edge terms

$$\langle r_a, v\rangle = \sum_{K\subset\omega_a} \langle r_a^T, v\rangle_K + \sum_{e\subset\gamma_a} \langle r_a^e, v\rangle_e, = \sum_{K\subset\omega_a} \langle \psi_a \left[f + \Delta U\right], v\rangle_K + \sum_{e\subset\gamma_a} \langle \psi_a [\![\nabla U]\!], v\rangle_e.$$

The triangle and edge functions (r_a^T, r_a^e) are actually broken polynomials from the spaces $\mathcal{P}_r^{-1}(\omega_a) \times \mathcal{P}_r^{-1}(\gamma_a)$ for some degree r.

Start with an element $K \subset \omega_a$, and consider the space $H_0^1(K)$. Since a function $v \in H_0^1(K)$ vanishes on the boundary, we can naturally extend it to a function $v \in H^1(\omega_a)$ by setting $v \equiv 0$ on $\omega_a \setminus K$. As v vanishes on all edges $e \subset \gamma_a$ and all triangles except K, we have $\langle r_a, v\rangle = \langle r_a^T, v\rangle_K$, and thus

$$\sup_{\{v\in H^1(\omega_a): v\not\equiv C_{st}\}} \frac{\langle r_a, v\rangle}{\|\nabla v\|_{\omega_a}} \geqslant \sup_{\{v\in H_0^1(K): v\neq 0\}} \frac{\langle r_a^T, v\rangle_K}{\|\nabla v\|_K} = \sup_{\{v\in H_0^1(K): \|\nabla v\|=1\}} \langle \psi_a \left[f + \Delta U\right], v\rangle_K$$

An estimate for this last term can be found using equivalence of norms (cf. [11, Ex 9.x.5]):

That is, we claim that $\|q\| := \sup_{\{v\in H_0^1(K): \|\nabla v\|=1\}} \langle \psi_a q, v\rangle_K$ defines a norm for polynomials q of degree r on K. The only troublesome axiom to be satisfied is that $\|q\| = 0$ implies $q = 0$. Suppose that q is a non-zero polynomial. Then there exists a point $x \in \mathrm{int}(K)$ for which we have $q(x) \neq 0$, and without loss of generality we suppose that $q(x) > 0$. Since q is continuous, we have $q(y) > q(x)/2 > 0$ for y in $V(x)$, a small neighborhood of x. The hat function satisfies $\psi_a > 0$ in the interior of K. Taking $V(x)$ small enough therefore guarantees that $\psi_a q > 0$ on $V(x)$. Finally, choose a continuous $v \in H_0^1(K)$ with $v(x) = 1$ and $v = 0$ outside of $V(x)$. For this v we then find

$$\|q\| \geqslant \frac{\langle \psi_a q, v\rangle_K}{\|\nabla v\|_K} = \frac{\langle \psi_a q, v\rangle_{V(x)}}{\|\nabla v\|_K} > 0,$$

because $\psi_a q \neq 0 \neq v$ in a small non-empty open neighborhood of x. Hence $\|\cdot\|$ is a norm on the space of polynomials with degree r.

The above also holds for a reference element $\hat{K}$. Using equivalence of norms, and the usual transformation lemma shows that for all elements $K \subset \omega_a$ we have

$$\sup_{\{v\in H_0^1(K): \|\nabla v\|=1\}} \langle \psi_a \left[f + \Delta U\right], v\rangle_K \gtrsim h_K \|f + \Delta U\|_K \,,$$

for a constant depending on the shape regularity and (unfortunately) on the polynomial-degree p used in $\mathbb{V}$. This provides a lower bound for the element related terms.

Next, fix an edge $e \subset \gamma_a$ and denote K_1, K_2 for the triangles that share this edge. Consider the following space of functions:

$$V_e := \left\{ v \in H_0^1(K_1 \cup K_2) : \langle v, P\rangle_{K_1} = \langle v, P\rangle_{K_2} = 0 \quad \forall P \in \mathcal{P}_r \right\}.$$

Again, we can naturally extend these functions to live in $H^1(\omega_a)$. Since $v \in V_e$ vanish on all edges except e, and is orthogonal to polynomials of degree r on K_1 and K_2, we have

$$\sup_{\{v\in H^1(\omega_a): v\not\equiv C_{st}\}} \frac{\langle r_a, v\rangle}{\|\nabla v\|_{\omega_a}} \geqslant \sup_{\{v\in V_e: v\neq 0\}} \frac{\langle r_a^e, v\rangle_e}{\|\nabla v\|_{K_1\cup K_2}} = \sup_{\{v\in V_e: v\neq 0\}} \frac{\langle \psi_a [\![\nabla U]\!], v\rangle_e}{\|\nabla v\|_{K_1\cup K_2}} \gtrsim h_e^{1/2} \|[\![\nabla U]\!]\|_e \,.$$

The last inequality is a variation of a well known result [11, Ex 9.x.7]. It can be proven by using the same technique we used for the element terms. The constant depends on both shape regularity and the polynomial-degree p.

The last step consist of summing the above inequalities over $K \subset \omega_a$ and $e \subset \gamma_a$, i.e.

$$\begin{aligned}
\|\underline{\sigma_a}\|_{\omega_a} \geqslant \|r_a\|_{H^1_\star(\omega_a)'} &\gtrsim \sum_{K\subset\omega_a} h_K \|f + \Delta U\|_K + \sum_{e\subset\gamma_a} h_e^{1/2} \|[\![\nabla U]\!]\|_e \\
&\geqslant \left(\min_{K\subset\omega_a} h_K\right) \sum_{K\subset w_a} \|f + \Delta U\|_K + \left(\min_{e\subset\gamma_a} h_e^{1/2}\right) \sum_{e\subset\gamma_a} \|[\![\nabla U]\!]\|_e \\
&\gtrsim h_{\omega_a} \|f + \Delta U\|_{\omega_a} + h_{\omega_a}^{1/2} \|[\![\nabla U]\!]\|_{\gamma_a} .
\end{aligned}$$

The hidden constants depend the shape regularity and the polynomial-degree p. The proof is similar for $a \in \mathcal{V}^{bdr}$. □

Bibliography

[1] M. Ainsworth, G. Andriamaro, and O. Davydov. "A Bernstein–Bézier Basis for Arbitrary Order Raviart–Thomas Finite Elements". In: *Constructive Approximation* 41.1 (2015), pp. 1–22.

[2] M. Ainsworth, G. Andriamaro, and O. Davydov. "Bernstein-Bézier finite elements of arbitrary order and optimal assembly procedures". In: *SIAM Journal on Scientific Computing* 33.6 (2011), pp. 3087–3109.

[3] M. Ainsworth and J. Coyle. "Hierarchic finite element bases on unstructured tetrahedral meshes". In: *International Journal for Numerical Methods in Engineering* 58.14 (2003), pp. 2103–2130.

[4] M. Alnæs et al. "The fenics project version 1.5". In: *Archive of Numerical Software* 3.100 (2015), pp. 9–23.

[5] S. Bartels and C. Carstensen. "Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids. Part II: Higher order FEM". In: *Mathematics of computation* 71.239 (2002), pp. 971–994.

[6] P. Binev, W. Dahmen, and R. DeVore. "Adaptive finite element methods with convergence rates". In: *Numerische Mathematik* 97.2 (2004), pp. 219–268.

[7] D. Boffi, F. Brezzi, and M. Fortin. *Mixed finite element methods and applications*. Vol. 44. Springer, 2013.

[8] D. Braess. *Finite elements: Theory, fast solvers, and applications in solid mechanics*. Cambridge University Press, 2007.

[9] D. Braess, V. Pillwein, and J. Schöberl. "Equilibrated residual error estimates are p-robust". In: *Computer Methods in Applied Mechanics and Engineering* 198.13 (2009), pp. 1189–1197.

[10] D. Braess and J. Schöberl. "Equilibrated residual error estimator for edge elements". In: *Mathematics of Computation* 77.262 (2008), pp. 651–672.

[11] S. Brenner and R. Scott. *The mathematical theory of finite element methods*. Vol. 15. Springer Science & Business Media, 2007.

[12] J. M. Cascón and R. H. Nochetto. "Quasioptimal cardinality of AFEM driven by nonresidual estimators". In: *IMA Journal of Numerical Analysis* 32.1 (2012), pp. 1–29.

[13] J. M. Cascon et al. "Quasi-optimal convergence rate for an adaptive finite element method". In: *SIAM Journal on Numerical Analysis* 46.5 (2008), pp. 2524–2550.

[14] L. Chen. “iFEM: an innovative finite element methods package in MATLAB”. In: *Preprint, University of Maryland* (2008).

[15] Z. Chen and H. Wu. *Selected topics in finite element methods*. Science Press, Beijing, 2010.

[16] V. Dolejší, A. Ern, and M. Vohralík. “hp-adaptation driven by polynomial-degree-robust a posteriori error estimates for elliptic problems”. In: (2015).

[17] W. Dörfler. “A convergent adaptive algorithm for Poisson’s equation”. In: *SIAM Journal on Numerical Analysis* 33.3 (1996), pp. 1106–1124.

[18] M. Dubiner. “Spectral methods on triangles and other domains”. In: *Journal of Scientific Computing* 6.4 (1991), pp. 345–390.

[19] A. Ern and M. Vohralík. “Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations”. In: *SIAM Journal on Numerical Analysis* 53.2 (2015), pp. 1058–1081.

[20] G. N. Gatica. *A simple introduction to the mixed finite element method: theory and applications*. Springer Science & Business Media, 2014.

[21] R. C. Kirby. “Algorithm 839: FIAT, a new paradigm for computing finite element basis functions”. In: *ACM Transactions on Mathematical Software (TOMS)* 30.4 (2004), pp. 502–516.

[22] C. Kreuzer and K. G. Siebert. “Decay rates of adaptive finite elements with Dörfler marking”. In: *Numerische Mathematik* 117.4 (2011), pp. 679–716.

[23] MATLAB. *version R2015b (8.6.0.267246)*. Natick, Massachusetts: The MathWorks Inc., 2015.

[24] K. Mekchay and R. H. Nochetto. “Convergence of adaptive finite element methods for general second order linear elliptic PDEs”. In: *SIAM Journal on Numerical Analysis* 43.5 (2005), pp. 1803–1827.

[25] P. Morin, R. H. Nochetto, and K. G. Siebert. “Convergence of adaptive finite element methods”. In: *SIAM review* 44.4 (2002), pp. 631–658.

[26] P. Morin, R. H. Nochetto, and K. G. Siebert. “Data oscillation and convergence of adaptive FEM”. In: *SIAM Journal on Numerical Analysis* 38.2 (2000), pp. 466–488.

[27] W. Prager and J. L. Synge. “Approximations in elasticity based on the concept of function space”. In: *Quart. Appl. Math* 5.3 (1947), pp. 241–269.

[28] P.-A. Raviart and J.-M. Thomas. “A mixed finite element method for 2-nd order elliptic problems”. In: *Mathematical aspects of finite element methods*. 1977.

[29] R. Rodríguez. “Some remarks on Zienkiewicz-Zhu estimator”. In: *Numerical Methods for Partial Differential Equations* 10.5 (1994), pp. 625–635.

[30] J. Schöberl. *NGSolve Finite Element Library*. `https://sourceforge.net/projects/ngsolve/`.

[31] J. Schöberl and S. Zaglmayr. “High order Nédélec elements with local complete sequence properties”. In: *COMPEL-The international journal for computation and mathematics in electrical and electronic engineering* 24.2 (2005), pp. 374–384.

[32] L. R. Scott and S. Zhang. “Finite element interpolation of nonsmooth functions satisfying boundary conditions”. In: *Mathematics of Computation* 54.190 (1990), pp. 483–493.

[33] R. Stevenson. “Optimality of a standard adaptive finite element method”. In: *Foundations of Computational Mathematics* 7.2 (2007), pp. 245–269.

[34] R. Stevenson. *Some notes with Adaptive Finite Elements*. `https://staff.fnwi.uva.nl/r.p.stevenson/notes1.pdf`.

[35] R. Stevenson. “The completion of locally refined simplicial partitions created by bisection”. In: *Mathematics of computation* 77.261 (2008), pp. 227–241.

[36] C. T. Traxler. “An algorithm for adaptive mesh refinement inn dimensions”. In: *Computing* 59.2 (1997), pp. 115–137.

[37] A. Veeser and R. Verfürth. “Poincaré constants for finite element stars”. In: *IMA Journal of Numerical Analysis* (2011), drr011.

[38] R. Verfürth. *A posteriori error estimation techniques for finite element methods*. Oxford University Press, 2013.

[39] R. Verfürth. *A review of a posteriori error estimation and adaptive mesh-refinement techniques*. John Wiley & Sons Inc, 1996.

[40] J. P. Webb. “Hierarchal vector basis functions of arbitrary order for triangular and tetrahedral finite elements”. In: *IEEE Transactions on Antennas and Propagation* 47.8 (1999), pp. 1244–1253.

[41] W. Zheng and H. Qi. “On Friedrichs–Poincaré-type inequalities”. In: *Journal of mathematical analysis and applications* 304.2 (2005), pp. 542–551.

[42] O. C. Zienkiewicz and J. Z. Zhu. “A simple error estimator and adaptive procedure for practical engineerng analysis”. In: *International Journal for Numerical Methods in Engineering* 24.2 (1987), pp. 337–357.

[43] O. C. Zienkiewicz and J. Z. Zhu. “The superconvergent patch recovery and a posteriori error estimates. Part 1: The recovery technique”. In: *International Journal for Numerical Methods in Engineering* 33.7 (1992), pp. 1331–1364.

[44] O. C. Zienkiewicz et al. *The finite element method*. Vol. 3. McGraw-hill London, 1977.