JUSTIFICATION OF THE SATURATION ASSUMPTION*

CARSTEN CARSTENSEN, DIETMAR GALLISTL, AND JOSCHA GEDICKE

ABSTRACT. The saturation assumption is widely used in computational science and engineering, usually without any rigorous theoretical justification and even despite of counterexamples for some coarse meshes known in the mathematical literature. On the other hand, there is overwhelming numerical evidence at least in an asymptotic regime for the validity of the saturation. In the generalized form, the assumption states, for any $0 < \varepsilon \leq 1$, that

(SA) $$|||u - \hat{U}|||^2 \leq (1 - \varepsilon/C)|||u - U|||^2 + \varepsilon \operatorname{osc}^2(f, \mathcal{N})$$

for the exact solution u and the first-order conforming finite element solution U (resp. $\hat{U}$) of the Poisson model problem with respect to a regular triangulation $\mathcal{T}$ (resp. $\hat{\mathcal{T}}$) and its uniform refinement $\hat{\mathcal{T}}$ within the class $\mathbb{T}$ of admissible triangulations. The point is that the patch-oriented oscillations $\operatorname{osc}(f, \mathcal{N})$ vanish for constant right-hand sides $f \equiv 1$ and may be of higher order for smooth f, while the strong reduction factor $(1 - \varepsilon/C) < 1$ involves some universal constant C which exclusively depends on the set of admissible triangulations and so on the initial triangulation only. This paper proves the inequality (SA) for the energy norms of the errors for any admissible triangulation $\mathcal{T}$ in $\mathbb{T}$ up to computable pathological situations characterized by failing the weak saturation test (WS). This computational test (WS) for some triangulation $\mathcal{T}$ states that the solutions U and $\hat{U}$ do not coincide for the constant right-hand side $f \equiv 1$. The set of possible counterexamples is characterized as $\mathcal{T}$ with no interior node or exactly one interior node which is the vertex of all triangles and $\hat{\mathcal{T}}$ is a particular uniform *bisec3* refinement. In particular, the strong saturation assumption holds for all triangulations with more than one degree of freedom. The weak saturation test (WS) is only required for zero or one degree of freedom and gives a definite outcome with $O(1)$ operations. The only counterexamples known so far are regular n-polygons. The paper also discusses a generalization to linear elliptic second-order PDEs with small convection to prove that saturation is somehow generic and fails only in very particular situations characterised by (WS).

1. Introduction

1.1. **Motivation.** It is well known that the criss-cross triangulation $\mathcal{T}_0$ of the unit square Ω is a counterexample to the saturation assumption if the refined mesh $\hat{\mathcal{T}}$ is refined everywhere by the newest vertex bisection (NVB) with refinement edges along the boundary $\partial\Omega$. The respective discrete solutions $U = \hat{U}$ for the Poisson model problem

(1.1) $\qquad -\operatorname{div}(A\nabla u) = f \text{ in } \Omega \quad \text{and} \quad u = 0 \text{ on } \partial\Omega$

coincide for the constant right-hand side $f \equiv 1$ and $A = 1_{2\times2}$ for the Laplace operator. This is often regarded as a pre-asymptotic artifact and contrasted with striking numerical evidence for the saturation assumption on finer triangulations. This paper provides a rigorous proof of this conjecture and characterises the very small set of counterexamples. For mesh-refinement techniques with an interior node property, the saturation assumption has been reasonably justified in [DN02], and is used in [AO00, BS93, Ver96, FLOP10]. This paper justifies the saturation assumption of [FLOP10, p.293] where it is warned that this assumption may fail to hold in general. A proof of the saturation assumption in the context of eigenvalue problems is included in [CGMM14] for sufficiently small mesh-sizes only.

The saturation assumption is established in [DN02] for a different situation with an increase of polynomial degrees from first to second order in the finite element spaces but on the same mesh. This increase of the finite element space allows for the same number of degrees of freedom as the mesh-refinements of this paper and hence appears competitive from a practical point of view. Despite the fact that [DN02] observe that *red* refinement leads to saturation in one example, they conclude that *quadratics do indeed encode finer information than refined linears* in [DN02, Remark 3.2] in view of the counterexample of the criss-cross triangulation of the unit square for *bisec3* refinement. The main results I and II of this paper, however, prove this statement in the negative and point out that piecewise quadratics possibly encode finer information on the oscillations than refined piecewise linear conforming finite element schemes; but the two improved solutions enjoy a similar saturation property up to an extremely limited number of geometries exclusively for the very first mesh with at most one degree of freedom.

The saturation property has to be considered in comparison to the error estimator reduction in the convergence analysis of adaptive finite element methods [CKNS08, Ste08]. In explicit residual-based error estimators, the mesh-size enters as a weight and hence reduces under refinement. This implies a reduction property of such error estimators and eventually leads to linear convergence of some total error which is a convex combination of the error estimator and the error. In contrast to this, the saturation property describes the reduction (SA) of the error terms without involving any error estimator contribution, but with immediate important applications in the context of hierarchical error estimators. The proofs are rather independent, e.g., the saturation property (SA) cannot be proved by simply reducing the mesh-size.

Let $\mathcal{T}$ be a shape-regular triangulation of Ω into triangles with the set of nodes $\mathcal{N}$ and the set of edges $\mathcal{E}$. Let $P_1(\mathcal{T})$ denote the piecewise linear
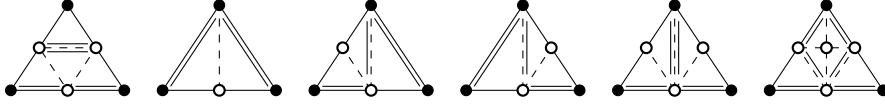
FIGURE 1. Refinement rules *red, green, blue-left, blue-right, bisec3* and *bisec5*. The reference edge, the bottom edge, is always refined. The new reference edges for the sub-triangles are indicated with a second line.

polynomials with respect to $\mathcal{T}$ and let the finite element space $V(\mathcal{T}) := P_1(\mathcal{T}) \cap H_0^1(\Omega)$ consist of all piecewise linear functions which are globally continuous and vanish along the boundary $\partial\Omega$.

Throughout this paper, let $A \in \mathbb{R}^{2\times 2}$ denote a symmetric positive definite constant diffusion matrix and let $\|\cdot\| := \|A^{1/2}\nabla\cdot\|_{L^2(\Omega)}$ be the induced energy norm in $V \equiv H_0^1(\Omega)$. The discrete problem seeks a piecewise linear function $u_\mathcal{T} \in V(\mathcal{T})$ such that

$$(1.2) \qquad \int_\Omega (A\nabla u_\mathcal{T}) \cdot \nabla v_\mathcal{T} \, dx = \int_\Omega f v_\mathcal{T} \, dx \quad \text{for all } v_\mathcal{T} \in V(\mathcal{T}).$$

Given an initial regular triangulation $\mathcal{T}_0$ of Ω into triangles with at least one interior vertex and the set of all admissible refinements $\mathbb{T}$ by successive application of the refinement rules from Figure 1 (see Definition 2.3 for more details) the main result concerns two notions of saturation for a triangulation $\mathcal{T}$ and its refinement $\hat{\mathcal{T}}$ where each edge is bisected and each triangle $T \in \hat{\mathcal{T}}$ is obtained by *red* or *bisec3* refinement as illustrated in Figure 1; written $\hat{\mathcal{T}} \in \text{unif}(\mathcal{T})$.

1.2. **Strong saturation.** Strong saturation for $\mathcal{T} \in \mathbb{T}$ and some uniform refinement $\hat{\mathcal{T}} \in \text{unif}(\mathcal{T})$ means that, for any $0 < \varepsilon \leq 1$, there exists $\varrho(\varepsilon) := 1 - \varepsilon/C(\mathcal{T}_0) < 1$, with a universal constant $C(\mathcal{T}_0)$ which exclusively depends on $\mathcal{T}_0$, such that: Given any right-hand side $f \in L^2(\Omega)$, the exact solution u of (1.1) and the discrete solution $U := u_\mathcal{T} \in V(\mathcal{T})$ (resp. $\hat{U} := u_{\hat{\mathcal{T}}} \in V(\hat{\mathcal{T}})$) of the discrete problem (1.2) with respect to $\mathcal{T}$ (resp. $\hat{\mathcal{T}}$) satisfy

$$(\text{SA}) \qquad \|u - \hat{U}\|^2 \leq \varrho(\varepsilon)\|u - U\|^2 + \varepsilon \, \text{osc}^2(f, \mathcal{N})$$

for the patch-oriented oscillations

$$(1.3) \qquad \text{osc}^2(f, \mathcal{N}) := \sum_{z \in \mathcal{N}(\Omega)} \text{diam}(\Omega_z)^2 \|f - \textstyle\fint_{\Omega_z} f \, dx\|_{L^2(\Omega_z)}^2$$

with the integral mean $\fint_{\Omega_z} f \, dx$ of f over the extended nodal patch Ω_z; more details on the oscillations follow in Section 3.

1.3. **Weak saturation.** Weak saturation for $\mathcal{T}$ and $\hat{\mathcal{T}} \in \text{unif}(\mathcal{T})$ means that for the constant right-hand side $f \equiv 1$, the discrete solutions $U \in V(\mathcal{T})$ and $\hat{U} \in V(\hat{\mathcal{T}})$ are different,

$$(\text{WS}) \qquad U \neq \hat{U}, \quad \text{and so} \quad \|u - \hat{U}\| < \|u - U\|.$$

The strong saturation property (SA) is a frequent assumption that a mesh-refinement procedure will eventually lead to q-linear convergence of the approximate finite element solution to the exact solution.

1.4. **Main results.**

Main Result I. *For some global constant $C(\mathcal{T}_0)$ which exclusively depends on $\mathcal{T}_0$ and given any $0 < \varepsilon \le 1$, for any $\mathcal{T} \in \mathbb{T}$ and $\hat{\mathcal{T}} \in \mathrm{unif}(\mathcal{T})$, (WS) implies (SA) with $\varrho(\varepsilon) = 1 - \varepsilon/C(\mathcal{T}_0)$.*

It appears interesting that the proof combines hard analysis (i.e., direct estimation with explicit constants) and soft analysis (i.e. functional analysis with compactness principles and unknown constants). The hard analysis concerns first the situation where the triangulation has some edge with a positive distance from the boundary $\partial\Omega$. This leads to a constant $C(\mathcal{T}_0)$ which depends only on a lower bound of the minimum angle $\min \angle \mathcal{T}_0$ in $\mathcal{T}_0$ (and hence in $\mathbb{T}$). The remaining situations are determined by a finite number of configurations like $\mathcal{T}_0$ and possibly a few others. For each of those pairs $\mathcal{T}$ and $\hat{\mathcal{T}}$, a compactness and equivalence of norm argument provides the assertion. The constants in the soft analysis depend very much on $\mathcal{T}$ and $\hat{\mathcal{T}}$ and of course on $\mathcal{T}_0$. The maximum of all those constants in the finite number of possible geometries in the second scenario concludes the proof.

Weak saturation is almost always true and fails only for one very particular situation with $\dim V(\mathcal{T}) = 1$ and one particular *bisec3* refinement pattern.

Main Result II. *Weak saturation can only fail for $\mathcal{T}$ and $\hat{\mathcal{T}} \in \mathrm{unif}(\mathcal{T})$ if $\mathcal{T}$ has exactly one interior node z and $\hat{\mathcal{T}}$ is obtained by* bisec3 *for each triangle $T \in \mathcal{T}(z) = \mathcal{T}$ such that the refinement edge of T is opposite to the vertex z on the boundary $\partial\Omega$.*

Notice that the exceptional case is with $\dim V(\mathcal{T}) = 1$ and *bisec3* refinement where all refinement edges are opposite to the free node. This case can be easily checked without difficulty for the geometry at hand. The known exceptional cases are regular polygons from [BC08] which include the criss-cross unit square. It is left as an open question whether there exist other counterexamples for $A = 1_{2\times 2}$.

1.5. **Outline.** The remaining parts of this paper are organised as follows. Section 2 studies a metric on the set of edges in a regular triangulation and thereby quantifies a uniform bound for the distance of some interior edge to a compactly interior edge. Section 3 establishes the strong saturation for all triangulations that contain one edge with positive distance to the boundary. Section 4 proves that weak saturation implies strong saturation. A characterisation of triangulations that allow for weak saturation follows in Section 5. Section 6 discusses the extension to general elliptic linear second-order PDEs with small convection. It is surprising that the weak saturation test applies to the main part (1.1)–(1.2) only where the coefficients of the lower-order terms have no influence.

Throughout this paper, standard notation on Lebesgue and Sobolev spaces and their norms is employed. The L^2 projection onto piecewise polynomials of degree $k \in \mathbb{N}_0$ is denoted by Π_k. The energy norm is denoted by $\|\cdot\| := \|A^{1/2}\nabla\cdot\|_{L^2(\Omega)}$. The formula $a \lesssim b$ represents an inequality $a \le Cb$ for some mesh-independent, positive generic constant C; $a \approx b$ abbreviates

$a \lesssim b \lesssim a$. The measure $|\cdot|$ is context-sensitive and refers to the number of elements of some finite set (e.g. the number $|\mathcal{T}|$ of triangles in a triangulation $\mathcal{T}$) or the length $|E|$ of an edge E or the area $|T|$ of some domain T and not just the modulus of a real number or the Euclidean length of a vector.

2. Edge-Connectivity

This section studies a metric on the set of interior edges for admissible triangulations to prove that the length of some polygonal path as in Figure 2 that links an arbitrary interior edge F to some edge F' which lies compactly in the domain through a finite chain of interior edges allows for some global bound of the number of edges in this chain. The technical challenge of this section consists in the large class of admissible triangulations for rather general refinements as depicted in Figure 1. Details are stated as Theorem 2.1 below, right after all the necessary notation is set up. It turns out that the aforementioned global bound and the shape-regularity determine the constant $C(\mathcal{T}_0)$ indicated in the introduction as the main result of this section.

Definition 2.1 (nodes and edges). *Given a regular triangulation $\mathcal{T}$, denote the set of edges by $\mathcal{E}$ and the set of nodes by $\mathcal{N}$. Let $\mathcal{N}(\Omega)$ and $\mathcal{E}(\Omega)$ denote the sets of interior nodes and interior edges. For an interior edge $E = \partial T_+ \cap T_- \in \mathcal{E}(\Omega)$ shared by two triangles T_+ and T_-, the edge-patch is defined as $\omega_E := \mathrm{int}(T_+ \cup T_-)$. Given a triangle $T \in \mathcal{T}$, denote its set of edges by $\mathcal{E}(T)$ and its set of nodes by $\mathcal{N}(T)$. For any edge $E = \mathrm{conv}\{z_1, z_2\} \in \mathcal{E}$, the set of endpoints reads $\mathcal{N}(E) = \{z_1, z_2\}$. Define the set of compactly interior edges as*

$$\mathcal{E}_c(\Omega) := \{E \in \mathcal{E}(\Omega) \mid E \cap \partial\Omega = \emptyset\}$$

and notice that E and $\partial\Omega$ are compact sets and so $E \cap \partial\Omega = \emptyset$ means that $E \subset\subset \Omega$ lies compactly in Ω with $\mathrm{dist}(E, \partial\Omega) > 0$. The set of interior edges whose two endpoints belong to the boundary $\partial\Omega$ reads

$$\mathcal{E}_b(\Omega) := \{E \in \mathcal{E}(\Omega) \mid \mathcal{N}(E) \subset \mathcal{N}(\partial\Omega)\}.$$

The set $\mathcal{E}_a(\Omega)$ of interior edges that belong to at least one triangle with an interior node is

$$\mathcal{E}_a(\Omega) := \{E \in \mathcal{E}(\Omega) \mid \exists T \in \mathcal{T} \text{ with } \mathcal{N}(T) \cap \mathcal{N}(\Omega) \neq \emptyset \text{ and } E \in \mathcal{E}(T)\}.$$

Denote by $\mathcal{E}_0$ (resp. $\mathcal{E}_0(\Omega)$) the set of edges (resp. interior edges) of the coarse initial triangulation $\mathcal{T}_0$. For an interior edge $E = \partial T_+ \cap T_- \in \mathcal{E}_0(\Omega)$ shared by two triangles $T_+, T_- \in \mathcal{T}_0$, the edge-patch reads $\omega_E^{(0)} := \mathrm{int}(T_+ \cup T_-)$.

Definition 2.2 (metric δ). *Assume that $\mathcal{E}(\Omega) \neq \emptyset$ and define a metric δ on the set $\mathcal{E}(\Omega)$ of interior edges, for $F, F' \in \mathcal{E}(\Omega)$ by*

$$\delta(F, F') := \min\left\{ J \in \mathbb{N}_0 \, \middle| \, \begin{array}{l} \exists F_1, \ldots, F_{J+1} \in \mathcal{E}(\Omega) \text{ with } F_1 = F, F_{J+1} = F' \\ \text{and } \forall j = 1, \ldots, J, \ F_j \cap F_{j+1} \neq \emptyset \end{array} \right\}.$$

Furthermore, let

$$\delta(F, \mathcal{E}_c(\Omega)) := \min\{\delta(F, F') \mid F' \in \mathcal{E}_c(\Omega)\}.$$
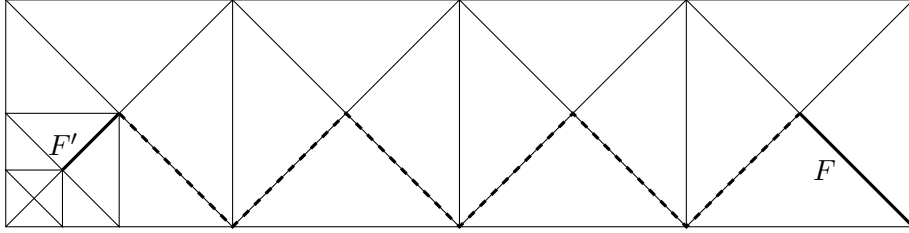
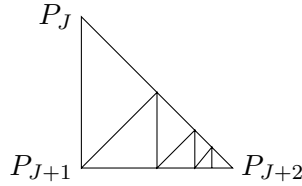FIGURE 2. An edge F and a compactly interior edge F' (thick) and a possible connecting path (dashed).

FIGURE 3. Coarse triangle $K \in \mathcal{T}_0$ and refinement without any compactly interior edge $\mathcal{E}_c(\mathrm{int}(K)) = \emptyset$ and $|\mathcal{E}_b(\mathrm{int}(K))| = 6$. Further refinements of this kind towards P_{J+2} prove that $|\mathcal{E}(\mathrm{int}(K))|$ can be arbitrarily large while $\mathcal{E}_c(\mathrm{int}(K)) = \emptyset$.

Example 2.1 (distances can be large). The example triangulation of Figure 2 shows $\delta(F, \mathcal{E}_c(\Omega)) = 7$. Further refinements towards the lower left corner of this rectangular domain indicate that the number of triangles in $\mathcal{T}$ may not be bounded by a universal constant which depends only on $\mathcal{T}_0$. At the same time, the edge-path which connects interior edges F and F' can be extremely large (add more and more of the criss-cross squares in the middle for a modified $\mathcal{T}_0$). Despite the warnings of this example, the number $\max_{F \in \mathcal{E}_a(\Omega)} \delta(F, \mathcal{E}_c(\Omega)) \leq C_1(\mathcal{T}_0)$ is bounded in terms of $\mathcal{T}_0$ for a broad class of admissible triangulations defined below in Definition 2.3.

Example 2.2 (no uniform bound for edges in $\mathcal{E}_b(\Omega)$). Figure 3 displays one critical triangle of a family of triangulations for which $\max_{F \in \mathcal{E}(\Omega)} \delta(F, \mathcal{E}_c(\Omega))$ is unbounded. Indeed, if the edge shared by nodes P_{J+1} and P_{J+2} and the edge shared by nodes P_{J+2} and P_J in Figure 3 belong to the boundary $\partial \Omega$, then arbitrarily many edges $\mathcal{E}_b(\Omega)$ may be added through refinement without changing the underlying finite element space.

Definition 2.3 (admissible triangulations). *Let $\mathcal{T}_0$ be a regular triangulation. For each $T \in \mathcal{T}_0$, one chooses one of its edges $\mathcal{E}(T)$ as a reference edge from the set of edges $\mathcal{E}_0$. The set $\mathbb{T} := \mathbb{T}(\mathcal{T}_0)$ of admissible triangulations contains all regular triangulations $\mathcal{T}$ that are refined from $\mathcal{T}_0$ with the refinement rules of Figure 1, where the new reference edges for the sub-triangles are indicated by a second line.*

Theorem 2.1. *There exists some constant $C_1(\mathcal{T}_0) < \infty$ such that any admissible triangulation $\mathcal{T} \in \mathbb{T}$ with the set $\mathcal{E}_c(\Omega) \neq \emptyset$ of compactly interior*
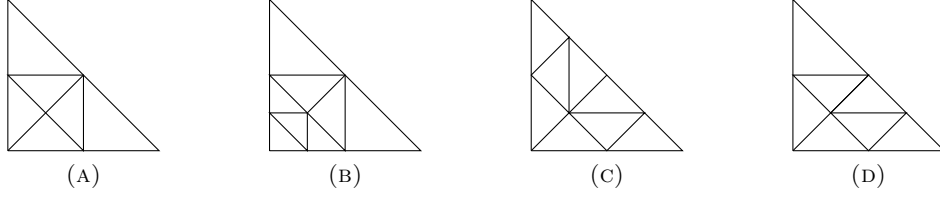
FIGURE 4. All possible coarsest refinements of K with one interior node (up to the reflection of Figure D across the diagonal line).

edges satisfies

$$\max_{F \in \mathcal{E}_a(\Omega)} \delta(F, \mathcal{E}_c(\Omega)) \leq C_1(\mathcal{T}_0).$$

The proof of Theorem 2.1 combines topological arguments (connecting edge-paths) with local geometrical facts (refinement rules for one initial triangle). The latter are summarized in the following lemma.

Lemma 2.1. *Let $K \in \mathcal{T}_0$ be a triangle of the initial triangulation $\mathcal{T}_0$ and let $\mathcal{T} \in \mathbb{T}$ denote an admissible triangulation with nodes $\mathcal{N}$ and edges $\mathcal{E}$.*

(i) *If $\mathcal{N}(\text{int}(K)) \neq \emptyset$, then $\mathcal{E}(K) \cap \mathcal{E} = \emptyset$. In other words, none of the edges of K belongs to $\mathcal{E}$.*

(ii) *$\mathcal{E}_c(\text{int}(K)) = \emptyset$ if and only if $|\mathcal{N}(\text{int}(K))| \leq 1$.*

(iii) *If $\mathcal{E}_c(\text{int}(K)) \neq \emptyset$ and $P \in \mathcal{N}(\text{int}(K))$, then $\mathcal{E}_c(\text{int}(K)) \cap \mathcal{E}(P) \neq \emptyset$.*

(iv) *If $\mathcal{N}(\text{int}(K)) \neq \emptyset$ and $E = (F_1 \cup F_2) \in \mathcal{E}(K) \subseteq \mathcal{E}_0$ is bisected into $F_1, F_2 \in \mathcal{E} \setminus \mathcal{E}_0$, then there exists some $y \in \mathcal{N}(\text{int}(K))$ such that*

$$\text{conv}\{y, \text{mid}(E)\} \in \mathcal{E}(\text{int}(K)).$$

Proof. A careful discussion of the refinement rules reveals that any triangulation of K with at least one interior node is some refinement of one of the triangulations of Figure 4. This proves (i).

Due to Definition 2.3 any further refinements of triangulations from Figure 4 with exactly one interior node has to bisect a triangle of $\mathcal{T}(\mathcal{N}(\text{int}(K)))$. Direct investigation of these possible bisections in the triangulations of Figure 4 prove (ii).

Suppose that all edges that contain the interior node $P \in \text{int}(K)$ are not compactly interior edges in K. Then, their respective endpoints belong to the boundary ∂K of the triangle K. All possibilities for this situation are displayed in Figure 4 and imply that $\mathcal{N}(\text{int}(K))$ contains exactly the vertex P. Then, $\mathcal{E}_c(\text{int}(K)) = \emptyset$. The contraposition implies (iii).

Direct investigations first verify (iv) for the triangulations of Figure 4 and second for their refinements. $\square$

Proof of Theorem 2.1. Suppose that $\mathcal{E}_c(\Omega) \neq \emptyset$. Any $F \in \mathcal{E}_a(\Omega) \cap \mathcal{E}_b(\Omega)$ is connected to some edge in $\mathcal{E}_a(\Omega) \setminus \mathcal{E}_b(\Omega)$ by an edge-path of length smaller or equal to 1. Therefore, without loss of generality, suppose that $F \in \mathcal{E}_a(\Omega) \setminus (\mathcal{E}_b(\Omega) \cup \mathcal{E}_c(\Omega))$. Hence, $F = \text{conv}\{P, Q\}$ for $P \in \mathcal{N}(\Omega)$ and $Q \in \mathcal{N}(\partial\Omega)$.

In the *first* step of the proof suppose that $P \in \mathcal{N}(\text{int}(K))$ for some $K \in \mathcal{T}_0$. Suppose that $\mathcal{E}_c(\text{int}(K)) \neq \emptyset$ and, thus, by Lemma 2.1.iii, P is connected to

some interior node $\mathcal{N}(\Omega)$. This together with Lemma 2.1.iii implies that P belongs to some edge in $\mathcal{E}_c(\Omega)$, $\delta(F, \mathcal{E}_c(\Omega)) = 1 \leq 3 + 2|\mathcal{E}_0(\Omega)|$. Otherwise, $\mathcal{E}_c(\text{int}(K)) = \emptyset$ and all neighbouring nodes of P in $\mathcal{T}$ belong to the boundary $\partial\Omega$ of the domain Ω. Lemma 2.1.ii implies that P is the only interior node of K, thus K is a refinement of the triangulations in Figure 4. Since all neighbouring nodes of P in $\mathcal{T}$ belong to the boundary, $\Omega = \text{int}(K)$. This contradicts $\mathcal{E}_c(\Omega) \neq \emptyset$.

In the *second* step of the proof suppose that P does not belong to exactly one $K \in \mathcal{T}_0$. Hence, P belongs to a common edge $E_1 \in \mathcal{E}_0(\Omega)$ of two coarse triangles. Since there exists $F' \in \mathcal{E}_c(\Omega)$, the edge-wise connectivity of Ω implies the existence of a finite set $E_1, \ldots, E_J \in \mathcal{E}_0(\Omega)$, $1 \leq J \leq |\mathcal{E}_0(\Omega)|$, of interior edges in the coarse triangulation $\mathcal{T}_0$ such that E_1 is as above with P lays on E_1 and $E_1, \ldots, E_J \in \mathcal{E}_0(\Omega)$ is a polygon with $\mathcal{N}(E_j) \cap \mathcal{N}(E_{j+1}) \neq \emptyset$ for $j = 1, \ldots J$ and the topological closure of $\omega_{E_J}^{(0)}$ contains F'. Without loss of generality let J be a minimal choice such that F' does not belong to the topological closure of $\omega_{E_1}^{(0)} \cup \cdots \cup \omega_{E_{J-1}}^{(0)}$ (understood as the empty set if $J = 1$). Moreover, suppose that $E_j = \text{conv}\{P_j, P_{j+1}\}$ for all $j = 1, \ldots, J$ for piecewise distinct nodes $P_1, \ldots, P_{J+1} \in \mathcal{N}_0$. In this situation, one designs some edge-path from $F = \text{conv}\{P, Q\}$ to $\mathcal{E}_c(\Omega)$ as follows. The edge $E_1 = \text{conv}\{P_1, P_2\} \in \mathcal{E}_0(\Omega) \setminus \mathcal{E}(\Omega)$ is refined and, hence, there exist pairwise distinct $F_1^{(1)}, \ldots, F_1^{(k_1)} \in \mathcal{E}(\Omega)$ with $k_1 \in \mathbb{N}$ and $F_1^{(j)} \cap F_1^{(j-1)} \neq \emptyset$ for all $j = 2, \ldots, k_1$ and $\text{conv}\{P, P_2\} = F_1^{(1)} \cup \cdots \cup F_1^{(k_1)}$. Suppose $P \in \mathcal{N}(F_1^{(1)})$ and $P_2 \in \mathcal{N}(F_1^{(k_1)})$ such that $(F, F_1^{(1)}, \ldots, F_1^{(k_1)})$ is an edge-path in $\mathcal{T}$ from Q to P_2. For any $j = 2, \ldots, J$, let

$$E_j = F_j^{(1)} \cup \cdots \cup F_j^{(k_j)} = \text{conv}\{P_j, P_{j+1}\}$$

for distinct $F_j^{(1)}, \ldots, F_j^{(k_j)}$ in $\mathcal{E}(\Omega)$ such that $(F_j^{(1)}, \ldots, F_j^{(k_j)})$ defines an edge-path from P_j to P_{j+1} along E_j in the fine triangulation $\mathcal{T}$. This composes to an edge-path from Q to P_{J+1} in $\mathcal{T}$.

Suppose that $F' \subseteq E_J$. In case $J = 1$, P is connected to F' and therefore $\delta(F, \mathcal{E}_c(\Omega)) = 1 \leq 3 + 2|\mathcal{E}_0(\Omega)|$. Otherwise $J \geq 2$ and $k_1 \in \{0, 1\}$. Since J is minimal, $k_2, \ldots, k_{J-1} \in \{1, 2\}$ and $F_J^{\min(k_J, 2)} \in \mathcal{E}_c(\Omega)$ implies that the edge-path

$$(F, F_1^{(k_1)}, F_2^{(1)}, F_2^{(k_2)}, F_3^{(1)}, \ldots, F_{J-1}^{(k_{J-1})}, F_J^{(1)}, F_J^{\min(k_J, 2)})$$

connects F to $\mathcal{E}_c(\Omega)$ in $\mathcal{T}$. This and $J \leq |\mathcal{E}_0(\Omega)|$ prove

$$\delta(F, \mathcal{E}_c(\Omega)) = \sum_{\ell=1}^{J-1} k_\ell + \min(k_J, 2) - 1 \leq 2J - 1 \leq 3 + 2|\mathcal{E}_0(\Omega)|.$$

In the remaining case F' is contained in $\omega_{E_J}^{(0)}$ but $F' \not\subseteq E_J$. Since J is minimal, $k_1 \in \{0, 1\}$ and $k_2, \ldots, k_{J-1} \in \{1, 2\}$ and so the edge-path $(F, F_1^{(1)}, \ldots, F_{J-1}^{(k_{J-1})})$ connects Q with $P_J = \mathcal{N}(E_{J-1}) \cap \mathcal{N}(E_J)$ and has length smaller than or equal to $2J$. Recall that $F' \in \mathcal{E}_c(\Omega)$ belongs to some $K = \text{conv}\{P_J, P_{J+1}, P_{J+2}\} \in \mathcal{T}_0$ with edge $E_J = \text{conv}\{P_J, P_{J+1}\}$ and opposite vertex $P_{J+2} \in \mathcal{N}_0$. The conclusion of the proof consists in the design of some edge-path $F_J^{(1)}, \ldots, F_J^{(k_J)}$ in $\mathcal{T}$ which connects P_J and $F_J^{(k_J)} \in \mathcal{E}_c(\Omega)$

with $k_J \leq 4$. Indeed, this implies that the edge-path $(F, F_1^{(1)}, \dots, F_J^{(k_J)})$ connects F and $\mathcal{E}_c(\Omega)$ and proves

$$\delta(F, \mathcal{E}_c(\Omega)) \leq 2J + k_J - 1 \leq 3 + 2J \leq 3 + 2|\mathcal{E}_0(\Omega)|.$$

Three cases have to be considered to design such an edge-path $F_J^{(1)}, \dots, F_J^{(k_J)}$ with $k_J \leq 4$.

(a) In the first case assume that $F' \in \mathcal{E}_c(\mathrm{int}(K))$. Hence, $\mathcal{N}(\mathrm{int}(K)) \neq \emptyset$ and Lemma 2.1.i implies that all edges of K are at least bisected. Since $F' \not\subseteq E_J$, E_J is at most bisected once. Thus, Lemma 2.1.iv leads to an edge-path of length $k_J = 2$ with $F_J^{(2)} = \mathrm{conv}\{y, \mathrm{mid}(E_J)\} \in \mathcal{E}_c(\Omega)$, for some $y \in \mathcal{N}(\mathrm{int}(K))$.

(b) In the second case assume $F' \subseteq \partial K \setminus E_J$. If $F' \subseteq \mathrm{conv}\{P_J, P_{J+2}\}$, one finds an edge-path from P_J to $\mathcal{E}_c(\Omega)$ of length $k_J \leq 2$. It remains the case that $F' \subseteq \mathrm{conv}\{P_{J+1}, P_{J+2}\}$. Since E_J is at most bisected once, an edge-path of length $k_J \leq 4$ connects P_J with $\mathcal{E}_c(\Omega)$.

(c) In the remaining case suppose $F' \in \mathcal{E}_c(\Omega) \setminus \mathcal{E}_c(\mathrm{int}(K))$ and so there is an edge $E \in \mathcal{E}(K)$ of K with $F' \cap E \neq \emptyset$. Moreover, F' has at least one vertex on the boundary ∂K of K and, hence, E is an interior edge $E \in \mathcal{E}(\Omega)$. If $E = E_J$ or $E = \mathrm{conv}\{P_J, P_{J+2}\}$, this leads to a path of length $k_J \leq 2$. If $E = \mathrm{conv}\{P_{J+1}, P_{J+2}\}$, the fact that E_J is at most bisected once, leads to an edge-path of length $k_J \leq 4$. $\qquad\square$

3. Discrete Efficiency

This section introduces the discrete efficiency of the explicit edge-residual based a posteriori error estimator η in the context of (1.1)–(1.2). For any interior edge $E \in \mathcal{E}(\Omega)$, there exist two adjacent triangles $T_+, T_- \in \mathcal{T}$ such that $E = \partial T_+ \cap \partial T_-$ and $\omega_E := \mathrm{int}(T_+ \cup T_-)$. Let ν_E denote the fixed normal vector of E that points from T_+ to T_-. Given any (possibly vector-valued) function v, define the jump of v across E by $[v]_E := v|_{T_+} - v|_{T_-}$ with the traces $v|_{T_+}$ and $v|_{T_-}$ on E with length $|E|$.

Define $\eta := \eta(\mathcal{E}(\Omega)) := \sqrt{\eta^2(\mathcal{E}(\Omega))}$ by

$$\eta^2(E) := |E| \|[A\nabla U]_E \cdot \nu_E\|_{L^2(E)}^2 \quad \text{for any } E \in \mathcal{E}(\Omega) \text{ and}$$

$$\eta^2(\mathcal{F}) := \sum_{E \in \mathcal{F}} \eta^2(E) \qquad \text{for any subset } \mathcal{F} \subseteq \mathcal{E}(\Omega).$$

A refined analysis of the results from [CV99, Rod94] shows that the error estimator η is reliable and efficient up to oscillations. The definition of the oscillations relies on the following extended nodal patches of [CB02].

Definition 3.1 (extended nodal patch)**.** *Let $(\varphi_z \mid z \in \mathcal{N})$ denote the nodal basis of $P_1(\mathcal{T}) \cap H^1(\Omega)$ with $\varphi_z(z) = 1$ and $\varphi_z(y) = 0$ for all $y \in \mathcal{N} \setminus \{z\}$. Assume that there is a map $\zeta : \mathcal{N} \to \mathcal{N}(\Omega)$ which assigns to each vertex $z \in \mathcal{N}$ an interior vertex $\zeta(z) \in \mathcal{N}(\Omega)$ where $\zeta(z) = z$ for all $z \in \mathcal{N}(\Omega)$. Define the functions*

$$\psi_z := \sum_{\substack{y \in \mathcal{N} \\ y = \zeta(z)}} \varphi_z \quad \text{for any } z \in \mathcal{N}(\Omega).$$

The functions $(\psi_z \mid z \in \mathcal{N}(\Omega))$ define a partition of unity. For each interior vertex $z \in \mathcal{N}(\Omega)$, the extended nodal patch reads

$$\Omega_z := \{x \in \Omega \mid 0 < \psi_z(x)\}.$$

Throughout this paper we assume that Ω_z is connected for any $z \in \mathcal{N}(\Omega)$.

For the extended nodal patch Ω_z and the integral mean $\fint_{\Omega_z} f\,dx$ of the right-hand side $f \in L^2(\Omega)$ of (1.1), the oscillations read

$$\mathrm{osc}^2(f, \Omega_z) := |\Omega_z| \|f - \textstyle\fint_{\Omega_z} f\,dx\|^2_{L^2(\Omega_z)} \text{ and}$$

$$\mathrm{osc}^2(f, \mathcal{N}) := \sum_{z \in \mathcal{N}(\Omega)} \mathrm{osc}^2(f, \Omega_z) \text{ with } \mathrm{osc}(f, \mathcal{N}) := \sqrt{\mathrm{osc}^2(f, \mathcal{N})}.$$

Theorem 3.1. *There exists some constant $C_{\mathrm{rel}} \approx 1$ which depends on the initial triangulation $\mathcal{T}_0$ and the coefficient matrix A such that any right-hand side $f \in L^2(\Omega)$ and any $\mathcal{T} \in \mathbb{T}$ with exact solution $u \in V$ to (1.1) and discrete solution $U \in V(\mathcal{T})$ to (1.2) satisfy*

$$\|u - U\|^2 \le C_{\mathrm{rel}}(\eta^2 + \mathrm{osc}^2(f, \mathcal{N})).$$

Proof. Let $e := u - U$. It is proven in [CB02, Thm. 2.1] that there exists a quasi-interpolation $\mathcal{I}e \in V(\mathcal{T})$ (which is essentially that of [Car99]) with

$$\|e - \mathcal{I}e\| \lesssim \|e\| \quad \text{and} \quad \int_\Omega f(e - \mathcal{I}e)\,dx \lesssim \|e\|\,\mathrm{osc}(f, \mathcal{N}).$$

This, the discrete problem and the integration by parts together with the techniques of [Ver96] lead to

$$\|e\|^2 = \int_\Omega f(e - \mathcal{I}e)\,dx - \sum_{E \in \mathcal{E}(\Omega)} \int_E (e - \mathcal{I}e)[A\nabla U]_E \cdot \nu_E\,ds$$

$$\lesssim \big(\mathrm{osc}(f, \mathcal{N}) + \eta\big)\|e\|.$$

This concludes the proof. $\qquad\qquad\square$

The further analysis of the discrete efficiency employs the following lemma on the oscillations.

Lemma 3.1 (oscillations). *Suppose $\mathcal{T}$ is a regular triangulation of the bounded Lipschitz domain Ω' into J triangles with $C_{reg} := \max_{T \in \mathcal{T}} |\Omega'|/|T|$. Assume that any triangle $T \in \mathcal{T}$ contains at least one interior vertex $\mathcal{N}(T) \cap \mathcal{N}(\Omega') \ne \emptyset$. Then any function $f \in L^2(\Omega')$ satisfies*

$$|\Omega'| \|f - \textstyle\fint_{\Omega'} f\,dx\|^2_{L^2(\Omega')} \le \frac{8 + J^3 C_{reg}^2}{4} \sum_{E \in \mathcal{E}(\Omega')} |\omega_E| \|f - \textstyle\fint_{\omega_E} f\,dx\|^2_{L^2(\omega_E)}$$

$$\le \frac{8 + J^3 C_{reg}^2}{4} \sum_{z \in \mathcal{N}(\Omega')} |\Omega'_z| \|f - \textstyle\fint_{\Omega'_z} f\,dx\|^2_{L^2(\Omega'_z)}.$$

Proof. First consider the special case of a piecewise constant function $f \in P_0(\mathcal{T})$. Let $\mathcal{T} = \{T_1, \ldots, T_J\}$ and denote $f_j := f|_{T_j}$, $\lambda_j := |T_j|/|\Omega'|$. Then

$$\bar{f} := \textstyle\fint_{\Omega'} f\,dx = \sum_{j=1}^J \lambda_j f_j \quad \text{and} \quad \sum_{j=1}^J \lambda_j = 1.$$

It follows that

$$\|f - \bar{f}\|_{L^2(\Omega')}^2 = \sum_{j=1}^{J}(f_j - \bar{f})^2 |T_j|$$

and

$$f_j - \bar{f} = (1 - \lambda_j)f_j - \sum_{\substack{k=1 \\ k \neq j}}^{J} \lambda_k f_k = \sum_{\substack{k=1 \\ k \neq j}}^{J} \lambda_k (f_j - f_k).$$

Hence,

$$\|f - \bar{f}\|_{L^2(\Omega')}^2 = |\Omega'| \sum_{j=1}^{J} \lambda_j \left(\sum_{\substack{k=1 \\ k \neq j}}^{J} \lambda_k (f_j - f_k) \right)^2$$

The Cauchy inequality in $\mathbb{R}^{J-1}$ implies for any $j \in \{1, \dots, J\}$ that

$$\left(\sum_{\substack{k=1 \\ k \neq j}}^{J} \lambda_k (f_j - f_k) \right)^2 \leq \left(\sum_{\substack{k=1 \\ k \neq j}}^{J} \lambda_k \right) \left(\sum_{\substack{k=1 \\ k \neq j}}^{J} \lambda_k (f_j - f_k)^2 \right)$$

$$\leq (1 - \lambda_j) \sum_{\substack{k=1 \\ k \neq j}}^{J} (f_j - f_k)^2.$$

The combination of the previous two displayed inequalities results in

$$\|f - \bar{f}\|_{L^2(\Omega')}^2 \leq |\Omega'| \sum_{j=1}^{J} \sum_{\substack{k=1 \\ k \neq j}}^{J} \lambda_j (1 - \lambda_j)(f_j - f_k)^2$$

(3.1)

$$\leq |\Omega'| \sum_{j=1}^{J} \sum_{\substack{k=1 \\ k \neq j}}^{J} (f_j - f_k)^2 \Big/ 4.$$

On the other hand, for any $E \in \mathcal{E}(\Omega')$ with $\overline{\omega}_E = T_+ \cup T_-$, $f_+ := f|_{T_+}$, $f_- := f|_{T_-}$, and $f_E := \fint_{\omega_E} f \, dx$ it holds that

$$\|f - f_E\|_{L^2(\omega_E)}^2 = \left(|f_+ - f_E|^2 |T_+| + |f_- - f_E|^2 |T_-| \right).$$

Elementary algebraic manipulations with $|T_+| + |T_-| = |\omega_E|$ prove

$$f_+ - f_E = \frac{|T_-|}{|\omega_E|}(f_+ - f_-) \quad \text{and} \quad f_- - f_E = \frac{|T_+|}{|\omega_E|}(f_- - f_+).$$

Hence,

(3.2)
$$|\omega_E| \|f - f_E\|_{L^2(\omega_E)}^2 = |\omega_E|^{-1}|f_+ - f_-|^2 \left(|T_+||T_-|^2 + |T_+|^2|T_-| \right)$$
$$= |T_+||T_-||f_+ - f_-|^2.$$

Given any $j, k \in \{1, \dots, J\}$ with $j \neq k$, there exists some $2 \leq \ell \leq J$ with pairwise distinct triangles $T_j =: T_{(1)}, T_{(2)}, \dots, T_{(\ell-1)}, T_{(\ell)} := T_k$ and edges
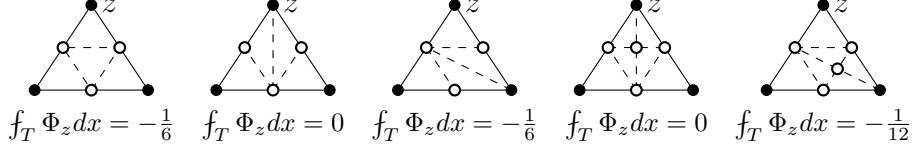
FIGURE 5. Sub-triangulations for a triangle $T \subseteq \omega_z$ in the proof of Theorem 3.2 with values of $\fint_T \Phi_z \, dx \le 0$.

$E_{(1)}, \dots, E_{(\ell-1)}$ with $T_{(m+1)} \cup T_{(m)} = \overline{\omega}_{E_{(m)}}$ for all $m = 1, \dots, \ell - 1$. Then the Cauchy inequality and (3.2) imply

$$|f_k - f_j|^2 = \left| \sum_{m=1}^{\ell-1} (f|_{T_{(m-1)}} - f|_{T_{(m)}}) \right|^2$$

$$\le |\Omega'|^{-2} \left(\sum_{m=1}^{\ell-1} \lambda_{(m)}^{-1} \lambda_{(m+1)}^{-1} \right) \left(\sum_{m=1}^{\ell-1} |T_{(m+1)}| |T_{(m)}| (f|_{T_{(m-1)}} - f|_{T_{(m)}})^2 \right)$$

$$\le |\Omega'|^{-2} (J-1) C_{reg}^2 \left(\sum_{m=1}^{\ell-1} |\omega_{E_{(m)}}| \| f - \fint_{\omega_{E_{(m)}}} f \, dx \|_{\omega_{E_{(m)}}}^2 \right).$$

This and (3.1) prove the assertion with constant $(J-1)^2 J C_{reg}^2 / 4$.

In the case of a general function $f \in L^2(\Omega')$, let $\bar{f} := \fint_{\Omega'} f \, dx$ and $\bar{f}_0 := \fint_{\Omega'} \Pi_0 f \, dx$, $\Pi_0 f \in P_0(\mathcal{T})$. Orthogonality leads to

$$\| f - \bar{f} \|_{L^2(\Omega')}^2 = \| f - \Pi_0 f \|_{L^2(\Omega')}^2 + \| \Pi_0 f - \bar{f}_0 \|_{L^2(\Omega')}^2 + \| \bar{f} - \bar{f}_0 \|_{L^2(\Omega')}^2.$$

For the last term on the right hand side, Hölder's inequality shows

$$\| \bar{f} - \bar{f}_0 \|_{L^2(\Omega')}^2 \le \| f - \Pi_0 f \|_{L^2(\Omega')}^2.$$

This together with

$$|\Omega'| \| f - \Pi_0 f \|_{L^2(\Omega')}^2 \le \sum_{E \in \mathcal{E}(\Omega')} |\omega_E| \| f - \fint_{\omega_E} f \, dx \|_{L^2(\omega_E)}^2$$

and the above estimate for piecewise constant functions proves the assertion with constant $(8 + (J-1)^2 J C_{reg}^2)/4$. The proof of the second stated inequality is immediate. $\qquad\square$

A key ingredient for the proof of the strong saturation is some fine-grid function Φ_z.

Definition 3.2. *Let $\varphi_z \in V(\mathcal{T})$ denote the nodal basis function associated with the node $z \in \mathcal{N}(\Omega)$. Define the set of edges that contain z by $\mathcal{E}(z) := \{ E \in \mathcal{E}(\Omega) \mid z \in E \}$. Let $\psi_E := \varphi_{\mathrm{mid}(E)}$ be the linear shape function of the refined triangulation $\hat{\mathcal{T}} \in \mathrm{unif}(\mathcal{T})$ associated with the midpoint of the edge $E \in \mathcal{E}(\Omega)$, and introduce*

$$\Phi_z := \varphi_z - \sum_{E \in \mathcal{E}(z)} \psi_E \in H_0^1(\omega_z) \subseteq V(\hat{\mathcal{T}}).$$

Theorem 3.2. *For any compactly interior edge* $E = \text{conv}\{a, b\} \in \mathcal{E}_c(\Omega)$ *with the vertices* $a, b \in \mathcal{N}(\Omega)$, *at least one vertex* $z \in \mathcal{N}(E) = \{a, b\}$ *satisfies*

$$\fint_{\omega_z} \Phi_z \, dx \approx 1 \approx \|\!|\Phi_z|\!\| \qquad and \qquad \fint_F \Phi_z \, ds = 0 \;\; for \; all \;\; F \in \mathcal{E}(\Omega).$$

Proof. The proof employs the technique of [CGMM14, Theorem 3.1]. The second assertion, namely $\fint_F \Phi_z \, ds = 0$ for all $F \in \mathcal{E}(\Omega)$, follows directly from the definition of Φ_z for any $z \in \mathcal{N}(E)$. For the proof of the first assertion, all possible configurations together with the values of $\fint_T \Phi_z \, dx$ for some $T \in \mathcal{T}(z) := \{K \in \mathcal{T} \mid z \in K\}$ are depicted in Figure 5. All values of $\int_T \Phi_z \, dx$ are nonpositive and so $\fint_{\omega_z} \Phi_z \, dx \approx 1$ or $\fint_{\omega_z} \Phi_z \, dx = 0$. The exceptional situation $\fint_{\omega_z} \Phi_z \, dx = 0$ implies the refinement pattern with *bisec3* and refinement edges opposite of z in all triangles $T \in \mathcal{T}(z)$. This pattern is possible for at most one vertex a or b. In other words $\fint_{\omega_a} \Phi_a \, dx = 0$ implies $\fint_{\omega_b} \Phi_b \, dx \neq 0$. This proves the assertion for at least one $z \in \{a, b\}$. The scaling $\|\!|\Phi_z|\!\| \approx 1$ follows from the shape-regularity and an inverse estimate. $\qquad \square$

Some hard analysis in the remaining parts of this section proves (SA) for a large class of triangulations $\mathbb{T}_H$.

Definition 3.3. $\mathbb{T}_H := \{\mathcal{T} \in \mathbb{T} \mid \mathcal{E}_c(\Omega) \neq \emptyset\}$

Theorem 3.3 (discrete efficiency). *There exists some constant* $C_{\text{deff}} \approx 1$ *which depends on the initial triangulation* $\mathcal{T}_0$ *and on the coefficient matrix* A *such that any* $\mathcal{T} \in \mathbb{T}_H$ *and* $\hat{\mathcal{T}} \in \text{unif}(\mathcal{T})$ *and any right-hand side* $f \in L^2(\Omega)$ *with discrete solutions* $U \in V(\mathcal{T})$ *and* $\hat{U} \in V(\mathcal{T})$ *to the Poisson model problem* $f + \text{div} \, A \nabla u = 0$ *in* Ω *satisfy*

$$(3.3) \qquad \eta^2 \leq C_{\text{deff}} \left(\|\!|\hat{U} - U|\!\|^2 + \text{osc}^2(f, \mathcal{N}) \right).$$

Proof. Note that, for any $E \in \mathcal{E}(\Omega) \backslash \mathcal{E}_a(\Omega)$, the error estimator contribution vanishes, $\eta^2(E) = 0$.

In the *first step* of the proof let $E = E_1 \in \mathcal{E}_a(\Omega)$. Theorem 2.1 implies that there exists a compactly interior edge $E' \in \mathcal{E}_c(\Omega)$ and a connected path of interior edges $E_1, \ldots, E_J$ with $J \leq C_1(\mathcal{T}_0) \approx 1$ such that $E \subseteq \overline{\omega}_{E_1}$ and $E' \subseteq \overline{\omega}_{E_J}$. Denote the endpoints of those edges by $P_1, \ldots, P_{J+1}$ such that $E_j = \text{conv}\{P_j, P_{j+1}\}$ for all $j = 1, \ldots J$ and note that the union of nodal patches

$$\Omega_E := \bigcup_{j=1}^{J+1} \omega_{P_j}$$

is a connected open set.

The *second step* consists in the proof of

$$(3.4) \qquad \eta(E) \lesssim \|A^{1/2} \nabla(\hat{U} - U)\|_{L^2(\Omega_E)} + \text{diam}(\Omega_E)\|f - f_{\Omega_E}\|_{L^2(\Omega_E)}.$$

Since $[A\nabla U]_E \cdot \nu_E$ is constant along the edge E of length $|E|$ with some sign $\pm$ as indicated below, it follows that

$$\pm\eta(E) = |E|[A\nabla U]_E \cdot \nu_E.$$

The edge-basis function ψ_E from Definition 3.2 satisfies $|E| = 2\int_E \psi_E\, ds$. Hence,

$$\pm\eta(E)/2 = \int_E \psi_E[A\nabla U]_E \cdot \nu_E\, ds.$$

Theorem 3.2 implies the existence of some node $z \in \mathcal{N}(E')$ such that $\int_E \Phi_z ds = 0$ and $\fint_{\omega_z} \Phi_z\, dx \neq 0$. With $\alpha := \int_{\omega_E} \psi_E\, dx / \int_{\omega_z} \Phi_z\, dx \approx 1$ (from shape-regularity) this implies

$$\pm\eta(E)/2 = \int_E (\psi_E + \alpha\Phi_z)\,[A\nabla U]_E \cdot \nu_E\, ds.$$

Note that the function $v_{\hat{\mathcal{T}}} := \psi_E + \alpha\Phi_z \in V(\hat{\mathcal{T}})$ satisfies $\int_F v_{\hat{\mathcal{T}}} ds = 0$ on all other edges $F \in \mathcal{E}\backslash\{E\}$. Therefore, the piecewise Gauss divergence theorem leads to

$$\pm\eta(E)/2 = \int_{\Omega_E} \nabla v_{\hat{\mathcal{T}}} \cdot A\nabla U dx$$

(In fact all the edge contributions and all other volume contributions vanish.) Consider the split

$$(3.5) \qquad \pm\eta(E)/2 = \int_{\Omega_E} \nabla v_{\hat{\mathcal{T}}} \cdot A\nabla(U - \hat{U})dx + \int_{\Omega_E} \nabla v_{\hat{\mathcal{T}}} \cdot A\nabla\hat{U}dx.$$

Recall $\|\Phi_z\| \approx 1$ from Theorem 3.2 and compute $\|\psi_E\| \approx 1$ to see that $|\alpha| \approx 1$ proves $\|v_{\hat{\mathcal{T}}}\| \lesssim 1$. This and the Cauchy-Schwarz inequality imply

$$(3.6) \qquad \int_{\Omega_E} \nabla v_{\hat{\mathcal{T}}} \cdot A\nabla(U - \hat{U})dx \lesssim \|A^{1/2}\nabla(U - \hat{U})\|_{L^2(\Omega_E)}$$

for the first term on the right-hand side of (3.5). Since $v_{\hat{\mathcal{T}}}$ is supported on $\overline{\Omega}_E$, the second term in (3.5) reads

$$\int_{\Omega_E} \nabla v_{\hat{\mathcal{T}}} \cdot A\nabla\hat{U}dx = \int_\Omega \nabla v_{\hat{\mathcal{T}}} \cdot A\nabla\hat{U}dx = \int_\Omega f v_{\hat{\mathcal{T}}}\, dx.$$

The choice of α shows $\fint_{\Omega_E} v_{\hat{\mathcal{T}}} dx = 0$. Hence $f_{\Omega_E} = \fint_{\Omega_E} f dx$ satisfies

$$\int_{\Omega_E} \nabla v_{\hat{\mathcal{T}}} \cdot A\nabla\hat{U}dx = \int_{\Omega_E} (f - f_{\Omega_E}) v_{\hat{\mathcal{T}}}\, dx.$$

The Friedrichs inequality and the aforementioned bounds show

$$\|v_{\hat{\mathcal{T}}}\|_{L^2(\Omega_E)} \lesssim \mathrm{diam}(\Omega_E)\|A^{1/2}\nabla v_{\hat{\mathcal{T}}}\|_{L^2(\Omega_E)} \lesssim \mathrm{diam}(\Omega_E).$$

The previous two displayed formulas and the Cauchy-Schwarz inequality imply

$$(3.7) \qquad \int_{\Omega_E} \nabla v_{\hat{\mathcal{T}}} \cdot A\nabla\hat{U}dx \lesssim \mathrm{diam}(\Omega_E)\|f - f_{\Omega_E}\|_{L^2(\Omega_E)}.$$

The combination of (3.5)–(3.7) conclude the proof of (3.4).

Step three is the conclusion of the proof of Theorem 3.3. Since $J \lesssim 1$ is uniformly bounded, Lemma 3.1 (applied to Ω_E) and (3.4) result in

$$(3.8) \qquad \eta^2(E) \lesssim \|A^{1/2}\nabla(U - \hat{U})\|_{L^2(\Omega_E)}^2 + \sum_{z\in\mathcal{N}(\Omega_E)} \mathrm{osc}^2(f, \Omega_z).$$

The design of Ω_E as the union of a finite number $\leq C(\mathcal{T}_0)$ of nodal patches and the shape-regularity of $\mathcal{T} \in \mathbb{T}$ imply the finite overlap

$$\max_{x \in \Omega} \left| \{ E \in \mathcal{E}(\Omega) \mid x \in \Omega_E \} \right| \lesssim 1$$

independent of $\mathcal{T} \in \mathbb{T}_H$, which concludes the assertion (3.3). $\qquad\square$

Theorem 3.4 (saturation property). *There exists some constant $C_H \in (1, \infty)$ which only depends on the interior angles of $\mathcal{T}_0$ and the coefficient matrix A such that for all $\mathcal{T} \in \mathbb{T}_H$ and for all $f \in L^2(\Omega)$ with exact solution $u \in V$ and discrete solutions $U, \hat{U}$ with respect to $\mathcal{T}$ and $\hat{\mathcal{T}} \in \mathrm{unif}(\mathcal{T})$, any $0 < \varepsilon \leq 1$ satisfies*

$$(3.9) \qquad \|u - \hat{U}\|^2 \leq (1 - \varepsilon/C_H(\mathcal{T}_0)) \|u - U\|^2 + \varepsilon \operatorname{osc}^2(f, \mathcal{N}).$$

Proof. Theorem 3.3 and the reliability of η imply that

$$\|u - U\|^2 \leq C_{\mathrm{rel}}(\eta^2 + \operatorname{osc}^2(f, \mathcal{N})) \leq C_{\mathrm{rel}}(C_{\mathrm{deff}} + 1)(\|\hat{U} - U\|^2 + \operatorname{osc}^2(f, \mathcal{N})).$$

This and the Galerkin-orthogonality

$$\|\hat{U} - U\|^2 = \|u - U\|^2 - \|u - \hat{U}\|^2$$

guarantee that $C_H(\mathcal{T}_0) := C_{\mathrm{rel}}(C_{\mathrm{deff}} + 1)$ satisfies

$$\|u - U\|^2 / C_H(\mathcal{T}_0) \leq \|u - U\|^2 - \|u - \hat{U}\|^2 + \operatorname{osc}^2(f, \mathcal{N}).$$

This proves the assertion for $\varepsilon = 1$. Multiply this inequality with $0 < \varepsilon \leq 1$ and add to the inequality $\|u - \hat{U}\|^2 \leq \|u - U\|^2$ (from Galerkin orthogonality) times $(1 - \varepsilon)$ to obtain

$$\|u - \hat{U}\|^2 \leq (1 - \varepsilon/C_H(\mathcal{T}_0)) \|u - U\|^2 + \varepsilon \operatorname{osc}^2(f, \mathcal{N}). \qquad\square$$

Remark 3.1 (adaptive mesh-refinement). Theorem 3.4 can be applied to adaptive mesh refinement as well, c.f. [CGMM14] for the Laplace eigenvalue problem. The adaptive refinement is based on the bulk criterion [Dör96] on nodal patches. For chosen bulk parameter $0 < \theta \leq 1$, let $\mathcal{M} \subset \mathcal{N}(\Omega)$ be the minimal subset, such that

$$\theta \eta^2 \leq \sum_{z \in \mathcal{M}} \eta^2(\mathcal{E}(z)).$$

Once a node is selected for refinement, all edges $\mathcal{E}(z)$ are refined by the *red-green-blue* refinement algorithm resulting in a locally refined $\hat{\mathcal{T}}$. Hence, once a triangle is refined at least one interior edge is bisected and therefore $\hat{\mathcal{T}} \in \mathbb{T}$. Note that all nodes in the set of marked nodes $\mathcal{M} \subseteq \mathcal{N}(\Omega)$ satisfy Theorem 3.2. Hence, the arguments of Theorem 3.3 apply to all edges $\mathcal{E}(z)$ for $z \in \mathcal{M}$ and so yield the discrete efficiency

$$\eta^2 \leq \theta^{-1} \sum_{z \in \mathcal{M}} \eta^2(\mathcal{E}(z)) \leq C_{\mathrm{deff}} \left(\|\hat{U} - U\|^2 + \operatorname{osc}^2(f, \mathcal{N}) \right).$$

This proves the saturation property (3.9) for adaptively refined meshes.

4. Weak Saturation implies Strong Saturation

The proof that weak saturation implies strong saturation (Main Result I) involves arguments from hard and soft analysis for the sets of triangulations

$$\mathbb{T}_H = \{\mathcal{T} \in \mathbb{T} \mid \mathcal{E}_c(\Omega) \neq \emptyset\} \quad \text{and} \quad \mathbb{T}_S := \mathbb{T} \setminus \mathbb{T}_H.$$

In view of Theorem 3.4, the assertion has to be verified only for $\mathbb{T}_S$. Note that the set $\mathbb{T}_S$ defines a finite set of finite element spaces

$$|\{V(\mathcal{T}) \mid \mathcal{T} \in \mathbb{T}_S\}| < \infty$$

while $\mathbb{T}_S$ may be infinite (cf. Example 2.2).

Theorem 4.1. *Suppose that $\mathcal{T} \in \mathbb{T}_S$ and $\hat{\mathcal{T}} \in \mathrm{unif}(\mathcal{T})$ satisfy WS. Then there exists some constant $C_S(V(\mathcal{T}))$ such that, for all $f \in L^2(\Omega)$, the exact solution u and the discrete solutions U and $\hat{U}$ satisfy*

$$\|u - \hat{U}\|^2 \leq (1 - \varepsilon/C_S(V(\mathcal{T}))) \, \|u - U\|^2 + \varepsilon \operatorname{osc}^2(f, \mathcal{N}).$$

The proof is based on a compactness argument.

Lemma 4.1. *Given $\mathcal{T} \in \mathbb{T}_S$ with (WS) there exists some $0 < C_S(V(\mathcal{T})) < \infty$ such that any $f_1 \in P_1(\mathit{bisec5}(\mathcal{T}))$ and exact (resp. discrete) solutions u (resp. U and $\hat{U}$ with respect to $\mathcal{T}$ and $\hat{\mathcal{T}} \in \mathrm{unif}(T)$) satisfy*

$$\|u - U\|^2 \leq C_S(V(\mathcal{T}))\big(\|\hat{U} - U\|^2 + \operatorname{osc}^2(f_1, \mathcal{N})\big).$$

Proof. For any $\mathcal{T} \in \mathbb{T}_S$, Theorem 3.1 implies

$$\|u - U\|^2 \leq C_{\mathrm{rel}} \left(\sum_{E \in \mathcal{E}(\Omega)} \eta^2(E) + \operatorname{osc}^2(f, \mathcal{N}) \right).$$

Let $\hat{\mathcal{T}} \in \mathrm{unif}(\mathcal{T})$ and define the following semi-norms

$$\vartheta_1(f_1) := \sqrt{\sum_{E \in \mathcal{E}(\Omega)} \eta^2(E)} \quad \text{and} \quad \vartheta_2(f_1) := \sqrt{\|\hat{U} - U\|^2 + \operatorname{osc}^2(f_1, \mathcal{N})}$$

for all f_1 in the space $P_1(\mathit{bisec5}(\mathcal{T}))$ and exact (resp. discrete) solutions u (resp. U and $\hat{U}$). If $\vartheta_2(f_1) = 0$, then $\operatorname{osc}(f_1, \mathcal{N}) = 0$ implies that f_1 equals a global constant. This and the weak saturation imply $f_1 \equiv 0$. Hence, equivalence of semi-norms $\vartheta_1 \lesssim \vartheta_2$ and the reliability lead to a constant $C_1(V(\mathcal{T}), V(\hat{\mathcal{T}}))$ such that

$$\|u - U\|^2 \leq C_1(V(\mathcal{T}), V(\hat{\mathcal{T}}))(\|\hat{U} - U\|^2 + \operatorname{osc}^2(f_1, \mathcal{N})).$$

It is correct that there is more than one realisation of $\hat{\mathcal{T}} \in \mathrm{unif}(\mathcal{T})$, but each of these leads to some constant $C_1(V(\mathcal{T})), V(\hat{\mathcal{T}}))$. This proves the lemma with $C_S(V(\mathcal{T})) = \max_{\hat{\mathcal{T}} \in \mathrm{unif}(\mathcal{T})} C_1(V(\mathcal{T}), V(\hat{\mathcal{T}}))$. $\square$

Proof of Theorem 4.1. Given $\mathcal{T} \in \mathbb{T}_S$ and $f \in L^2(\Omega)$, let $f_1 \in P_1(\mathit{bisec5}(\mathcal{T}))$ denote its L^2 projection onto piecewise affine (but possibly discontinuous) functions with respect to $\mathit{bisec5}(\mathcal{T})$. Denote the solution to problem (1.1) with right-hand side f_1 as $u(f_1)$ and note that U and $\hat{U}$ also solve (1.2) with right-hand side f_1. With $v := u - u(f_1)$, the triangle inequality reads

$$\|u - U\| \leq \|v\| + \|u(f_1) - U\|.$$

Define the first-order oscillations of f by

$$\operatorname{osc}_1(f, \mathcal{T}) = \sqrt{\sum_{T \in \mathcal{T}} |T| \, \|f - \Pi_1 f\|_{L^2(T)}^2}.$$

A piecewise Poincaré inequality shows for the constant C_P that

$$\|v\|^2 = \int_\Omega (f - \Pi_1 f)(v - \Pi_1 v) \, dx \le C_P \operatorname{osc}_1(f, \mathcal{T}) \|v\|.$$

Since $\operatorname{osc}_1(f, \mathcal{T}) + \operatorname{osc}(f_1, \mathcal{N}) \lesssim \operatorname{osc}(f, \mathcal{N})$, Lemma 4.1 proves leads to some constant $C_S(V(\mathcal{T}))$ with

$$\|u - U\|^2 \le C_S(V(\mathcal{T}))(\|\hat{U} - U\|^2 + \operatorname{osc}^2(f, \mathcal{N})).$$

This and the Galerkin orthogonality $\|\hat{U} - U\|^2 = \|u - U\|^2 - \|u - \hat{U}\|^2$ prove, for any $0 < \varepsilon \le 1$, that

$$\|u - \hat{U}\|^2 \le \frac{C_S(V(\mathcal{T})) - \varepsilon}{C_S(V(\mathcal{T}))} \|u - U\|^2 + \varepsilon \operatorname{osc}^2(f, \mathcal{N}). \qquad \square$$

Proof of Main Result I. Although the set $\mathbb{T}_S$ may be infinite as indicated in Example 2.2, the set $\{V(\mathcal{T}) \mid \mathcal{T} \in \mathbb{T}_S\}$ is finite, whence

$$\max_{\mathcal{T} \in \mathbb{T}_S} C_S(V(\mathcal{T})) < \infty$$

and Theorem 4.1 implies for any $0 < \varepsilon \le 1$ that

$$\|u - \hat{U}\|^2 \le \left(1 - \varepsilon / \left(\max_{\mathcal{T} \in \mathbb{T}_S} C_S(V(\mathcal{T}))\right)\right) \|u - U\|^2 + \varepsilon \operatorname{osc}^2(f, \mathcal{N}).$$

Thus, Theorem 3.4 proves, for any $\mathcal{T} \in \mathbb{T}$ and

$$C_2(\mathcal{T}_0) := \max \left\{ C_H(\mathcal{T}_0), \max_{\mathcal{T} \in \mathbb{T}_S} C_S(V(\mathcal{T})) \right\}$$

the strong saturation (SA) with $\rho(\varepsilon) := 1 - \varepsilon / C_2(\mathcal{T}_0)$. $\qquad \square$

Remark 4.1. The constant $C_H(\mathcal{T}_0)$ depends only on the smallest angle γ in $\mathbb{T}$, while the constant $C_S(V(\mathcal{T}))$ for $\mathcal{T} \in \mathbb{T}_S$ depends on $V(\mathcal{T})$ and so implicitly on $\mathcal{T}_0$. Since the entries in the global stiffness matrix depend on the geometric data in a continuous way, small perturbations in the positions of the vertices of $\mathcal{T}$ will preserve the weak saturation property.

5. A Characterisation of Domains with the Weak Saturation Property

This section is devoted to the proof of the Main Result II based on the subsequent two lemmas.

Lemma 5.1. *Let $\mathcal{T}$ be a regular triangulation of a bounded polygonal Lipschitz domain Ω with exactly one interior vertex z. Then the solution U to (1.2) with $f \equiv 1$ satisfies $U(z) > 0$.*

Proof. Let φ_z denote the local basis function associated with z. Since there is only one degree of freedom, the solution to (1.2) reads $U = U(z)\varphi_z$. The one-dimensional discrete linear system of equations reads

$$U(z) \int_\Omega (A \nabla \varphi_z) \cdot \nabla \varphi_z \, dx = \int_\Omega f \varphi_z \, dx.$$

Since $f > 0$ is constant and A is positive definite, $U(z)$ is positive. $\qquad\square$

Lemma 5.2. *Suppose that $\mathcal{E}_c(\Omega) = \emptyset$ and there exists an interior edge $E \in \mathcal{E}_b(\Omega)$ (with both end points on the boundary $\partial\Omega$). Then either* red *or* bisec3 *uniform refinement leads to weak saturation* (WS).

Proof. Consider an edge $E \in \mathcal{E}_b(\Omega)$ with end points $z_1, z_2 \in \mathcal{N}$ on the boundary $\partial\Omega$ that is refined. Since E is an interior edge, there are two adjacent triangles $T_+ = \mathrm{conv}\{z_+, z_1, z_2\}$ and $T_- = \mathrm{conv}\{z_-, z_1, z_2\}$.

In the case that $z_+, z_- \in \mathcal{N}(\partial\Omega)$, it holds $U|_{\omega_E} \equiv 0$. Lemma 5.1 shows that $U \neq \hat{U}$.

In the case that some $z \in \{z_+, z_-\} \cap \mathcal{N}(\Omega)$ is an interior vertex, then all edges $E = \mathrm{conv}\{y, z\}$ with one vertex z have the second end-point $y \in \mathcal{N}(\partial\Omega)$ on the boundary, because $\mathcal{E}_c(\Omega) = \emptyset$. Therefore the local problem on ω_z with only one degree of freedom decouples from the global system and Lemma 5.1 implies $U(z) > 0$. Therefore $U|_{\omega_E} \geq 0$. Since U is zero at the endpoints z_1, z_2, U vanishes along E and since U is non-negative in Ω it follows $[\nabla U]_E \cdot \nu_E \leq 0$ along E. Note that the jump $[\nabla U]_E$ is a multiple of the normal ν_E. The matrix A is positive definite and thus $[A\nabla U]_E \cdot \nu_E \leq 0$. Let φ_E denote the hat-function of the refined triangulation associated with the midpoint of E. Suppose for contradiction that $U \equiv \hat{U}$. Then

$$0 < \int_\Omega f\varphi_E \, dx = \int_{\omega_E} (A\nabla U) \cdot \nabla\varphi_E \, dx = |E|([A\nabla U]_E \cdot \nu_E)/2.$$

This implies $f \not\equiv 1$ which is a contradiction. $\qquad\square$

Proof of Main Result II. Theorem 3.3 proves the strong saturation (SA) for any $\mathcal{T} \in \mathbb{T}_H$, that is if there exists a compactly interior edge $\mathcal{E}_c(\Omega) \neq \emptyset$. In case that any interior edge has one endpoint on the boundary $\partial\Omega$ and there exists one interior edge $E \in \mathcal{E}(\Omega)$ with both endpoints on the boundary $\mathcal{N}(E) \subseteq \mathcal{N}(\partial\Omega)$, Lemma 5.2 proves weak saturation (WS). The remaining configuration is that each interior edge has exactly one endpoint on the boundary, which implies that the domain Ω equals the nodal patch ω_z of the only interior node z. If at least one $T \in \mathcal{T}$ is not refined using *bisec3* or if not all refinement edges in $\mathcal{T}$ are opposite to z, the discrete test function Φ_z from Definition 3.2 satisfies

$$\fint_{\omega_z} \Phi_z \, dx \approx 1 \qquad \text{and} \qquad \fint_F \Phi_z \, ds = 0 \text{ for all } F \in \mathcal{E}(\Omega).$$

Hence, the discrete efficiency technique of Theorem 3.3 leads to strong saturation. The only remaining case is that all triangles $T \in \mathcal{T}$ are refined by *bisec3* with refinement edges opposite to z. $\qquad\square$

The Main Result II reduces possible counterexamples of (WS) to configurations with 1 degree of freedom and *bisec3* refinement with all refinement edges opposite to the one interior node. The following result illustrates that under certain angle conditions even in this situation (WS) is valid.

Theorem 5.1. *Suppose that there exists exactly one interior vertex which is contained by all interior edges and that there exists an interior edge $E \in \mathcal{E}(\Omega)$ with one interior end point $z \in \mathcal{N}(\Omega)$ and one end point $P_j \in \mathcal{N}(\partial\Omega)$*
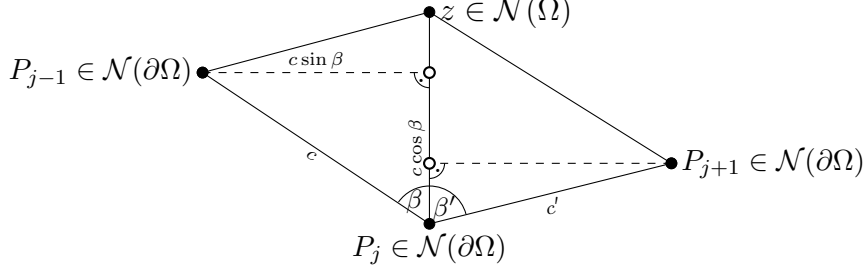
FIGURE 6. Angles in an edge patch.

on the boundary $\partial\Omega$ such that the angles depicted in Figure 6 satisfy $\cot\beta + \cot\beta' \leq 0$. Then (WS) holds.

Proof. Suppose that $U \equiv \hat{U}$. Then the local basis function φ_E that is 1 at $\text{mid}(E)$ and 0 at all other nodes of the refined triangulation satisfies

$$(5.1) \qquad 0 < \int_\Omega f\varphi_E \, dx = \int_{\omega_E} (A\nabla U) \cdot \nabla\varphi_E \, dx = |E|([A\nabla U]_E \cdot \nu_E)/2.$$

Note that all nodes of $\omega_E \cap \mathcal{N}$ except z are boundary nodes. Hence,

$$|E|([A\nabla U]_E \cdot \nu_E)/2 = U(z)|E|([A\nabla\varphi_z]_E \cdot \nu_E)/2.$$

The value of $\nabla\varphi_z \cdot \nu_E$ with respect to the angles of the two triangles as depicted in Figure 6 reads

$$|E|([\nabla\varphi_z]_E \cdot \nu_E) = \cot\beta + \cot\beta'.$$

Since the jump $[\nabla\varphi_z]_E$ is a multiple of the normal ν_E and the matrix A is positive definite, the case $\cot\beta + \cot\beta' \leq 0$ leads to a contradiction of (5.1) to $f \equiv 1$. □

Remark 5.1. Theorem 5.2 shows error reduction for triangulations with non-convex corners. Note that the criss-cross counter-example for error reduction does not fulfill the condition of Theorem 5.1 for error reduction.

Remark 5.2. In particular, Main Result I & II imply that the following saturation test can be employed to decide whether the strong saturation property is valid.

compute discrete solutions U and $\hat{U}$ with respect to $\mathcal{T}$ and $\hat{\mathcal{T}} \in \text{unif}(\mathcal{T})$ and $f \equiv 1$
if $U = \hat{U}$ **then** no saturation **else** (SA) for all $f \in L^2(\Omega)$ **end if**

This test is very simple because it is only performed for $f \equiv 1$ and has to be only performed for configurations with one degree of freedom.

6. Linear Second-Order Elliptic Problems

This section extends the results of the foregoing sections to elliptic linear second-order equations with constant coefficients, namely with a symmetric positive definite $A \in \mathbb{R}^{2\times 2}$, $b \in \mathbb{R}^2$, and $\gamma \in \mathbb{R}$. Given $f \in L^2(\Omega)$, the general second-order linear PDE assumes the form

$$\mathcal{L}u := -\operatorname{div}(A\nabla u) + b \cdot \nabla u + \gamma u = f.$$

Its weak formulation seeks $u \in V := H_0^1(\Omega)$ such that, for all $v \in V$,

$$(6.1) \qquad \mathcal{B}(u,v) := \int_\Omega ((A\nabla u) \cdot \nabla v + vb \cdot \nabla u + \gamma uv) \, dx = \int_\Omega fv \, dx.$$

Assume that the constant coefficients b and γ are such that the bilinear form $\mathcal{B}$ is V-elliptic, i.e., there exists some constant $c_{\text{ell}} \approx 1$ such that, for all $v \in V$,

$$(6.2) \qquad c_{\text{ell}}(\|v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2) \leq \mathcal{B}(v,v) =: \|v\|_{\mathcal{B}}^2.$$

(For constant coefficient b, $\int_\Omega vb \cdot \nabla v \, dx = \int_\Omega b \cdot \nabla |v|^2/2 \, dx = 0$. Hence, it suffices to consider $\gamma > -C_F(\Omega)$ for the Friedrichs constant $C_F(\Omega) \leq \text{diam}(\Omega)/\pi$.) Under the above ellipticity condition, $\|\cdot\|_{\mathcal{B}} := \mathcal{B}(\cdot, \cdot)^{1/2}$ defines the energy norm. The finite element method computes a unique $u_{\mathcal{T}} \in V(\mathcal{T})$ such that, for all $v_{\mathcal{T}} \in V(\mathcal{T})$,

$$(6.3) \qquad \mathcal{B}(u_{\mathcal{T}}, v_{\mathcal{T}}) = \int_\Omega fv_{\mathcal{T}} \, dx.$$

The following generalization of the Main Result I states that ellipticity plus small skew-symmetry implies saturation. The saturation test from Remark 5.2 and the notion of weak saturation still concern the reduced problem $-\text{div}(A\nabla u) = f$, whereas strong saturation for problem (6.1) states that for any $0 < \varepsilon \leq 1$ there exists $\varrho(\varepsilon) := 1 - \varepsilon/C(\mathcal{T}_0) < 1$ such that

$$(\text{SA'}) \qquad \|u - \hat{U}\|_{\mathcal{B}}^2 \leq \varrho(\varepsilon)\|u - U\|_{\mathcal{B}}^2 + \varepsilon \, \text{osc}(f - b \cdot \nabla U - \gamma U, \mathcal{N})$$

holds with a universal constant $C(\mathcal{T}_0)$ which exclusively depends on $\mathcal{T}_0$ and the coefficients A, b, γ.

Theorem 6.1 (weak saturation implies strong saturation)**.** *Assume $\mathcal{B}$ is elliptic with (6.2) and the convection parameter b satisfies $|b| < C_{\text{ell}}$. Then there exists a global constant $C(\mathcal{T}_0)$ which depends only on $\mathcal{T}_0$ such that for any $\mathcal{T} \in \mathbb{T}$ and $\hat{\mathcal{T}} \in \text{unif}(\mathcal{T})$, (WS) implies (SA') with $\varrho(\varepsilon) = 1 - \varepsilon/C(\mathcal{T}_0)$.*

The proof will be given throughout the remaining parts of this section.

Lemma 6.1. *For any $\mathcal{T} \in \mathbb{T}$ with $\hat{\mathcal{T}} \in \text{unif}(\mathcal{T})$, the exact and discrete solutions $u, U, \hat{U}$ to (6.1) and (6.3) with right-hand side $f \in L^2(\Omega)$ satisfy*

$$(6.4) \qquad \|u - U\|_{\mathcal{B}} \lesssim \eta + \text{osc}(f - b \cdot \nabla U - \gamma U, \mathcal{N}).$$

If, in addition $\mathcal{T} \in \mathbb{T}_H$, it holds that

$$\|u - U\|_{\mathcal{B}} \lesssim \|\hat{U} - U\|_{\mathcal{B}} + \text{osc}(f - b \cdot \nabla U - \gamma U, \mathcal{N}).$$

Proof. Note that U solves (1.2) with right-hand side $f - b \cdot \nabla U - \gamma U$ instead of f. The reliability of the error estimator from Theorem 3.1 translates directly into

$$\|u - U\|_{\mathcal{B}}^2 \lesssim \eta^2 + \text{osc}^2(f - b \cdot \nabla U - \gamma U, \mathcal{N})$$

for the oscillations $\text{osc}(\cdot, \mathcal{N})$ of Section 3. For any $\mathcal{T} \in \mathbb{T}_H$, the discrete efficiency techniques from Theorem 3.3 imply

$$\eta \lesssim \|\hat{U} - U\|_{\mathcal{B}} + \text{osc}(f - b \cdot \nabla \hat{U} - \gamma \hat{U}, \mathcal{N}).$$

The triangle inequality reveals

$$\operatorname{osc}(f - b \cdot \nabla \hat{U} - \gamma \hat{U}, \mathcal{N}) \leq \operatorname{osc}(f - b \cdot \nabla U - \gamma U, \mathcal{N})$$
$$+ \operatorname{osc}(b \cdot \nabla(U - \hat{U}) - \gamma(U - \hat{U}), \mathcal{N}).$$

The combination with the reliability (6.4) leads to the second stated estimate. $\square$

For the finite set $\{V(\mathcal{T}) \mid \mathcal{T} \in \mathbb{T}_S\}$, the following analogue of Lemma 4.1 follows from a compactness argument.

Lemma 6.2. *For any $\mathcal{T} \in \mathbb{T}_S$ with weak saturation (WS), there exists some constant $C_S(V(\mathcal{T}))$ such that any $f_1 \in P_1(bisec5(\mathcal{T}))$ with exact and discrete solutions $u, U, \hat{U}$ to (6.1) and (6.3) with right-hand side f_1 satisfies*

$$\|u - U\|_{\mathcal{B}} \leq C_S(V(\mathcal{T})) \left(\|\hat{U} - U\|_{\mathcal{B}} + \operatorname{osc}(f_1 - b \cdot \nabla \hat{U} - \gamma \hat{U}, \mathcal{N}) \right).$$

Proof. For all f_1 in the space $P_1(bisec5(\mathcal{T}))$, define the semi-norms

$$\vartheta_1(f_1) := \sqrt{\sum_{E \in \mathcal{E}(\Omega)} \eta^2(E)} \quad \text{and}$$

$$\vartheta_2(f_1) := \sqrt{\|\hat{U} - U\|_{\mathcal{B}}^2 + \operatorname{osc}^2(f_1 - b \cdot \nabla U - \gamma U, \mathcal{N})}.$$

If $\vartheta_2(f_1) = 0$, then $\hat{U} \equiv U$ and $\operatorname{osc}(f_1 - b \cdot \nabla U - \gamma U, \mathcal{N}) = 0$ imply that $f_1 - b \cdot \nabla U - \gamma U \in P_0(\Omega)$ is constant. Hence, $U = \hat{U}$ solves (1.2) with constant right-hand side and the weak saturation implies $U = \hat{U} \equiv 0$ and $f_1 \in P_0(\Omega)$ vanishes. Equivalence of semi-norms in the finite-dimensional space $P_1(bisec5(\mathcal{T}))$ and the reliability (6.4) lead to a constant $C_S(V(\mathcal{T}))$ such that $\vartheta_1 \leq C_S(V(\mathcal{T}))\vartheta_2$. This is the assertion. $\square$

Proof of Theorem 6.1. Equation (6.4) and Lemma 6.2 plus the arguments from the proof of Main Result I imply that there exists a constant $C \approx 1$ such that any $\mathcal{T}$ and $\hat{\mathcal{T}} \in \operatorname{unif}(\mathcal{T})$ with (WS) satisfy

$$\|u - U\|_{\mathcal{B}} \lesssim \|\hat{U} - U\|_{\mathcal{B}} + \operatorname{osc}(f - b \cdot \nabla U - \gamma U, \mathcal{N}).$$

Since the bilinear form $\mathcal{B}$ is not symmetric, it holds

$$\|\hat{U} - U\|_{\mathcal{B}}^2 = \|u - U\|_{\mathcal{B}}^2 - \|u - \hat{U}\|_{\mathcal{B}}^2 + 2 \int_{\Omega} (\hat{U} - U) b \cdot \nabla(u - \hat{U}) \, dx.$$

The Cauchy and Young inequalities imply

$$(6.5) \quad 2 \int_{\Omega} (\hat{U} - U) b \cdot \nabla(u - \hat{U}) \, dx \leq C_{\text{ell}}^{-1} |b| \left(\|u - \hat{U}\|_{\mathcal{B}}^2 + \|\hat{U} - U\|_{\mathcal{B}}^2 \right).$$

Provided $|b| < C_{\text{ell}}$, these terms can be absorbed. This results in

$$(6.6) \quad \|\hat{U} - U\|_{\mathcal{B}}^2 \lesssim \|u - U\|_{\mathcal{B}}^2 - \|u - \hat{U}\|_{\mathcal{B}}^2.$$

The techniques from the proof of Theorem 3.4 conclude the proof of Theorem 6.1. $\square$

Remark 6.1. For a general parameter b and $\gamma \in \mathbb{R}$ such that $\mathcal{L}$ is injective, the estimate (6.6) follows for sufficiently small mesh-size. The proof is a combination of (6.5) and the higher-order convergence of the error in the L^2 norm, which is also employed in [SW96]. This leads to saturation in an asymptotic regime.

References

[AO00]   Mark Ainsworth and J. Tinsley Oden, *A posteriori error estimation in finite element analysis*, Pure and Applied Mathematics (New York), Wiley-Interscience [John Wiley & Sons], New York, 2000.

[BC08]   Sören Bartels and Carsten Carstensen, *A convergent adaptive finite element method for an optimal design problem*, Numer. Math. **108** (2008), no. 3, 359–385.

[BS93]   Randolph E. Bank and R. Kent Smith, *A posteriori error estimates based on hierarchical bases*, SIAM J. Numer. Anal. **30** (1993), no. 4, 921–935.

[Car99]   Carsten Carstensen, *Quasi-interpolation and a posteriori error analysis in finite element methods*, M2AN Math. Model. Numer. Anal. **33** (1999), no. 6, 1187–1202.

[CB02]   Carsten Carstensen and Sören Bartels, *Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids. I. Low order conforming, nonconforming, and mixed FEM*, Math. Comp. **71** (2002), no. 239, 945–969.

[CGMM14] C. Carstensen, J. Gedicke, V. Mehrmann, and A. Miedlar, *An adaptive finite element method with asymptotic saturation for eigenvalue problems*, Numer. Math. **128** (2014), no. 4, 615–634.

[CKNS08]   J.M. Cascon, Ch. Kreuzer, R. H. Nochetto, and K. G. Siebert, *Quasi-optimal convergence rate for an adaptive finite element method*, SIAM J. Numer. Anal. **46** (2008), no. 5, 2524–2550.

[CV99]   Carsten Carstensen and Rüdiger Verfürth, *Edge residuals dominate a posteriori error estimates for low order finite element methods*, SIAM J. Numer. Anal. **36** (1999), no. 5, 1571–1587.

[DN02]   Willy Dörfler and Ricardo H. Nochetto, *Small data oscillation implies the saturation assumption*, Numer. Math. **91** (2002), no. 1, 1–12.

[Dör96]   W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal. **33** (1996), 1106–1124.

[FLOP10]   S. Ferraz-Leite, C. Ortner, and D. Praetorius, *Convergence of simple adaptive Galerkin schemes based on $h - h/2$ error estimators*, Numer. Math. **116** (2010), no. 2, 291–316.

[Rod94]   Rodolfo Rodríguez, *A posteriori error analysis in the finite element method*, Finite element methods (Jyväskylä, 1993), Lecture Notes in Pure and Appl. Math., vol. 164, Dekker, New York, 1994, pp. 389–397.

[Ste08]   Rob Stevenson, *The completion of locally refined simplicial partitions created by bisection*, Mathematics of Computation **77** (2008), no. 261, 227–241.

[SW96]   Alfred H. Schatz and Jun Ping Wang, *Some new error estimates for Ritz-Galerkin methods with minimal regularity assumptions*, Math. Comp. **65** (1996), no. 213, 19–27.

[Ver96]   R. Verfürth, *A review of a posteriori error estimation and adaptive mesh-refinement techniques*, Advances in numerical mathematics, Wiley, 1996.

(C. Carstensen) Institut für Mathematik, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

E-mail address: cc@mathematik.hu-berlin.de

(D. Gallistl) Institut für Numerische Simulation, Universität Bonn, Wegelerstrasse 6, 53115 Bonn, Germany

E-mail address: gallistl@ins.uni-bonn.de

(J. Gedicke) Interdisziplinäres Zentrum für Wissenschaftliches Rechnen (IWR), Universität Heidelberg, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany

E-mail address: joscha.gedicke@iwr.uni-heidelberg.de