Averaging Techniques for Reliable and Efficient A Posteriori Finite Element Error Control: Analysis and Applications

Carsten Carstensen and Max Jensen

Abstract. Local averaging techniques, which are used to postprocess discrete flux or stress approximations of low-order finite element schemes for elliptic boundary value problems, are applied for error control and adaptive mesh refinement. We put particular emphasis on the explicit calculation of all constants, arising in the proofs of reliability and efficiency, in terms of the known data and quantify the equivalence of local averaging techniques. We highlight and discuss a wide selection of applications for which averaging-based estimators provide highly accurate error control.

1. Introduction

In this section we present central concepts of a posteriori finite element error control. In Subsection 1.1 we give a brief introduction to reliability and efficiency of a posteriori error estimators. In Subsections 1.2 and 1.3 we consider error estimators as termination criteria for error control and in adaptive mesh-refining.

1.1. Reliability and Efficiency. To introduce error estimators we do not need to specify an underlying boundary value problem of the computation; in later sections, however, we will be concerned with differential equations

\[ Lu = f, \]

paired with appropriate boundary conditions. Here \( L \) is in general an elliptic differential operator of second order, acting on functions with the domain \( \Omega \).

In the abstract framework it suffices to consider a function \( p := p(\nabla u) \), which depends on the derivative of the unknown exact solution \( u \), as well as the discrete counterpart \( p_h := p(\nabla u_h) \), which depends on the known finite element solution \( u_h \). Typically, \( p \) has the physical interpretation of a flux or stress.

Given a norm \( \| \cdot \| \), the aim is to approximate the unknown error \( \| p - p_h \| \) by a computable quantity \( \eta \), called error estimator.

Definition 1.1. (Error Estimator). A quantity \( \eta \), which is thought of being an approximation to \( \| p - p_h \| \), is called a posteriori error estimator, or estimator for brevity, if it is a computable function of known quantities such as \( f, \Omega, \partial \Omega \) and \( u_h, p_h \).

The definition of error estimators is vague and not very useful on its own. The purpose of an estimator is to provide lower and upper error bounds. This is made precise by the terms reliability and efficiency. To introduce these two concepts, we consider a family of finite element meshes \( (T_h)_{h \in \mathcal{H}} \), where \( h \in \mathcal{H} \) is a parameter representative for the mesh-size in \( T_h \). For quasi-uniform triangulations we assume that the index set \( \mathcal{H} \) is a subset of \( (0, \infty) \) and \( h \) is the diameter of the largest element.

2000 Mathematics Subject Classification. 65N15, 65N30.

©0000 (copyright holder)
**Definition 1.2. (Reliability).** An estimator \( \eta \) is called **reliable** if there is a constant \( C_{\text{rel}} \) and a bound

\[
\|p - p_h\| \leq C_{\text{rel}} \eta + \text{h.o.t.}_{\text{rel}}
\]

such that the function \( \text{h.o.t.}_{\text{rel}} \) is \( o(\|p - p_h\|) \).

**Definition 1.3. (Efficiency).** An estimator \( \eta \) is called **efficient** if there is a constant \( C_{\text{eff}} \) and a bound

\[
\eta \leq C_{\text{eff}} \|p - p_h\| + \text{h.o.t.}_{\text{eff}}
\]

such that the function \( \text{h.o.t.}_{\text{eff}} \) is \( o(\|p - p_h\|) \).

We say a function \( \text{h.o.t.} \) is \( o(\|p - p_h\|) \) if, and only if, \( \lim_{H \ni h \to 0} \text{h.o.t.}/\|p - p_h\| = 0 \).

**Definition 1.4. (Asymptotic Exactness).** An estimator is called **asymptotically exact** if it is reliable and efficient.

We emphasize that \( C_{\text{rel}} \) and \( C_{\text{eff}} \) are multiplicative constants which do not depend on the mesh-size of the underlying finite element mesh \( T \). In the above definitions the abbreviation \( \text{h.o.t.} \) stands for ‘higher-order terms’. In practical computations these terms are typically much smaller than \( \eta \) and \( \|p - p_h\| \), but their size usually depends on the (unknown) smoothness of the exact solution or the (known) smoothness of the given data.

It is the mathematical task of a posteriori error analysis to provide sufficient and necessary conditions for the reliability and efficiency of error estimators as well as to characterise and estimate the constants \( C_{\text{rel}}, C_{\text{eff}} \) and the higher-order terms \( \text{h.o.t.}_{\text{rel}}, \text{h.o.t.}_{\text{eff}} \).

### 1.2. A Posteriori Error Control

There are at least two important areas where estimators are applied in practice: error estimation and adaptive mesh-refinement. For the first area, one is interested in a termination criterion for an algorithm of successively adapted mesh-refinements which guarantees that a given tolerance \( \text{Tol} > 0 \) is not exceeded:

\[
\|p - p_h\| \leq \text{Tol}.
\]

Since the term \( \|p - p_h\| \) is unknown, it is replaced by its upper bound (1.1) which leads to the criterion

\[
C_{\text{rel}} \eta + \text{h.o.t.}_{\text{rel}} \leq \text{Tol}.
\]

For (1.3) to hold, it is evident that we require not only quantitative knowledge of \( \eta \) but also of \( C_{\text{rel}} \) and of \( \text{h.o.t.}_{\text{rel}} \). Notice that one may call the computable upper bound

\[
\tilde{\eta} := C_{\text{rel}} \eta + \text{h.o.t.}_{\text{rel}}
\]

an error estimator. Clearly, \( \tilde{\eta} \) satisfies (1.1) with the reliability constant 1 and vanishing higher-order terms. If \( \tilde{\eta} \) is not computable, the error control is incomplete and therefore useless. Fortunately, in the examples below, \( \tilde{\eta} \) is computable and thus enables guaranteed error control.

Observe that the above error bound gives control only with respect to the norm \( \| \cdot \| \) and not on other target functionals. In this paper we focus on energy norms and so ignore goal-oriented error control. The latter is also important and the reader is referred to [1, 3], where arguments are detailed on how to reduce the more general problem to energy-norm control, and to [7] for a survey of the work of Rannacher et al. on a very successful computational approach.

### 1.3. Adaptive Mesh-Refining

Error estimators usually not only give a bound on the global error, but also contain information on the local approximation quality of \( p_h \) to \( p \). For instance, \( \eta \) can take the form

\[
\eta^2 = \sum_{T \in \mathcal{T}} \eta_T^2,
\]

where \( \eta_T \) are computable elementwise contributions to \( \eta \). The second area of application of error estimators relies on the observation. One then interprets \( \eta_T \) as a local *indicator* and refines the
element $T$ if the associated value $\eta_T$ is relatively large. For example, one can flag all elements $T$ for refinement which satisfy the inequality

$$1/2 \max \{\eta_K : K \in T\} \leq \eta_T.$$

Notice that further strategies may be required to avoid hanging nodes and degenerated elements. The use of $\eta_T$ as a refinement criterion is often based on heuristics; one should then speak of a refinement indicator $\eta_T$ and not of an error indicator.

Notice that $C_{\text{rel}}, C_{\text{eff}}$ do not enter in (1.5) while $\text{h.o.t. rel}$, $\text{h.o.t. eff}$ are simply ignored.

The rigorous justification of adaptive mesh-refining algorithms started with [19, 20, 22] by a proof of error-reduction properties. In these publications, even the rate of convergence of the numerical approximations to the exact solution is specified. However, the number of elements generated by the considered adaptive algorithms is not controlled, which may affect the performance of the method. In [8] coarsening steps are introduced to prove optimality properties in respect to the performance of the adaptive mesh-refinement. However, coarsening seems to be, for elliptic problems, theoretically motivated in first place and appears to be unnecessary in practice.

2. Preliminaries and Notation

Let $\Omega$ be a bounded Lipschitz domain in $\mathbb{R}^2$ with a piecewise affine boundary $\Gamma$. Assume that $\Omega$ is exactly covered by a mesh $T$, i.e. $\mathcal{U} = \Omega$. We only consider closed triangular elements $T \in T$.

The vertices $a,b,c \in \mathbb{R}^2$ of the element $T = \text{conv}\{a,b,c\}$ are called nodes; $\mathcal{N}$ denotes the set of all nodes. We let $\mathcal{N}_T := \Gamma \cap \mathcal{N}$ be the set of exterior and $\mathcal{N}_\Omega := \mathcal{N} \setminus \mathcal{N}_T$ be the set of interior nodes. We write $\mathcal{E}$ for the set of edges $E$; $\mathcal{E}_\Omega$ denotes the interior edges and $\mathcal{E}_\Gamma := \{E \in \mathcal{E} : E \subset \Gamma\} = \mathcal{E} \setminus \mathcal{E}_\Omega$ denotes the boundary edges. Intersecting distinct elements share either one vertex or an edge. Hanging nodes are excluded for the ease of the presentation. For each node $z \in \mathcal{N}$ let $\mathcal{E}_z := \{E \in \mathcal{E} : z \in E \cap \mathcal{N}\}$ and $\mathcal{T}_z := \{T \in \mathcal{T} : z \in T \cap \mathcal{N}\}$. To each edge $E$ a unit normal vector $\nu_E$ with fixed orientation is associated; if $E \subset \partial \Omega$, set $\nu_E = \nu$, the outer unit normal along $\partial \Omega$. The length of $E \in \mathcal{E}$ is denoted by $h_E = \text{diam}(E)$ and the diameter of $\mathcal{T}_z$, $z \in \mathcal{N}$, by $h_T$. Similarly, the diameter of an element $T$ is denoted by $h_T$. The area of $T \in \mathcal{T}$ is abbreviated by $|T| = \mathcal{L}^2(T)$. Here $\mathcal{L}^2(T)$ is the two-dimensional Lebesgue measure of $T$.

We let $P_k(T)$ be the set of algebraic polynomials with total degree less than or equal to $k$. Furthermore, $P_k(T, \mathbb{M})$ is the set of $\mathbb{M}$-valued elements, where $\mathbb{M}$ is a space of real scalars, vectors or matrices. We set

$$P_k(T, \mathbb{M}) := \{v_h \in L^\infty(\Omega) : \forall T \in \mathcal{T}, v_h|_T \in P_k(T, \mathbb{M})\},$$

$$P_k(T) := P_k(T, \mathbb{R}).$$

We let $P_h$ be a space which contains all possible values of the $p_h$ we introduced in the previous section. We make hereby the assumption that $P_h$ is a subspace of $P_1(T, \mathbb{M})$. Typically, $P_h = P_0(T, \mathbb{M})$ or $P_h = P_1(T, \mathbb{M})$. However, in principle, also other choices of $P_h$ are admissible; for instance, to take the effect of the boundary conditions into account.

We shall see in the subsequent sections that the size of the error $||p - p_h||$ can be estimated by comparing $p_h$ with functions in the space

$$Q_h := \{v_h \in P_1(T, \mathbb{M}) : v_h \text{ continuous}\}.$$

Given a subset $\omega$ of $\Omega$, we abbreviate

$$P_h|_\omega := \{p_h|_\omega : p_h \in P_h\} \quad \text{and} \quad Q_h|_\omega := \{q_h|_\omega : q_h \in Q_h\}.$$

The nodal basis functions $\varphi_z \in P_1(T)$, $z \in \mathcal{N}$, are defined by

$$\varphi_z(z') = \begin{cases} 1 : & z = z', \\ 0 : & z \neq z'. \end{cases} \quad \forall z' \in \mathcal{N}.$$

Consequently,

$$0 \leq \varphi_z \leq 1, \quad \text{supp} \varphi_z = \mathcal{T}_z, \quad \text{and} \quad \sum_{z \in \mathcal{N}} \varphi_z = 1.$$
2.1. Averaging Operators. We construct averaging operators $A : \mathcal{P}_h \to \mathcal{Q}_h$ from local linear operators $A_z : P_1(T_z; \mathcal{M}) \to \mathcal{M}$ with the formula
\begin{equation}
A(p_h) := \sum_{z \in N} A_z(p_h|_{T_z}) \varphi_z, \quad p_h \in \mathcal{P}_h.
\end{equation}

We assume that the local averaging operators $A_z$ are preserving, that is
\begin{equation}
A_z(f) = f(z)
\end{equation}
for all $f \in N(T_z) := (\mathcal{P}_h|_{T_z}) \cap (\mathcal{Q}_h|_{T_z})$.

**Example 2.1 (Averaging of Nodal Values).** Amongst the easiest local averaging techniques is the averaging of the function values of $p_h$ at a node $z \in N$; that is, $A_z$ is defined as weighted mean of all the different values $(v|_{T})(z)$ where $z \in T$:
\begin{equation}
A_z(v) := \sum_{T \in T_z} \lambda_{z,T} v|_{T}(z) \quad \text{for all } v \in P_1(T_z; \mathcal{M}), \; v \in N.
\end{equation}

Here the $\lambda_{z,T}$ are real coefficients for which condition (2.2) is equivalent to
\begin{equation}
\forall z \in N : \sum_{T \in T_z} \lambda_{z,T} = 1.
\end{equation}

**Example 2.2 (ZZ Averaging).** The special case
\begin{equation}
\lambda_{z,T} := \frac{|T|}{|T_z|} \quad \text{for all } T \in T_z, \; z \in N,
\end{equation}
where $| \cdot |$ denotes the area, is our interpretation of a gradient recovery. The case $\mathcal{P}_h = \mathcal{P}_0(T)$ is due to Zienkiewicz and Zhu [23]. The corresponding operator $Z := A$ with $A_z := Z_z$ reads
\begin{equation}
Z(p_h) = \sum_{z \in N} \left( \sum_{T \in T_z} \frac{|T|}{|T_z|} (p_h|_{T})(z) \right) \varphi_z.
\end{equation}

Since the choice (2.5) immediately implies (2.4), condition (2.2) is satisfied.

2.2. Estimators. We define, for any fixed $p_h \in \mathcal{P}_h$, the averaging estimators
\begin{equation}
\eta_M := \min_{r_h \in \mathcal{Q}_h} \| p_h - r_h \|_{L^2(\Omega; \mathcal{M})} \leq \eta_A := \| p_h - A(p_h) \|_{L^2(\Omega; \mathcal{M})}.
\end{equation}

For the ZZ averaging operator from Example 2.2, we introduce
\begin{equation}
\eta_Z := \| p_h - Z(p_h) \|_{L^2(\Omega; \mathcal{M})}.
\end{equation}

Finally, given any $p_h \in \mathcal{P}_h$ and $E \in \mathcal{E}$, let $[p_h]|_E$ denote the jump of $p_h$ across the edge $E$ with the $L^2$-norm $\| [p_h]|_E \|_{L^2(E)}$ along $E$; then,
\begin{equation}
\eta_E := \left( \sum_{E \in \mathcal{E}} h_E \| [p_h]|_E \|_{L^2(E)}^2 \right)^{1/2}.
\end{equation}

We introduce the convention $[p_h]|_E := 0$ for boundary edges $E \in \mathcal{N}_\Omega$ to unify the treatment of interior and exterior edges in some of the forthcoming proofs.

3. Efficiency

There is no need to specify a particular boundary value problem in order to analyze the efficiency of the above estimators. In this section we prove that averaging estimators are efficient for any problem which concerns the approximation of smooth functions by elements in $\mathcal{P}_h$.

We show first that $\eta_M$ is efficient in the sense of
\begin{equation}
\eta_M \leq \| p - p_h \| + \text{h.o.t.}
\end{equation}

The surprising fact is that $p$ can be any smooth function, e.g. $p \in H^1(\Omega; \mathbb{R}^n)$, and $\| \cdot \|$ any norm such that
\begin{equation}
\min_{q_h \in \mathcal{Q}_h} \| p - q_h \| = o(\| p - p_h \|).
\end{equation}
The proof of (3.1) only requires the triangle inequality and is therefore applicable to a wide range of applications: for any $q_h \in Q_h$ we have
\[ \eta_M \leq \|p_h - q_h\| \leq \|p - p_h\| + \|p - q_h\|. \]
Thus, efficiency (3.1) holds without any reference to the underlying boundary value problem. For instance, if $p_h$ denotes some flux approximation in a first-order conforming or nonconforming or lowest-order mixed FEM, then the $L^2$-error $\|p - p_h\|$ is of first order, while (3.2) is of second order provided the exact flux $p$ is sufficiently smooth.

We now show that $\eta_M$ and $\eta_A$ are equivalent up to multiplicative constants.

**Theorem 3.1.** Let $A$ be an averaging operator defined as in (2.1) with preserving local averaging operators $A_z$. Moreover, we assume that the $A_z$ are uniformly bounded in the following sense
\[ \exists C_{\text{uni}} > 0 \forall z \in \mathcal{N}, \forall p_z \in P_1(T_z, M) : C_{\text{uni}} \|p_z\|_{L^2(T_z, M)} \geq \|A_z p_z\|_M \sqrt{|T_z|}. \]
Then
\[ \eta_M \leq \eta_A \leq (1 + \sqrt{3/2} C_{\text{uni}}) \eta_M. \]
for all $p_h \in P_h$.

**Proof.** The first inequality is obvious and the proof concerns the second. The operator norm of $A$ is bounded by all $\sqrt{3/2} C_{\text{uni}}$ because for all $p_h \in P_1(T, M)$
\[ \|Ap_h\|_{L^2(\Omega, M)}^2 = \int_\Omega \left| \sum_{z \in \mathcal{N}} A_z(p_h|T_z) \varphi_z \right|^2 dx \leq \int_\Omega 3 \sum_{z \in \mathcal{N}} \|A_z(p_h|T_z)\|_M^2 \varphi_z^2 dx \]
\[ \leq 3 \sum_{z \in \mathcal{N}} C_{\text{uni}}^2 \|p_h\|_{L^2(T_z, M)}^2 \int_\Omega \varphi_z^2 dx = 3 \sum_{z \in \mathcal{N}} C_{\text{uni}}^2 \|p_h\|_{L^2(T_z, M)}^2 \frac{|T_z|}{|T|} \]
\[ = 3/2 C_{\text{uni}}^2 \|p_h\|_{L^2(\Omega, M)}^2. \]
where we used $\|\varphi_z\|_{L^2(T_z)}^2 = |T_z|/6$ for $T \in T_z$. Now let $p_h \in P_h$. There is a unique decomposition of $p_h$ into a component $p_c$ in $Q_h$ and into a component $p_d$ in its orthogonal complement in $L^2(\Omega; M)$. Because the local averaging operators $A_z$ are preserving, we have $A p_c = p_c$. Thus
\[ \|p_h - A p_h\|_{L^2(\Omega, M)}^2 = \|p_d - A p_d\|_{L^2(\Omega, M)}^2 \leq \|p_d\|_{L^2(\Omega, M)}^2 + \|A p_d\|_{L^2(\Omega, M)} \leq (1 + \sqrt{3/2} C_{\text{uni}}) \eta_M, \]
where we used that $\eta_M = \|p_d\|_{L^2(\Omega)}$. \(\Box\)

Recall the averaging of nodal values discussed in Example 2.1. Computing the minimum of the polynomial
\[ (\alpha, \beta) \mapsto \int_0^1 \int_0^{1-y} (\alpha x + \beta y + 1)^2 dx dy \]
and scaling show that $\|p_z|T(z)\|_M$ is bounded by $6 \|p_z\|_{L^2(T_z, M)}/\sqrt{|T|}$. Hence for $p_z \in P_1(T_z, M)$
\[ \sqrt{|T|} \|A_z p_z\|_M \leq 6 \sum_{T \in T_z} |\lambda_{T,z}| \sqrt{|T^z|} \|p_z\|_{L^2(T_z, M)} \leq 6 \left( \sum_{T \in T_z} |\lambda_{T,z}|^2 \frac{|T^z|}{|T|} \right)^{1/2} \|p_z\|_{L^2(T_z, M)}. \]
Consequently,
\[ C_{\text{uni}} = \max_{z \in \mathcal{N}} 6 \left( \sum_{T \in T_z} |\lambda_{T,z}|^2 \frac{|T^z|}{|T|} \right)^{1/2}. \]
For the ZZ estimator we conclude
\[ C_{\text{uni}} = \max_{z \in \mathcal{N}} 6 \left( \sum_{T \in T_z} \frac{|T^z|^2}{|T|^2} \frac{|T_z|}{|T|} \right)^{1/2} = 6. \]
Notice that we do not require shape regularity.
Remark 3.1. In [10] it is shown that the above bound can be sharpened with a more elaborate argument based on Ascoli’s lemma and the explicit calculation of stiffness matrices. There it is demonstrated that

\[ \eta_Z \leq \sqrt{10} \eta_M \approx 3.2 \eta_M \leq (1 + \sqrt{3/26}) \eta_M \approx 8.3 \eta_M. \]

In this paper also the three dimensional setting is analyzed.

We now turn to the equivalence of \( \eta_M, \eta_Z \) and \( \eta_E \). The estimators \( \eta_M, \eta_Z \) and \( \eta_E \) then satisfy the relationship

\[ (C_{en} C_{sr})^{-1/2} \eta_Z \leq \eta_E \leq (24 C_{sr})^{1/2} \eta_M. \]

Proof. Suppose that the jumps in the patch \( T_z \) are computed in clockwise direction, i.e. given an edge \( E \in \mathcal{E}_z \) which is the boundary between the elements \( T_1, T_2 \in \mathcal{T}_z \), we set \( \| p_h \|_E = p_h|_{T_1} - p_h|_{T_2} \) if \( T_1 \) is clockwise positioned ahead of \( T_2 \) with respect to \( z \). Then

\[ \sum_{E \in \mathcal{E}_z} \| p_h \|_E (z) = 0. \]

Thus, for all \( T \in \mathcal{T}_z \), we have

\[ \| p_h |_{T} (z) - Z(p_h |_{T_2}) \|_M \leq \max_{S \in \mathcal{E}_z} \| p_h |_{T} (z) - p_h |_{S} (z) \|_M \leq \frac{1}{2} \sum_{E \in \mathcal{E}_z} \| p_h \|_E (z) \|_M. \]

For all affine functions \( p \in \mathcal{E}_z \) the bound

\[ h_E \| p(z) \|_{E}^{2} \leq 4 \| p \|_{L^2(E,M)}^{2} \]

holds. Thus

\[ \eta_Z^2 = \int_{\Omega} \left( \sum_{z \in \mathcal{N}} (p_h - Z(p_h |_{T_z})) \varphi_z \right)^2 \| \|_M dx = 3 \int_{\Omega} \sum_{z \in \mathcal{N}} \| p_h - Z(p_h |_{T_z}) \|_M^2 \varphi_z^2 dx \]

\[ \leq 3 \int_{\mathcal{O}} \sum_{z \in \mathcal{N}} \left( \frac{1}{2} \sum_{E \in \mathcal{E}_z} \| p_h \|_E (z) \|_M \right)^2 \varphi_z^2 dx \leq \frac{3}{4} \int_{\Omega} \sum_{z \in \mathcal{N}} C_{en} \sum_{E \in \mathcal{E}_z} \| p_h \|_E (z) \|_M^2 \varphi_z^2 dx \]

\[ = \frac{3 C_{en}}{24} \sum_{z \in \mathcal{N}} \sum_{E \in \mathcal{E}_z} \| p_h \|_E (z) \|_M^2 |T_z| \leq \frac{C_{en} C_{sr}}{8} \sum_{z \in \mathcal{N}} \sum_{E \in \mathcal{E}_z} \| p_h \|_E (z) \|_M^2 h_E^2 \leq C_{en} C_{sr} \eta_E^2, \]

taking for the last inequality into account that the double sum passes over each edge twice. With \( p_d \) like in the proof of the last theorem, the chain of inequalities

\[ \eta_E^2 = \sum_{E \in \mathcal{E}} h_E \int_{\mathcal{O}} \left( \sum_{z \in \mathcal{N}} \| p_d \|_E (z) \varphi_z \right)^2 \| \|_M dx \leq \sum_{z \in \mathcal{N}} \sum_{E \in \mathcal{E}_z} h_E \| p_d \|_E (z) \|_M \| \varphi_z \|_{L^2(E)} \]

\[ = \sum_{z \in \mathcal{N}} \sum_{E \in \mathcal{E}_z} \frac{h_E^2}{3} \| p_d \|_E (z) \|_M^2 \leq \sum_{z \in \mathcal{N}} \sum_{T \in \mathcal{T}_z} \frac{2 h_E^2}{3} \| p_d \|_T (z) \|_M^2 \leq \sum_{z \in \mathcal{N}} \sum_{T \in \mathcal{T}_z} \frac{2 h_E^2}{3} \| p_d \|_{L^2(T,M)}^2 \frac{2}{|T|} \]

\[ \leq 24 C_{sr} \eta_M^2 \]

completes the argument.
4. Local Approximation Operators

The averaging operators $A$, considered in the last section, belong to the important class of the local approximation operators. By local approximation operators we mean operators of the structure

\[ J : V \to V_h, \quad p \mapsto \sum_{z \in \mathcal{N}} ((\Pi_z \circ A_z)(p|_{T_z})) \varphi_z. \]

The operators $J$ have through $V$, $V_h$ and $\Pi_z$ additional flexibility compared to $A$, which is used to incorporate additional constraints to the approximation if needed. For example, one may choose $V$, $V_h$ and $\Pi_z$ such that $J(v)$ satisfies the boundary conditions of the underlying boundary value problem exactly. However, also in other contexts constraints arise which can be incorporated into $J$; for instance, obstacle conditions in the field of variational inequalities. To give consideration to a wide range of applications, we cover the general case.

To be compatible with the framework of the operators $A$, the space $V$ needs to contain, in general, the set $\mathcal{N}$. In other situations local approximation operators are used to project the exact solution of the boundary value problem into the finite element space. For these applications $V$ needs to contain functions of $H^1(\Omega)$-type. For this reason we assume that $V$ is a subset of the broken Sobolev space

\[ H^1(\mathcal{T}, \mathbb{M}) := \{ v \in L^2(\Omega, \mathbb{M}) : \forall T \in \mathcal{T} : v|_T \in H^1(\mathcal{T}, \mathbb{M}) \}. \]

The space $H^1(\mathcal{T}, \mathbb{M})$ is equipped with the semi-norm

\[ |p|^2_{H^1(\mathcal{T}, \mathbb{M})} := \sum_{T \in \mathcal{T}} |(p|_T)|^2_{H^1(\mathcal{T}, \mathbb{M})} + \sum_{E \in \mathcal{E}_h} \| p|_E \|_{L^2(\mathcal{E}, \mathbb{M})}^2. \]

and the norm

\[ \| p \|^2_{H^1(\mathcal{T}, \mathbb{M})} := \| p \|_{L^2(\Omega, \mathbb{M})} + |p|^2_{H^1(\mathcal{T}, \mathbb{M})}. \]

We now turn to $V_h$. For all $z \in \mathcal{N}$ let $A_z$ be a non-empty, convex and closed subset of $\mathbb{M}$ and let $\Pi_z : \mathbb{M} \to A_z$ be the orthogonal projection onto $A_z$ in the canonical scalar product of $\mathbb{M}$. We assume the $A_z$ are chosen such that the following compatibility condition between $V$ and $V_h$ is satisfied:

\[ \exists C_{pf} > 0 \forall z \in \mathcal{N} \forall p \in V|_{T_z} : \| p - \Pi_z \bar{p} \|_{L^2(T_z, \mathbb{M})} \leq C_{pf} \| p \|_{H^1(T_z, \mathbb{M})}. \]

where

\[ h : \Omega \to \mathbb{R}, \quad x \in T \mapsto \text{diam}(T) \text{ and } \bar{p} := \frac{1}{\int_{T_z}} p \, dx. \]

Here and throughout the text we identify real numbers, like $\Pi_z \bar{p}$ in the above formula, with the constant function attaining this function value. In our analysis, (4.2) is used similar to a Poincaré-Friedrichs inequality. If (4.2) is fulfilled we define

\[ V_h := \{ \sum_{z \in \mathcal{N}} a_z \varphi_z : a_z \in A_z \} \subset P_1(\mathcal{T}, \mathbb{M}). \]

Instead of the orthogonal projection other choices for $\Pi_z$ are possible. We remark that in some of the subsequent proofs the expansion factor of the projection has to be included explicitly for more general $\Pi_z$. For the orthogonal projection the factor is equal to 1:

\[ \forall p_1, p_2 \in \mathbb{M} : \| \Pi_z p_1 - \Pi_z p_2 \|_\mathbb{M} \leq \| p_1 - p_2 \|_\mathbb{M}. \]

We now turn to the operators $A_z : V_h |_{T_z} \to \mathbb{M}$. We assume they satisfy the condition that

\[ \exists C_j > 0 \forall z \in \mathcal{N} \forall p \in V|_{T_z} : \| A_z p - \bar{p} \|_{L^2(T_z, \mathbb{M})} \leq C_j \| p \|_{H^1(T_z, \mathbb{M})}. \]

We show that the ZZ estimator satisfies (4.4).

**Theorem 4.1.** Suppose that $\mathcal{T}$ is shape-regular with $C_s$ and $C_{en}$ like in Theorem 3.2. Then

\[ \| Z p - \bar{p} \|_{L^2(T_z, \mathbb{M})} \leq \max\{ \sqrt{C_{en}} C_s, C_p \} | h p |_{H^1(T_z, \mathbb{M})} \]

where $C_p$ is the constant in the Poincaré inequality for the elements of the triangulation $\mathcal{T}$. 

\[ \]
PROOF. We have
\[
\|Zp - \tilde{p}\|_{L^2(T, \mathbb{M})} \leq \|p - Zp\|_{L^2(T, \mathbb{M})} + \|p - \tilde{p}\|_{L^2(T, \mathbb{M})} \\
\leq \sqrt{\sum_{T \in \mathcal{T}} C_T \eta_T^2} + \sum_{T \in \mathcal{T}_S} C_T \text{diam}(T) \|p\|_{H^1(T, \mathbb{M})} \\
\leq \max\{\sqrt{\sum_{T \in \mathcal{T}_S} C_T}, C_T\} \|p\|_{H^1(T, \mathbb{M})},
\]
according to Theorem 3.2. \qed

The \(A_z\) are not necessarily pointwise interpolation operators. It is self-evident that
\[
A_z p = \tilde{p}
\]
satisfies (4.4). We remark that similarly the important class of quasi-interpolation operators introduced by Clément fits into the framework of local approximation operators, provided the definition of \(A_z\) is relaxed to \(A_z : \Omega \rightarrow \mathbb{R}\), where \(\Omega\) is a superset of \(T_z\). While in the interior of the domain Clément operators also utilise (4.5), for boundary nodes \(z\) the average of \(p\) in \(T_z\) is not assigned to the coefficient of \(\varphi_z\) but added to the coefficient of a neighbouring interior node.

In the next theorem we consider the rate of convergence of local approximation operators.

**Theorem 4.2.** Local approximation operators \(J_z\), satisfying (4.2) and (4.4), are locally first-order approximating, which means there is a constant \(C > 0\) such that for all \(p \in V\)
\[
\|p - J_z p\|_{L^2(T_z, \mathbb{M})} \leq C \|p\|_{H^1(T, \mathbb{M})}.
\]
For \(J_z\) the constant \(C\) is equal to \(\sqrt{3} (C_{pf} + C_j)\). If \(T\) is shape-regular, then all \(p \in V\) also satisfy
\[
\|h^{-1}(p - J_z p)\|_{L^2(T_z, \mathbb{M})} \leq \sqrt{3} C_T (C_{pf} + C_j) \|p\|_{H^1(T, \mathbb{M})},
\]
where
\[
C_T = \max_{z \in N} \max_{S, \mathcal{T} \in \mathcal{T}_z} \frac{\text{diam}(S)}{\text{diam}(\mathcal{T})}.
\]

**Proof.** Given \(p \in V\) let \(p_z := J_z(p|_{T_z})\) and \(\tilde{p}_z := \int_{T_z} p\, dx\). Applying (4.3), we deduce
\[
\|p - \Pi_z(p_z)\|_{L^2(T_z, \mathbb{M})} \leq \|p - \Pi_z(\tilde{p}_z)\|_{L^2(T_z, \mathbb{M})} + \|\Pi_z(\tilde{p}_z) - \Pi_z(p_z)\|_{L^2(T_z, \mathbb{M})} \\
\leq (C_{pf} + C_j) \|p\|_{H^1(T, \mathbb{M})}.
\]
Therefore
\[
\|p - J_z p\|_{L^2(T_z, \mathbb{M})} = \int_{T_z} \sum_{z \in N} \varphi_z \|p - \Pi_z p_z\|_{\mathbb{M}}^2 dx \\
\leq \sum_{z \in N} (C_{pf} + C_j)^2 \|p\|_{H^1(T, \mathbb{M})}^2 \\
\leq 3 (C_{pf} + C_j)^2 \|p\|_{H^1(T, \mathbb{M})}^2.
\]
Now suppose that \(T\) is shape-regular. Then
\[
\|h^{-1}(p - J_z p)\|_{L^2(T_z, \mathbb{M})} = \sqrt{3} C_T (C_{pf} + C_j) \|p\|_{H^1(T, \mathbb{M})}.
\]
completes the proof. \qed

Local approximation operators are stable in the \(H^1(\Omega)\)-norm.

**Theorem 4.3.** Let \(J\) be a local approximation operator satisfying (4.2) and (4.4) and let \(T\) be shape-regular with \(C_{sr}\) like in Theorem 3.2. Then for all \(p \in V\)
\[
|J p|_{H^1(\Omega, \mathbb{M})} \leq \sqrt{3} C_T (C_{pf} + C_j) \|p\|_{H^1(T, \mathbb{M})}.
\]
RELIABLE AND EFFICIENT A POSTERIORI FINITE ELEMENT ERROR CONTROL

PROOF. Since \( \sum_{z \in \mathcal{N}} \varphi_z = 1 \) on \( \Omega \) it follows that \( \sum_{z \in \mathcal{N}} \nabla \varphi_z \) vanishes. Thus
\[
|Jp|^2_{H^1(\Omega,M)} = \int_{\Omega} \sum_k \left| \sum_{z \in \mathcal{N}} (\Pi_z(p) - p) \frac{\partial \varphi_z}{\partial n} \right|^2 dx \leq 3 \int_{\Omega} \sum_k ||\Pi_z(p) - p||^2_M (\partial_n \varphi_z)^2 dx
\]
\[
\leq 3 \max_{x,z} \left( \sum_k (h \partial_n \varphi_z)^2 \right) \sum_{z \in \mathcal{N}} \|h^{-1}(\Pi_z(p) - p)\|^2_{L^2(T,M)}.
\]
To compute \( \sum_k (h \partial_n \varphi_z)^2 \) we can assume without loss of generality that \( T \) has the vertices \((0,0), (0,\xi_1)\) and \((\eta_2,\xi_2)\) and that \( \varphi_z(0,0) = 1 \) and \( \varphi_z(0,\xi_1) = \varphi_z(\eta_2,\xi_2) = 0 \). Then
\[
\varphi_z(x,y) = \frac{\xi_2/\xi_1 - 1}{\eta_2} x - \frac{1}{\xi_1} y + 1.
\]
Since \( |T| = |\xi_1\eta_2|/2 \), it follows that
\[
\sum_k (h \partial_n \varphi_z)^2 = \left( \frac{h \xi_2/\xi_1 - h}{\eta_2} \right)^2 \left( \frac{h}{\xi_1} \right)^2 \left( \frac{\xi_2 - \xi_1}{2|T|^2} \right)^2 = \frac{h^2((\xi_2 - \xi_1)^2 + \eta_2^2)}{4|T|^2} \leq \frac{h^2}{2}|T|^2 \leq \frac{C_{tr}^2}{2}.
\]
We already demonstrated in (4.6) that
\[
\sum_{z \in \mathcal{N}} \|h^{-1}(\Pi_z(p) - p)\|^2_{L^2(T,M)} \leq 3 C_{tr}^2 (C_{pf} + C_j)^2 |p|^2_{H^1(T,M)},
\]
which concludes the proof. \( \square \)

For shape-regular triangulations \( T \) there is a constant \( C_{tr} \) such that for all \( T \in \mathcal{T} \) and \( p \in H^1(T,M) \)
\[
\|v\|^2_{L^2(\Omega,T,M)} \leq C_{tr}(\|h^{-2}v\|^2_{L^2(T,M)} + |v|^2_{H^1(T,M)}),
\]
cf. [11]. Together with the stability of \( J \) in \( H^1(\Omega) \) this implies that the approximation of \( p \) by \( Jp \) is on the elemental boundaries of order one half.

THEOREM 4.4. Let \( J \) be a local approximation operator satisfying (4.2) and (4.4) and let \( T \) be shape-regular with \( C_{sr} \) like in Theorem 3.2. Then for all \( p \in V \)
\[
(4.7) \quad \sum_{T \in \mathcal{T}} \|h^{-1}(p - Jp)\|^2_{L^2(T,M)} \leq C_{tr}(1 + C_{tr}^2)(2 + 3 C_{sr}^2 C_{tr}^2 (C_{pf} + C_j)^2) |p|^2_{H^1(T,M)}.
\]

PROOF. The chain of inequalities
\[
\sum_{T \in \mathcal{T}} \|h^{-1}(p - Jp)\|^2_{L^2(T,M)} \leq C_{tr}^2 \sum_{T \in \mathcal{T}} \|h^{-2}(p - Jp)\|^2_{L^2(T,M)} + |p - Jp|^2_{H^1(T,M)}
\]
\[
\leq C_{tr}^2 (1 + C_{tr}^2) \sum_{T \in \mathcal{T}} |p - Jp|^2_{H^1(T,M)}
\]
\[
\leq C_{tr}^2 (1 + C_{tr}^2)(2|p|_{H^1(T,M)} + 2|Jp|_{H^1(\Omega,M)})^2
\]
\[
\leq C_{tr}^2 (1 + C_{tr}^2)(2 + 3 C_{sr}^2 C_{tr}^2 (C_{pf} + C_j)^2) |p|^2_{H^1(T,M)}
\]
shows (4.7). \( \square \)

5. Reliability

While up to now we have not assumed that \( p \) and \( p_h \) are related to each other in a particular way, in applications these two functions are connected by an underlying boundary value problem and by the chosen numerical method.

In this section we focus on the proof of reliability of conforming finite element methods for elliptic problems. The analogous results for nonconforming methods and for saddle-point problems are deduced, for instance, in [15].

Applying the Einstein summation convention, we consider differential equations of the form
\[
(5.1) \quad -\partial_k C_{ijkl} \partial_i u_j = f_t,
\]
with \( u = (u_1, u_2) \in H^1(\Omega, \mathbb{R}^2) \), \( f \in L^2(\Omega, \mathbb{R}^2) \) and \( C \in W^{1,\infty}(\Omega, (\mathbb{R}^2)^{2 \times 2}) \). We require that \( C \) is symmetric and uniformly positive definite on \( \mathbb{R}^{2 \times 2} \); that is \( C_{ijkl} = C_{klij} \) and there are two positive constants \( \mu \) and \( \bar{\mu} \) such that
\[
\mu v_{ij} v_{ij} \leq v_{ij} C_{ijkl}(x) v_{kl} \leq \bar{\mu} v_{ij} v_{ij},
\]
for all \( x \in \Omega \) and \( v \in \mathbb{R}^{2 \times 2} \). Denoting by \( \lambda(x) \) the smallest and by \( \bar{\lambda}(x) \) the largest eigenvalue of \( C \), we may select \( \mu = \min_{x \in \Omega} \lambda(x) \) and \( \bar{\mu} := \max_{x \subseteq \Omega} \bar{\lambda}(x) \).

In the context of structural mechanics \( C \) is interpreted as elasticity tensor. For these applications the error control on the flux \( p - p_h \), where \( p_{ij} := C_{ijkl} \partial_l u_{ij} \) and \( (p_h)_{ij} := C_{ijkl} \partial_l (u_{h})_{ij} \), is often more relevant than the error control on \( u \). Since \( C \) is, in general, not piecewise linear, \( p_h \) might not be contained in \( P_h \) as defined in Section 2, a property we do not require for the proof of reliability of \( \eta \). Towards the end of the section we comment on the condition \( p_h \in P_h \) again.

We define the residual of the finite element method as \( r_l := -\partial_h C_{ijkl}(p - p_h)_{ij} = f_l - \partial_h C_{ijkl}(p_h)_{ij} \).

**Theorem 5.1.** Suppose we have the Galerkin orthogonality
\[
\int_{\Omega} (p_{kl} - (p_h)_{kl}) \partial_k (f(u - u_h))_l \, dx = 0,
\]
where \( J \) is a suitable local approximation operator. Then
\[
\|p - p_h\|_{L^2(\Omega, M)} \leq \sqrt{3} C_t \left( C_{pf} + C_j \right) \frac{h}{\mu} \|r\|_{L^2(\Omega, M)} + C_E \frac{\bar{h} \eta}{\mu},
\]
where
\[
C_E := C_{tr} (1 + C_{pf}^2)(2 + 3 C^2_M C^2_t (C_{pf} + C_j)^2) \max_{T \in T} \frac{h_T}{\eta_T}
\]
\[
\text{PROOF. Let } \zeta := u - u_h. \text{ Using that } p \in H^1(\Omega, \mathbb{R}^{2 \times 2}) \text{, we calculate}
\]
\[
\int_{\Omega} (p_{kl} - (p_h)_{kl}) \partial_k \zeta_1 \, dx = \int_{\Omega} (p_{kl} - (p_h)_{kl}) \partial_k (\zeta_1 - (J \zeta_1)_l) \, dx
\]
\[
\text{= } \int_{\Omega} \partial_k (p_{kl} - (p_h)_{kl}) (\zeta_1 - (J \zeta_1)_l) \, dx + \sum_{T \in T} \int_{\partial T} (p_{kl} - (p_h)_{kl}) (\zeta_1 - (J \zeta_1)_l) \, dx
\]
\[
\text{= } \int_{\Omega} r_l (\zeta_1 - (J \zeta_1)_l) \, dx + \sum_{E \subseteq E} \int_{E} (p_{kl} - (p_h)_{kl}) (\zeta_1 - (J \zeta_1)_l) (\nu_E)_k \, dx
\]
\[
\text{= } \|r\|_{L^2(\Omega, M)} \|h^{-1} (\zeta - (J \zeta)_1)\|_{L^2(\Omega, M)} + \eta_E \left( \sum_{E \subseteq E} h^{-1}_E \|\zeta - (J \zeta)_1\|^2_{L^2(E)} \right)^{1/2}
\]
\[
\leq \|r\|_{L^2(\Omega, M)} \sqrt{3} C_t (C_{pf} + C_j) \|\zeta\|_{H^1(T, M)} + \eta_E C_E \|\zeta\|_{H^1(T, M)}.
\]
The constant \( C_E \) contains besides the coefficients in (4.7) a factor for the transition from \( h_T \) to the side length \( h_E \). Assuming that \( \|\zeta\|_{H^1(T, M)} = \|\zeta\|_{H^1(T, M)} \neq 0 \),
\[
\|p - p_h\|_{L^2(\Omega, M)} \leq \frac{1}{\mu} \left( \left( \int_{\Omega} (p_{kl} - (p_h)_{kl}) \partial_k \zeta_1 \, dx \right)^{1/2} \right) \leq \frac{1}{\mu} \frac{\int_{\Omega} (p_{kl} - (p_h)_{kl}) \partial_k \zeta_1 \, dx}{\|\zeta\|_{H^1(\Omega, M)}}.
\]
Combining both bounds proves the theorem. \( \square \)

Condition (5.2) can, for instance, be satisfied for the standard \( P_h \) finite element method with essential Dirichlet boundary conditions. If \( u_h \) coincides with \( u \) at \( N_T \) then, also in the case of non-homogeneous boundary conditions, the term \( J(u - u_h) \) vanishes if \( J \) is defined as nodal interpolation operator. If natural or mixed boundary conditions are imposed, additional terms can arise, for details we refer to [15].

We use (5.3) to show the reliability of \( \eta \). Recall that to apply \( \eta \) for error control, we need to express \( C_{rel} \) and h.o.t. defined in the introduction in terms of computable quantities. We do
so with help of patchwise data oscillations

\[ \text{osc}(r; T) := \left( \sum_{z \in N_I} \text{diam}(T_z)^2 \left\| r - \int_{T_z} r \, dx \right\|^2_{L^2(T_z, M)} \right)^{1/2}. \]

Observe that the summation ranges only over the internal nodes. Our next step is to bound \( r \) locally in terms of \( r \) and \( \text{osc}(r; N) \). The choice of the numerical method is again incorporated into the analysis by (5.5) in form of a Galerkin orthogonality, which implicitly leads to the assumption that

\[ \sum_{z \in N_I} a_z \varphi_z : a_z \in \mathbb{R} \]

is a subset of the approximation space of the finite element method. Observe that we require in the next theorem \( p \in H^1(\Omega; \mathbb{R}^{2 \times 2}) \).

**Theorem 5.2.** Let \( z \in N_\Omega \) be an interior node. Suppose that \( p \in H^1(\Omega; \mathbb{R}^{2 \times 2}) \) and that

\[ \int_{\Omega} f \cdot \varphi_z \, dx = \int_{\Omega} p_h \cdot \text{div} \varphi_z \, dx. \]

Then for the residual \( r \) we have the local bound

\[ \| h r \|_{L^2(T_z, M)} \leq (1 + \sqrt{8}/27 |T_z|) \text{osc}(r; T_z)^2 + \sqrt{8}/27 \| h \|_{L^\infty(T_z)}^2 \left( \sum_{E \in E_z} \sqrt{h_E} \| [p_h] \cdot \nu_E \|_{L^2(E)} \right)^2. \]

**Proof.** We denote the \( L^2(\Omega, M) \) scalar product by \( \langle \cdot, \cdot \rangle_{\Omega,M} \). Using that \( [p]_{\partial E} = 0 \) for \( E \in E_z \) and (5.5) we conclude that

\[ \langle r_1, \varphi_z \rangle_{\Omega,M} = \langle (p - p_h)_{kl} \partial_k \varphi_z \rangle_{\Omega,M} + \sum_{E \in E_z} \int_E [p_h]_{kl} \varphi_z \langle \nu_E \rangle_k \, ds \]

\[ = \sum_{E \in E_z} \int_E [p_h]_{kl} \varphi_z \langle \nu_E \rangle_k \, ds \leq \sum_{E \in E_z} \sqrt{h_E} \| [p_h] \cdot \nu_E \|_{L^2(E)}. \]

The last bound followed from the Cauchy-Schwarz inequality. We denote the identity matrix by \( I \).

For the next computation we take into account that \( \bar{r}_z \) and \( (\varphi_z - \bar{\varphi}_z)I \) are \( L^2(\Omega, M) \)-perpendicular to each other, where \( \bar{r}_z := \int_{T_z} r \, dx \) and \( \bar{\varphi}_z := \int_{T_z} \varphi_z \, dx \). With \( 1 = \bar{\varphi}_z(I, I)_{\Omega,M}/(\varphi_z, I, I)_{\Omega,M} \),

\[ \| \bar{r}_z \|_{L^2(T_z, M)}^2 = \frac{\langle r, I \rangle_{\Omega,M}^2}{\langle \varphi_z, I, I \rangle_{\Omega,M}^2} (\langle r, \varphi_z - \bar{\varphi}_z \rangle_{\Omega,M} - (r - \bar{r}_z, (\varphi_z - \bar{\varphi}_z)I)_{\Omega,M})^2. \]

From Cavaliere's principle it follows that \( \langle \varphi_z I, I \rangle_{\Omega,M} = \sqrt{2} \| \varphi_z \|_{L^1(T_z)} = \sqrt{2} |T_z|/3 \). We calculate that \( ||(\varphi_z - \bar{\varphi}_z)I||_{L^2(\Omega,M)}^2 = 2 |T_z|/18 \). Consequently,

\[ \| \bar{r}_z \|_{L^2(T_z, M)}^2 \leq \frac{\langle I, I \rangle_{\Omega,M}^2}{\langle \varphi_z, I, I \rangle_{\Omega,M}^2} \left( 2 \| r, \varphi_z \|_{L^2(T_z, M)}^2 \right) \left( \| r, \varphi_z \|_{L^2(T_z, M)}^2 \right) \| r - \bar{r}_z \|_{L^2(T_z, M)}^2 \]

\[ = \sqrt{8}/3 \| r, \varphi_z \|_{\Omega,M}^2 + \sqrt{8}/27 |T_z| \| r - \bar{r}_z \|_{L^2(T_z, M)}^2. \]

Therefore,

\[ \| h r \|_{L^2(T_z, M)}^2 \leq \| h \|_{L^\infty(T_z)}^2 (\| r - \bar{r}_z \|_{L^2(T_z, M)} + \| \bar{r}_z \|_{L^2(T_z, M)} + \| r - \bar{r}_z \|_{L^2(T_z, M)}) \]

\[ \leq \| h \|_{L^\infty(T_z)}^2 (1 + \sqrt{8}/27 |T_z|) \| r - \bar{r}_z \|_{L^2(T_z, M)} \]

\[ \leq (1 + \sqrt{8}/27 |T_z|) \text{osc}(r; T_z)^2 + \sqrt{8}/3 \| h \|_{L^\infty(T_z)}^2 \left( \sum_{E \in E_z} \sqrt{h_E} \| [p_h] \cdot \nu_E \|_{L^2(E)} \right)^2. \]

Notice that only the data oscillations over the subgrid \( T_z \) are needed. \( \square \)
We remark that by a Poincaré inequality we have for \( r \in H^1(\Omega, M) \), i.e. \( f \in H^1(\Omega, M) \),
\[
\text{osc}(r; T_j) \leq C_p \, \text{diam}(T_j) \|\nabla r\|_{L^2(T_j, M)},
\]
possibly by enlarging \( C_p \). Thus the data oscillations are higher-order terms in the sense of (1.1), provided the residuum of the numerical method is bounded from above, independently of \( h \).

We assume that the patches \( T_z \) of internal nodes \( z \) cover \( \Omega \). This allows us to derive an a posteriori bound without explicitly referring to the boundary conditions. In principle when computing the reliability constant \( C_{\text{rel}} \) only in terms of the internal nodes one should, for the sharpness of the estimate, take into account that while interior elements are always covered by three patches \( T_z, z \in \mathcal{N}_T \), at the boundary there are fewer patches covering the elements. However, for the sake of simplicity we do not make this distinction in the now following analysis.

Finally, for the computation of the reliability constant we use the similar structure between \( \eta_E \) and the jump terms in (5.6).

**Theorem 5.3.** Let \( \mathcal{T} \) be shape-regular. Provided that (5.2) and (5.5) hold and that patches of interior elements cover \( \Omega \), we have the bound
\[
\|p - p_h\|_{L^2(\Omega, M)} \leq C_{\text{rel}} \eta_E + \text{h.o.t.}_{\text{rel}}
\]
where
\[
C_{\text{rel}}^2 := 2 \frac{\mu^2}{\mu^2} \left( C^2_g + C_{\text{en}} \sqrt{8/27} \|h\|_{L^\infty(\Omega)}^2 \right),
\]
\[\text{h.o.t.}_{\text{rel}} := 18 \left( 1 + \sqrt{8/27} \max_{z \in \mathcal{N}_T} |T_z| \right) \frac{\mu^2}{\mu^2} C_p^2 C^2_g (C_{\text{pf}} + C_J)^2 \text{diam}(T)^2 \|\nabla r\|_{L^2(\Omega, M)}^2.
\]

**Proof.** Covering \( \Omega \) with interior patches gives
\[
\|h r\|_{L^2(\Omega, M)} \leq \sum_{z \in \mathcal{N}_T} \|h r\|_{L^2(T_z, M)}^2 \
\leq \sum_{z \in \mathcal{N}_T} (1 + \sqrt{8/27} |T_z|) C_p^2 \text{diam}(T)^2 \|\nabla r\|_{L^2(\Omega, M)}^2 + C_{\text{en}} \sqrt{8/27} \|h\|_{L^\infty(\Omega)}^2 \sum_{T \in \mathcal{E}_e} h_E \|\|p_h\|_{E} - \nu_E\|_E^2
\leq 3 \left( 1 + \sqrt{8/27} \max_{z \in \mathcal{N}_T} |T_z| \right) C_p^2 \text{diam}(T)^2 \|\nabla r\|_{L^2(\Omega, M)}^2 + C_{\text{en}} \sqrt{8/27} \|h\|_{L^\infty(\Omega)}^2 \eta_E.
\]

Combining this bound with (5.3) gives
\[
\|p - p_h\|_{L^2(\Omega, M)}^2 \leq 6 C_g^2 (C_{\text{pf}} + C_J)^2 \frac{\mu^2}{\mu^2} \|h r\|_{L^2(\Omega, M)}^2 + 2 C_g^2 \frac{\mu^2}{\mu^2} C_p^2 \eta_E^2
\]
\[
\leq 18 \left( 1 + \sqrt{8/27} \max_{z \in \mathcal{N}_T} |T_z| \right) \frac{\mu^2}{\mu^2} C_p^2 C^2_g (C_{\text{pf}} + C_J)^2 \text{diam}(T)^2 \|\nabla r\|_{L^2(\Omega, M)}^2
\]
\[
+ 2 \frac{\mu^2}{\mu^2} \left( C_g^2 + C_{\text{en}} \sqrt{8/27} \|h\|_{L^\infty(\Omega)}^2 \right) \eta_E^2,
\]
which demonstrates the reliability of \( \eta_E \). \( \square \)

If \( p_h \in \mathcal{P}_h \), e.g. \( \mathcal{C} \) is elementwise linear and \( \mathcal{P}_h = \mathcal{P}_h(T, M) \), then the reliability of \( \eta_A \), \( \eta_Z \) and \( \eta_M \) follows from the equivalence proofs in the section on efficiency. Since the equivalence proofs were in essence based on compactness arguments and did not depend on the linearity of \( p_h \) one can show \( \eta_A \approx \eta_Z \approx \eta_M \approx \eta_E \) in a more general context. That one can regard the adaptation to differential operators with variable coefficients and higher-order polynomial approximation spaces, we refer to [6].

**6. Applications**

This section gives an overview of some applications to the Stokes and Lamé equations and to elastoplastic, obstacle and degenerated problems in which averaging techniques are known to work. The arguments for the reliability and efficiency proof are partly those from the previous section, partly involve new ideas to cover nonconforming methods and saddle-point problems.
6.1. Stokes Equations. The stationary viscous flow inside a bounded volume $\Omega \subset \mathbb{R}^2$ of viscosity $\mu > 0$ is described by the velocity field $u : \Omega \rightarrow \mathbb{R}^2$ and the pressure variable $p : \Omega \rightarrow \mathbb{R}$. We assume Dirichlet boundary conditions. Then the Stokes problem reads: Given $f \in L^2(\Omega)^2$ and $g \in H^{1/2}(\partial \Omega)^2$, find $(u,p) \in H^1(\Omega)^2 \times L^2(\Omega)$ with
\[
\begin{align*}
\text{div } \sigma + f &= 0 \quad \text{in } \Omega, \\
\text{div } u &= 0 \quad \text{in } \Omega, \\
u &= g \quad \text{on } \Gamma,
\end{align*}
\]
where $\int_{\Omega} p(x) \, dx = 0$ and $\sigma := 2\mu \varepsilon(u) - pI$ with $\varepsilon(u) := (\nabla u + \nabla u^T)/2$. For some of the results below smoother $f$ and $g$ need to be considered. In the literature one often finds the non-symmetric form $\sigma = 2\mu \nabla u - pI$, which is equivalent to the symmetric model if Dirichlet conditions are imposed.

The weak formulation of the above Stokes problem is straightforward and leads to a mixed FEM with unknowns $u_h$ and $p_h$. The proper choice of finite spaces for the discrete velocities $u_h$ and the discrete pressures $p_h$ is less trivial, particularly for piecewise polynomials of low-order. We point, for instance, to Kouhia and Stenberg who considered $u_h \in S_k \times S_k^{nc}$ and piecewise constant $p_h$.

It is known that, for sufficiently fine meshes, unique discrete solutions $(u_h,p_h)$ for which with quasi-optimal a priori error bounds hold, cf. [21]. A posteriori error estimates are studied in [4, 17]. In particular, the reliability and efficiency of the estimators
\[
\eta_A := \|\sigma_h - A\sigma_h\|_{L^2(\Omega)} \leq \eta_M := \min \|\sigma_h - \tau_h\|_{L^2(\Omega)} \quad \text{for } \tau_h \in S^1(T).
\]
has been verified for the discrete stress field
\[
\sigma_h := 2\mu \varepsilon_T(u_h) - p_h I.
\]
For proofs and details see [17]. In that publication also mixed boundary conditions are examined. A surprising consequence from the mathematical analysis is that the averaging concerns the stress field $\sigma_h$ only and not the variables $\varepsilon_T(u_h)$ and $p_h$ separately. The numerical examples in [17] underline the high accuracy of $\eta_A$ in this application.

6.2. Linear Elasticity. The small deformations $u : \Omega \rightarrow \mathbb{R}^2$ of a 2D elastic body are modelled by the Navier-Lamé equations: Given $f \in L^2(\Omega)^2$, $u_D \in H^{1/2}(\Gamma_D)^2$, $g \in L^2(\Gamma_N)^2$, find $u \in H^1(\Omega)^2$ with
\[
\begin{align*}
\text{div } \sigma + f &= 0 \quad \text{in } \Omega, \\
\text{div } u &= 0 \quad \text{in } \Omega, \\
u &= u_D \quad \text{on } \Gamma_D, \\
\sigma v &= g \quad \text{on } \Gamma_N,
\end{align*}
\]
where
\[
\sigma := C \varepsilon(u) := \lambda \text{tr } (\varepsilon(u)) I + 2\mu \varepsilon(u),
\]
\[
\varepsilon(u) := (\nabla u + \nabla u^T)/2.
\]
Here $\Gamma_D$ is a closed non-empty subset of $\Gamma$ on which Dirichlet boundary conditions are imposed and $\Gamma_N$ is a relatively open non-empty subset with Neumann boundary conditions. We assume that $\Gamma_D \cup \Gamma_N = \Gamma$. A new aspect is compressibility with the material constant $\lambda$ which tends to infinity as the Poisson ratio tends to 1/2 for rubber-like materials. It is known that in the incompressible limit for $\lambda \rightarrow \infty$ that the elastic problem turns into the Stokes problem as $\lambda \text{tr } (\varepsilon(u)) \rightarrow p$ and $\text{tr } \varepsilon(u) = \text{div } (u) \rightarrow 0$. In principle, the linear elastic problem can be discretised by $u_h$ in $S_1 \times S_1$. This leads to quasi-optimal convergence in the energy norm,
\[
\| C^{-1/2}(\sigma - \sigma_h) \| \leq C(\lambda) h_{\text{max}} \| D^2 u \|_{L^2(\Omega)}
\]
for a smooth exact solution $u$. Therein, the multiplicative constant $C(\lambda)$ might deteriorate if $\lambda \rightarrow \infty$. Figure 1 displays a numerical example from [16] for three different materials with Poisson ratios $nu = 0.3$, 0.49, and 0.499. For a uniform sequence of meshes on an $L$-shaped
domain $\Omega$ with known singular exact solution $u$, we computed discrete solutions $u_h \in \mathcal{S}_1 \times \mathcal{S}_1$ within a standard $P_1$ FE. The error of the corresponding discrete stress $\sigma_h := C \varepsilon(u_h)$, with the fourth-order material tensor $C$ and $\sigma = C \varepsilon(u)$, reads, in its energy norm

$$\| C^{-1/2}(\sigma - \sigma_h) \|^2 := \int_{\Omega} (\sigma - \sigma_h) : \varepsilon(u - u_h) \, dx.$$  

The quantity $\| C^{-1/2}(\sigma - \sigma_h) \|$ is plotted as a function of the number of degrees of freedom in Figure 1. For uniform meshes we observe a suboptimal convergence rate caused by the singularity of $u$. The experimental convergence rate appears to be independent of the Poisson ratio $\nu$ (i.e. of $\lambda$) in contrast to the multiplicative constant $C(\lambda)$. This phenomenon is called locking [9]. In Figure 1, the numerical result for a uniform mesh with $N = 10000$ degrees of freedom and $\nu = 0.499$ is worse than that for the coarsest mesh with $N = 16$ degrees of freedom for $\nu = 0.3$. The situation is even more dramatic for larger $\lambda \to \infty$.

Figure 1 displays three sequences of adaptively refined meshes for $\nu = 0.3$, $0.49$, and $0.499$ as well. The coarse meshes coincide with the results for the uniformly refined ones but improves with a convergence rate larger than 1 until the error is much smaller. Said differently, the effect of the multiplicative constant $C(\lambda)$ is seen in the beginning for $N \leq 100$ and decreases for larger $N$. Does this indicate the conjecture that adaptivity overcomes locking?

The error control by averaging schemes via

$$\eta_M := \min_{\tau_h \in S_0(T)} \| C^{-1/2}(\sigma_h - \tau_h) \|_{L^2(\Omega)} \leq \eta_A := \| C^{-1/2}(A\sigma - \sigma_h) \|_{L^2(\Omega)}$$

is displayed in Figure 1 as well. It is proved in [16] that $\eta_M \leq \eta_A$ is reliable up to $\lambda$-depending constants and, clearly, $\eta_M$ is efficient with respect to $\lambda$-independent constants. As observed in Figure 1, even a very poor finite element solution is estimated very accurately.

The preceding discussion focused on conforming FEM and large errors caused by locking in the incompressible limit $\lambda \to \infty$. More appropriate FEMs can overcome this locking. The first
It is elementary to verify that the error $e$ expressed by means of residuals $\epsilon$ the FE discretisation replaces $K$ for the exact and discrete stress field $\sigma$ and $\sigma_h$, respectively, have the same definition as in Subsection 7.2 for each time step. It can be proved for an implicit time-discretisation that

$$\| C^{-1/2}(\sigma - \sigma_h) \|^2 \leq \int_\Omega (\sigma - \sigma_h) : \varepsilon(u - u_h) \, dx = \int_\Omega (\sigma - \sigma_h) : \varepsilon(u - u_h - v_h) \, dx \quad \text{for all } v_h \in S_h \times S_h.$$ 

This term is an upper bound for error terms such as the stress error in the energy norm. Moreover, if hardening is present, the stress error controls the displacement error $\| u - u_h \|_{H^1(\Omega)}$ up to hardening depending multiplicative constants. For details and proofs we refer to [2, 13, 14]. Therefore, one can proceed as in linear elasticity to derive reliability and efficiency of $\eta_M \leq \eta_A$. The constant $C_{rel}$, however, depends crucially on the hardening; the estimates are not valid (in this form) for perfect plasticity.

It should be emphasized that reliability holds solely for the spatial discretisation; the accumulated error in time is not controlled by $\eta_M \leq \eta_A$; the result holds for Hencky materials only. The control of the time-discretisation error appears to be an important open question. However, the numerical results in [14] provide numerical evidence that $\eta_A$ is indeed a very accurate (spatial) error estimator.

6.3. Elastoplasticity. This section briefly describes the perspectives and limitations of averaging techniques in elastoplastic evolution problems. Therein, a time-discretisation is performed followed by a spatial discretisation in each time-step. Averaging error estimators $\eta_M$ for incompressibility. The choice of $S_h \times S_h^{nc}$ due to Kouhia and Stenberg [21] is appropriate for that and leads in [16] to robust and accurate estimates. Therein, the finite element schemes as well as their error estimators are highly accurate and $\lambda$-independent.

6.4. Obstacle Problems. This section briefly highlights the surprising result that for nonlinear variational inequalities, in certain settings, the same averaging estimator

$$\eta_M := \min_{q_h \in \mathcal{Q}_h} \| p_h - q_h \| \leq \eta_A := \| p_h - Ap_h \|$$

is, as for the variational equation, reliable and efficient: An affine obstacle has no substantial influence! We illustrate the situation with a linear model problem. Let $K$ denote the set of admissible deformations,

$$K := \{ v \in H^1_0(\Omega) : 0 \leq v \text{ almost everywhere in } \Omega \}$$

with $H^1_0(\Omega) := \{ v \in H^1(\Omega) : v = 0 \text{ on } \partial\Omega \}$. Then the weak form of the obstacle problem reads: Given $f \in L^2(\Omega)$ find $u \in K$ with

$$\int_\Omega \nabla u \cdot \nabla (u - v) \, dx \leq \int_\Omega f(u - v) \, dx \quad \text{for all } v \in K.$$

The FE discretisation replaces $K$ by the discrete version

$$K_h := K \cap P_k(T)$$

and hence determines $u_h \in K_h$ with

$$\int_\Omega \nabla u_h \cdot \nabla (u_h - v_h) \, dx \leq \int_\Omega f(u_h - v_h) \, dx \quad \text{for all } v_h \in K_h.$$ 

The main difference of the variational inequality and the model example of Section 3 can be expressed by means of residuals $\varrho \in H^{-1}(\Omega)$ and $\varrho_h \in P_k(T)$,

$$\varrho(v) := \int_\Omega v f \, dx - \int_\Omega \nabla u \cdot \nabla v \, dx \quad \text{for all } v \in H^1(\Omega),$$

$$\varrho_h := \sum_{z \in K} \left( \int_\Omega f \varphi_z \, dx - \int_\Omega \nabla u_h \cdot \nabla \varphi_z \, dx \right) \psi_z / \int_\Omega \varphi_z \, dx \in P_k(T).$$

It is elementary to verify that the error $e := u - u_h$ in the energy norm reads

$$|e|_1^2 = \int_\Omega f(e - e_h) \, dx - \int_\Omega \nabla u_h \cdot \nabla (e - e_h) \, dx + \int_\Omega \varrho_h e \, dx - \lambda g.$$
for some $e_h := \sum_{z \in {\mathcal N}_h} \left( \int_T e_h z^2(x) \, dx \right) \phi_z / \int_T \phi_z \, dx \in P_1(T)$. The approximation error $e - e_h$ can be analyzed as previously in the text. The additional terms with $\vartheta_h$ and $\varphi$ reflect the variational inequality. Indeed, one can show that $0 \leq \vartheta(e)$ and

$$\vartheta_h(z) \leq 0 = \vartheta_h(z) u_h(z) \leq u_h(z)$$

for all $z \in {\mathcal N}_T$.

Hence the arguments of Section 5 lead to the a posteriori error estimate

$$\|e\|_{1,2}^2 \leq C_{\eta_M}^2 + \text{h.o.t.}(f) - \int_{\Omega} \vartheta_h u_h \, dx.$$ 

The last term can be analyzed further. Indeed, $\vartheta_h u_h$ vanishes on an element $T \in T$ or, at least, $\vartheta_h(a) < 0 = u_h(a) = \vartheta_h(b) < u_h(b)$ for two nodes $a$ and $b$ of $T$. Inverse inequalities based on $\vartheta_h \leq 0 \leq u_h$ yield a bound of $\|\vartheta_h u_h\|_{L^1(T)}$ in terms of the mesh-size, $|\vartheta_h|_{1,2}$, and $\|\nabla u_h - A(\nabla u_h)\|_2$. The details and proofs can be found in [5]. The final result reads

$$\|e\|_{1,2} \leq C_{\text{rel}} \eta_M + \text{h.o.t.}(f).$$

For non-affine obstacles and nonconforming discretisations (i.e. $K_h \not\subset K$) some consistency terms arise and may dominate the upper bound, cf. [5]. Numerical results in [5] provide empirical evidence for a surprisingly high accuracy of $\eta_A$.

### 6.5. Degenerate Problems

The preceding examples concerned uniformly convex minimization problems on affine or convex subsets. The p-Laplacian is a first nonlinear equation with less strong convexity which requires a particular analysis. This is based on an appropriate quasi-norm, a metric that depends on the exact or discrete solution. The techniques of Section 5 and 6, however, can be adopted to this setting and then yield reliable and efficient error estimators [18]. The situation is even more difficult and essentially open for convexified problems where the energy minimization functional is not strictly convex. Very much as a surprise came numerical evidence in a 2-well benchmark example allowing for microstructures that $\eta_A$ yields an accurate stress error estimation [12].

### References


10. C. Carstensen, All first-order averaging techniques for a posteriori finite element error control on unstructured grids are efficient and reliable, Math. Comp. 73 (2004), no. 247, 1153–1165.


Department of Mathematics, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin, Germany
E-mail address: cc@mathematik.hu-berlin.de
E-mail address: jensens@mathematik.hu-berlin.de