SOME NOTES WITH NUMERICAL METHODS FOR STATIONARY PDES

ROB STEVENSON

1. Interpolation estimates in Sobolev spaces

Theorem 1.1 (Bramble-Hilbert "lemma")**.** *Let $\Omega \subset \mathbb{R}^n$ be a Lipschitz domain, and for some $m \in \mathbb{N}$, $q \in [1, \infty]$ and a normed space Y, let $L : W_q^m(\Omega) \to Y$ be a bounded linear mapping with $P_{m-1} \subset \mathrm{Ker} L$. Then $\exists C = C(\Omega)$ such that*

$$\|Lv\|_Y \leq C \|L\|_{W_q^m(\Omega) \to Y} |v|_{W_q^m(\Omega)} \quad (v \in W_q^m(\Omega)).$$

Lemma 1.2 (transformation lemma)**.** *Let $G(\hat{x}) = B\hat{x} + c$ with $\det B \neq 0$, and $\hat{\Omega}$ and Ω be Lipschitz domains in $\mathbb{R}^n$ with $G(\hat{\Omega}) = \Omega$.*
For $m \geq 0$, $p \in [1, \infty]$ and $v \in W_p^m(\Omega)$, $\hat{v} := v \circ G \in W_p^m(\hat{\Omega})$.
$\exists C = C(n, m, p)$ with

$$\begin{aligned} |\hat{v}|_{W_p^m(\hat{\Omega})} &\leq C \|B\|_2^m |\det B|^{-1/p} |v|_{W_p^m(\Omega)} \quad (v \in W_p^m(\Omega)), \\ |v|_{W_p^m(\Omega)} &\leq C \|B^{-1}\|_2^m |\det B|^{1/p} |\hat{v}|_{W_p^m(\hat{\Omega})} \quad (\hat{v} \in W_p^m(\hat{\Omega})). \end{aligned}$$

Theorem 1.3. *Let $\Omega, \hat{\Omega} \subset \mathbb{R}^n$ and G as in Lemma 1.2. Let*

$$\begin{aligned} h_\Omega &:= \inf\{\mathrm{diam}(S) : S \text{ ball containing } \Omega\} \\ \rho_\Omega &:= \sup\{\mathrm{diam}(S) : S \text{ ball in } \Omega\} \end{aligned}$$

and let $h_{\hat{\Omega}}$ and $\rho_{\hat{\Omega}}$ be defined similarly. Then $\|B\|_2 \leq \frac{h_\Omega}{\rho_{\hat{\Omega}}}$, $\|B^{-1}\|_2 \leq \frac{h_{\hat{\Omega}}}{\rho_\Omega}$.

Theorem 1.4. *Let $\Omega, \hat{\Omega} \subset \mathbb{R}^n$ and G as in Lemma 1.2.*
Let $k, m \in \mathbb{N}_0$ and $p, q \in [1, \infty]$ be such that $W_p^{k+1}(\hat{\Omega}) \hookrightarrow W_q^m(\hat{\Omega})$, and let $\hat{\Pi} : W_p^{k+1}(\hat{\Omega}) \to W_q^m(\hat{\Omega})$ be a bounded linear mapping that preserves polynomials of degree k.
Define Π by $\Pi(v) \circ G = \hat{\Pi}(v \circ G)$.
Then $\exists C = C(\hat{\Pi}, \hat{\Omega})$, thus independent of Ω, *such that*

$$|v - \Pi v|_{W_q^m(\Omega)} \leq C (\mathrm{vol}(\Omega))^{\frac{1}{q} - \frac{1}{p}} \frac{h_\Omega^{k+1}}{\rho_\Omega^m} |v|_{W_p^{k+1}(\Omega)} \quad (v \in W_p^{k+1}(\Omega)).$$

Date: May 17, 2016.

2. Application to estimate local interpolation errors

Theorem 2.1. *Let $(\hat{K}, \hat{P}, \hat{N})$ be a finite element with s denoting the maximal order of partial derivatives occurring in the definition of $\hat{N}$. For some $m, k \in \mathbb{N}_0$, $p, q \in [1, \infty]$, let*

$$\begin{aligned} &W_p^{k+1}(\hat{K}) \hookrightarrow C^s(\hat{K}) \\ &W_p^{k+1}(\hat{K}) \hookrightarrow W_q^m(\hat{K}) \\ &P_k(\hat{K}) \subset \hat{P} \subset W_q^m(\hat{K}) \end{aligned}$$

Then $\exists C = C(\hat{K}, \hat{P}, \hat{N})$ such that for all (K, P, N) that are affine interpolation equivalent to $(\hat{K}, \hat{P}, \hat{N})$,

$$|v - I_K v|_{W_q^m(K)} \leq C(\mathrm{vol}(K))^{\frac{1}{q}-\frac{1}{p}} \frac{h_K^{k+1}}{\rho_K^m} |v|_{W_p^{k+1}(K)} \quad (v \in W_p^{k+1}(K)).$$

Remark 2.2. Condition $W_p^{k+1}(\hat{K}) \hookrightarrow C^s(\hat{K})$ is imposed so that the interpolant $I_{\hat{K}}$ is a bounded mapping on $W_p^{k+1}(\hat{K})$.

Definition 2.3. A family of finite elements (K, P, N) is called *uniformly shape regular* when $\sup_K h_K/\rho_K < \infty$.

Corollary 2.4. *For a family of uniformly shape regular affine interpolation equivalent finite elements, result from Theorem 2.1 reads as*

$$|v - I_K v|_{W_q^m(K)} \leq C(\mathrm{vol}(K))^{\frac{1}{q}-\frac{1}{p}} h_K^{k+1-m} |v|_{W_p^{k+1}(K)} \quad (v \in W_p^{k+1}(K)).$$

3. Application to estimate global interpolation errors

Theorem 3.1. *Consider family $(\mathcal{T}_h)_h$ of subdivisions of a domain $\Omega \subset \mathbb{R}^n$ into element domains that are uniformly shape regular, and such that all finite elements are affine interpolation equivalent to a reference element $(\hat{K}, \hat{P}, \hat{N})$. Then under the conditions of Theorem 2.1 with $p = q$,*

(1)
$$\Big(\sum_{K \subset \mathcal{T}_h} h_K^{p(m-k-1)} \|v - I_K v\|_{W_p^m(K)}^p\Big)^{1/p} \lesssim |v|_{W_p^{k+1}(\Omega)} \quad (v \in W_p^{k+1}(\Omega)).$$

Define $I_{\mathcal{T}_h}$ by $(I_{\mathcal{T}_h} v)|_K := I_K v|_K$. Then if $\Im I_{\mathcal{T}_h} \subset C^{m-1}(\bar{\Omega})$, then with $h := \sup_{K \in \mathcal{T}_h} h_K$,

$$\|v - I_{\mathcal{T}_h} v\|_{W_p^m(\Omega)} \lesssim h^{k+1-m} |v|_{W_p^{k+1}(\Omega)} \quad (v \in W_p^{k+1}(\Omega)). \tag{2}$$

Remark 3.2. In these notes, by $C \lesssim D$ we will mean that C can be bounded on some absolute multiple of D, independently of parameters which C and D may depend on. Obviously, $C \gtrsim D$ is defined as $D \lesssim C$, and $C \eqsim D$ as $C \lesssim D$ and $C \gtrsim D$.

Remark 3.3. [homogeneous Dirichlet boundary conditions] In the situation of Theorem 3.1, if $\Im I_{\mathcal{T}_h} \subset C^0(\bar{\Omega})$, and $I_{\mathcal{T}_h}$ preserves the lowest order homogeneous Dirichlet boundary conditions, then $V_{\mathcal{T}_h,0} := \Im I_{\mathcal{T}_h}(H^{k+1}(\Omega) \cap H_0^1(\Omega)) \subset H_0^1(\Omega)$, and (2) for $m \in \{0,1\}$, $p = 2$ reads as

$$\|v - I_{\mathcal{T}_h} v\|_{H^m(\Omega)} \lesssim h^{k+1-m} |v|_{H^{k+1}(\Omega)} \quad (v \in H^{k+1}(\Omega) \cap H_0^1(\Omega)).$$

Using the Lax-Milgram lemma and Cea's lemma, we arrive at the following corollary.

Theorem 3.4. *Consider the situation of Theorem 3.1 with* $\Im I_{\mathcal{T}_h} \subset C^0(\bar{\Omega})$. *Let* $a : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ *be bilinear, bounded, coercive,* $F : H^1(\Omega) \to \mathbb{R}$ *linear and bounded. Let* $u \in H^1(\Omega)$, $u_{\mathcal{T}_h} \in V_{\mathcal{T}_h}$ *be the solutions of*

$$a(u,v) = F(v) \quad (v \in H^1(\Omega),$$
$$a(u_h, v_h) = F(v_h) \quad (v_h \in V_{\mathcal{T}_h}),$$

respectively. Then

$$\|u - u_h\|_{H^1(\Omega)} \lesssim h^k |u|_{H^{k+1}(\Omega)}$$

assuming $u \in H^{k+1}(\Omega)$.

Remark 3.5. Same conclusion when variational problem is formulated on $H_0^1(\Omega)$ and $V_{\mathcal{T}_h}$ reads as $V_{\mathcal{T}_h,0}$.

Under additional assumptions, higher order convergence can be demonstrated in the weaker $L_2(\Omega)$-norm:

Theorem 3.6 (Aubin-Nitsche duality 'trick')**.** *Let* $a(\,,\,)$ *be as in Thm 3.4. Suppose that for* $f \in L_2(\Omega)$, *the solution* $u_f \in H^1(\Omega)$ *(or in* $H_0^1(\Omega)$ *in case of hom. Dir.) of the adjoint problem* $a(v, u_f) = \int_\Omega f v dx$ $(v \in H^1(\Omega))$ $(H_0^1(\Omega))$ *is in* $H^2(\Omega)$ *with*

$$\|u_f\|_{H^2(\Omega)} \lesssim \|f\|_{L_2(\Omega)} \tag{3}$$

(this is known as a regularity condition*). Let* $(V_{\mathcal{T}_h})_h$ $((V_{\mathcal{T}_h,0})_h)$ *be such that*

(4)

$$\inf_{v_h \in V_{\mathcal{T}_h}} \|w - v_h\|_{H^1(\Omega)} \lesssim h \|w\|_{H^2(\Omega)} \text{ for all } w \in H^2(\Omega)\ (H^2(\Omega) \cap H_0^1(\Omega)).$$

Then for u *and* u_h *as in Thm 3.4, we have*

$$\|u - u_h\|_{L_2(\Omega)} \lesssim h \|u - u_h\|_{H^1(\Omega)}.$$

Proof. Let $w \in H^1(\Omega)$ $(H_0^1(\Omega))$ be the solution of the adjoint problem $a(v,w) = (u - u_h, v)_{L_2(\Omega)}$ $(v \in H^1(\Omega))$ $(H_0^1(\Omega))$. Then for any $w_h \in V_{\mathcal{T}_h}$ $(V_{\mathcal{T}_h,0})$,

$$\|u - u_h\|_{L_2(\Omega)}^2 = a(u - u_h, w) = a(u - u_h, w - w_h) \lesssim \|u - u_h\|_{H^1(\Omega)} \|w - w_h\|_{H^1(\Omega)}$$

Using that $\inf_{w_h} \|w - w_h\|_{H^1(\Omega)} \lesssim h \|w\|_{H^2(\Omega)} \lesssim h \|u - u_h\|_{L_2(\Omega)}$, the proof is completed. □

Example 3.7. If $\Omega \subset \mathbb{R}^2$ has a C^2 boundary or is convex, then for $f \in L_2(\Omega)$, the solution $u \in H_0^1(\Omega)$ of $\int_\Omega \nabla u \cdot \nabla v dx = \int_\Omega f v dx$ ($v \in H_0^1(\Omega)$) is in $H^2(\Omega)$ and satisfies $\|u\|_{H^2(\Omega)} \lesssim \|f\|_{L_2(\Omega)}$. (Without such conditions on Ω, this regularity result is generally *not* true).

4. Inverse inequality

Theorem 4.1. *Let $(V_{\mathcal{T}_h})_h$ be a family of affine equivalent f.e. spaces w.r.t. family $(\mathcal{T}_h)_h$ of uniformly shape regular subdivisions of $\Omega \subset \mathbb{R}^n$. Let $h_{\min} := \min_{K \in \mathcal{T}_h} \operatorname{diam}(K)$. Let $V_{\mathcal{T}_h} \subset W_p^m(\Omega)$. Then on $V_{\mathcal{T}_h}$,*

$$\| \cdot \|_{W_p^m(\Omega)} \lesssim h_{\min}^{-m} \| \cdot \|_{L_p(\Omega)}.$$

Proof. By the transformation lemma, equivalence of norms on finite dimensional spaces, and again the transformation lemma, for $v \in V_{\mathcal{T}_h}$ we have

$$\begin{aligned} |v|_{W_p^m(\Omega)}^p &= \sum_{K \in \mathcal{T}_h} |v|_K|_{W_p^m(K)}^p \lesssim \sum_{K \in \mathcal{T}_h} \|B^{-1}\|^{mp} |\det B| |\widehat{v|_K}|_{W_p^m(\hat{K})}^p \\ &\eqsim \sum_{K \in \mathcal{T}_h} \|B^{-1}\|^{mp} |\det B| \|\widehat{v|_K}\|_{L_p(\hat{K})}^p \lesssim \sum_{K \in \mathcal{T}_h} \|B^{-1}\|^{mp} \|v|_K\|_{L_p(K)}^p \\ &\lesssim \sum_{K \in \mathcal{T}_h} (\frac{\hat{h}}{\rho_K})^{mp} \|v|_K\|_{L_p(K)}^p \lesssim h_{\min}^{-pm} \|v\|_{L_p(\Omega)}^p. \end{aligned}$$

□

Literature with Sections 1–4: [Cia78]

5. Matrix-vector formulation of finite element discretization

Let V be some finite dimension subspace of some real Hilbert space H, let $a : V \times V \to \mathbb{R}$ be bilinear, bounded and coercive, and let $f : V \to \mathbb{R}$ be linear and bounded (e.g., a and f are restrictions to V of (bi)linear forms on H having those properties). We consider the problem of finding $u \in V$ s.t.

$$a(u, v) = f(v) \quad (v \in V) \tag{5}$$

Defining $A : V \to V'$ by $(Au)(v) = a(u, v)$ an equivalent formulation is given by

$$Au = f. \tag{6}$$

Let $\Phi = \{\phi_1, \dots, \phi_N\}$ be a basis for V. The corresponding dual basis $\Phi' = \{\phi_1', \dots, \phi_N'\}$ for V' is defined by $\phi_i'(\phi_j) = \delta_{ij}$.

Exercise -1. Let $\mathbf{A} \in \mathbb{R}^{N \times N}$ be defined by $\mathbf{A}_{ij} = a(\phi_j, \phi_i)$, called the *stiffness matrix.*

- Show that $\mathbf{A}$ is the representation of A w.r.t. primal and dual bases of V and V', respectively, i.e., if $v = \sum_j \mathbf{v}_j \phi_j$, then $Av = \sum_i (\mathbf{A}\mathbf{v})_i \phi_i'$. Conclude that an equivalent formulation of (5) or (6) is given by $\mathbf{A}\mathbf{u} = \mathbf{f}$, where $u = \sum_i \mathbf{u}_i \phi_i$, $f = \sum_i \mathbf{f}_i \phi_i'$.
- With $u = \sum_i \mathbf{u}_i \phi_i$, $v = \sum_i \mathbf{v}_i \phi_i$, $f = \sum_i \mathbf{f}_i \phi_i'$, i.e., $\mathbf{f}_i = f(\phi_i)$, and $\langle \cdot, \cdot \rangle$ the standard scalar product on $\mathbb{R}^N$, show that $\langle \mathbf{A}\mathbf{u}, \mathbf{v} \rangle = a(u, v)$ and $f(v) = \langle \mathbf{f}, \mathbf{v} \rangle$.

Unless stated otherwise, with the norm $\|\cdot\|$ on $\mathbb{R}^N$ (or on $\mathbb{R}^{N\times N}$) we will always mean the standard norm (or the corresponding operator norm).

Exercise 0.
- Show that $a(\cdot,\cdot)$ is symmetric iff $\mathbf{A} = \mathbf{A}^T$.
- Show that $a(v, v) > 0$ for all $0 \neq v \in V$ iff $\mathbf{A}$ is positive definite (denoted as $\mathbf{A} > 0$), i.e. $\langle \mathbf{A}\mathbf{v}, \mathbf{v} \rangle > 0$ for all $0 \neq \mathbf{v} \in \mathbb{R}^N$.

Remark 5.1. With the notations of Exercise -1, we have

$$\langle \mathbf{A}\mathbf{u}, \mathbf{v} \rangle = a(\sum_j \mathbf{u}_j \phi_j, \sum_i \mathbf{v}_i \phi_i) = \sum_{ij} \mathbf{u}_j \mathbf{v}_i \sum_K a(\phi_j|_K, \phi_i|_K).$$

The (set of non-zero entries of) the matrix $a(\phi_j|_K, \phi_i|_K)$ is known as the *element stiffness matrix.*

6. Conditioning of the stiffness matrix

Let $V \subset L_2(\Omega)$. $\mathbf{M} \in \mathbb{R}^{N\times N}$ defined by $\mathbf{M}_{ij} = (\phi_j, \phi_i)_{L_2(\Omega)}$ is called the *mass matrix.* Note that $\mathbf{M}$ is symmetric, positive definite.

Lemma 6.1. *If $\{\psi_1, \ldots, \psi_m\}$ is an independent set in a normed space $(V, \|\cdot\|)$, then $\|\sum_i \mathbf{c}_i \psi_i\|^2 \eqsim \sum_i |\mathbf{c}_i|^2$ (i.e. uniformly in $\mathbf{c} \in \mathbb{R}^m$).*

Proof. $\mathbf{c} \mapsto \|\sum_i \mathbf{c}_i \psi_i\|$ is continuous, so it attains a maximum and minimum on the unit ball in $\mathbb{R}^m$. By the independence of the set, the minimum is strictly positive. □

Theorem 6.2. *Let $(V_{\mathcal{T}_h})_h$ be a family of affine equivalent f.e. spaces w.r.t. a family of quasi-uniform, uniformly shape regular subdivisions of $\Omega \subset \mathbb{R}^n$. Then $\mathbf{M} = \mathbf{M}_h$ corresponding to the* nodal basis *is uniformly well-conditioned, i.e., $\sup_h \kappa(\mathbf{M}) < \infty$, where $\kappa(\mathbf{M}) = \|\mathbf{M}\|\|\mathbf{M}^{-1}\| = \frac{\rho(\mathbf{M}^\top\mathbf{M})^{\frac{1}{2}}}{\rho(\mathbf{M}^{-\top}\mathbf{M}^{-1})^{\frac{1}{2}}}$ is the* spectral condition number *of* $\mathbf{M}$.

Proof. By the choice of the basis, in the relation $v = \sum_i \mathbf{v}_i \phi_i$ we have $\mathbf{v}_i = N_i(v)$ where N_i denotes the ith global degree of freedom. With

$h>0$ such that $h \eqsim \min_{K\in\mathcal{T}_h} \operatorname{diam}(K) \eqsim \max_{K\in\mathcal{T}_h} \operatorname{diam}(K)$, we have

$$\begin{aligned}\langle \mathbf{M}\mathbf{v},\mathbf{v}\rangle = \|v\|^2_{L_2(\Omega)} &= \sum_K \|v|_K\|^2_{L_2(K)} = \sum_K |\det B| \|\widehat{v|_K}\|^2_{L_2(\hat{K})}\\ &\eqsim h^n \sum_K \|\widehat{v|_K}\|^2_{L_2(\hat{K})} \overset{\text{Lemma 6.1}}{\eqsim} h^n \sum_K \sum_j |\hat{N}^{\text{loc}}_j(\widehat{v|_K})|^2\\ &\overset{\text{affine eq.}}{=} h^n \sum_K \sum_j |N^{\text{loc}}_j(v|_K)|^2 \eqsim h^n \sum_i |N_i(v)|^2 = h^n\|\mathbf{v}\|^2. \square\end{aligned}$$

Theorem 6.3. *For $\Omega \subset \mathbb{R}^n$, let $(V_{\mathcal{T}_h})_h \subset H^m(\Omega)$ (or $\subset H^m_0(\Omega)$) be a family of f.e. spaces. Let $a(\cdot,\cdot): H^m(\Omega)\times H^m(\Omega) \to \mathbb{R}$ be bil., bound. and coercive (or with $H^m(\Omega)$ reading as $H^m_0(\Omega)$). Then the stiffness matrix $\mathbf{A} = \mathbf{A}_h$ w.r.t. a basis of $V_{\mathcal{T}_h}$ satisfies $\|\mathbf{A}\| \lesssim h_{\min}^{-2m}\|\mathbf{M}\|$ and $\|\mathbf{A}^{-1}\| \lesssim \|\mathbf{M}^{-1}\|$, with $\mathbf{M}$ being the corresponding mass matrix.*

Proof. Using Theorem 4.1, we have

$$\begin{aligned}|\langle \mathbf{A}\mathbf{v},\mathbf{w}\rangle| = |a(v,w)| &\lesssim \|v\|_{H^m(\Omega)}\|w\|_{H^m(\Omega)} \lesssim h_{\min}^{-2m}\|v\|_{L_2(\Omega)}\|w\|_{L_2(\Omega)}\\ &\lesssim h_{\min}^{-2m}\lambda_{\max}(\mathbf{M})\|\mathbf{v}\|\|\mathbf{w}\|,\end{aligned}$$

or $\|\mathbf{A}\| \lesssim h_{\min}^{-2m}\lambda_{\max}(\mathbf{M})$. On the other hand

$$\langle \mathbf{A}\mathbf{v},\mathbf{v}\rangle \gtrsim \|v\|^2_{H^m(\Omega)} \geq \|v\|^2_{L_2(\Omega)} \gtrsim \lambda_{\min}(\mathbf{M})\|\mathbf{v}\|^2,$$

and so

$$\|\mathbf{A}^{-1}\mathbf{v}\|^2 \lesssim \lambda_{\min}(\mathbf{M})^{-1}\langle \mathbf{v}, \mathbf{A}^{-1}\mathbf{v}\rangle \leq \lambda_{\min}(\mathbf{M})^{-1}\|\mathbf{v}\|\|\mathbf{A}^{-1}\mathbf{v}\|$$

or $\|\mathbf{A}^{-1}\mathbf{v}\| \lesssim \lambda_{\min}(\mathbf{M})^{-1}\|\mathbf{v}\|$ or $\|\mathbf{A}^{-1}\| \lesssim \lambda_{\min}(\mathbf{M})^{-1}$. $\square$

Remark 6.4. If the basis in Theorem 6.3 is the *nodal* basis, then under the conditions of Theorem 6.2 we have $\kappa(\mathbf{A}) \lesssim h_{\min}^{-2m}$. Generally, this estimate is sharp.

7. A posteriori error estimation

For *simplicity*: Poisson on a polytopal domain Ω, usually in $n=2$ dimensions, homogeneous Dirichlet boundary conditions.

$\mathcal{T}$ is a uniformly shape regular, conforming partition into n-simplices. $\mathcal{S}_\mathcal{T}$ is Lagrange f.e. space of degree k. $\mathcal{E}(\mathcal{T})$ is the set of the interior edges of $\mathcal{T}$.

For $T\in\mathcal{T}$, $v\in\mathcal{S}_\mathcal{T}$, $f\in L_2(\Omega)$, the (squared) error indicator for v on T reads as

$$\eta(v,T)^2 := h_T^2\|f+\Delta v\|^2_{L_2(T)} + h_T\|[\![\nabla v]\!]\|^2_{L_2(\partial T\setminus\partial\Omega)},$$

where $[\![\nabla v]\!]$ is jump of normal derivative of v over interface, $h_T := |T|^{1/n}$.

The (squared) oscillation of f on T is defined as

$$\operatorname{osc}(f,T)^2 := h_T^2\|f-P_T^r f\|^2_{L_2(T)},$$

where, for some fixed $\mathbb{N}_0 \ni r \geq k-2$, P_T^r is the $L_2(T)$-orthogonal projector onto $\mathcal{P}_r(T)$.

Note that $\operatorname{osc}(f,T)^2 \leq \eta(v,T)^2$, because $P_T^r \Delta v = \Delta v$. (Usually $\sum_{T\in\mathcal{T}} \operatorname{osc}(f,T)^2 \ll \sum_{T\in\mathcal{T}} \eta(v,T)^2$, cf. Example 11.3).

For $\mathcal{M} \subset \mathcal{T}$,

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

$\mathcal{T} \leq \tilde{\mathcal{T}}$ means that $\tilde{\mathcal{T}}$ is a refinement of $\mathcal{T}$. $R_{\mathcal{T}\to\tilde{\mathcal{T}}} := \mathcal{T} \setminus \tilde{\mathcal{T}}$, i.e., the set of those $T \in \mathcal{T}$ that were refined when passing from $\mathcal{T}$ to $\tilde{\mathcal{T}}$.

$u_{\mathcal{T}}$ will denote the Galerkin solution from $\mathcal{S}_{\mathcal{T}}$.

Theorem 7.1 (local upper bound provided by error estimator)**.** *For $\mathcal{T} \leq \tilde{\mathcal{T}}$, it holds that*

$$|u_{\tilde{\mathcal{T}}} - u_{\mathcal{T}}|^2_{H^1(\Omega)} \lesssim \eta(u_{\mathcal{T}}, R_{\mathcal{T}\to\tilde{\mathcal{T}}})^2.$$

In particular

$$|u - u_{\mathcal{T}}|^2_{H^1(\Omega)} \lesssim \eta(u_{\mathcal{T}}, \mathcal{T})^2.$$

Proof. It holds that

$$|u_{\tilde{\mathcal{T}}} - u_{\mathcal{T}}|_{H^1(\Omega)} = \sup_{0\neq w_{\tilde{\mathcal{T}}}\in\mathcal{S}_{\tilde{\mathcal{T}}}} \frac{a(u_{\tilde{\mathcal{T}}} - u_{\mathcal{T}}, w_{\tilde{\mathcal{T}}})}{|w_{\tilde{\mathcal{T}}}|_{H^1(\Omega)}}. \tag{7}$$

For any $w_{\mathcal{T}} \in \mathcal{S}_{\mathcal{T}}$, we have

$$\begin{aligned}
&a(u_{\tilde{\mathcal{T}}} - u_{\mathcal{T}}, w_{\tilde{\mathcal{T}}}) = a(u_{\tilde{\mathcal{T}}} - u_{\mathcal{T}}, w_{\tilde{\mathcal{T}}} - w_{\mathcal{T}})\\
&= \int_\Omega f(w_{\tilde{\mathcal{T}}} - w_{\mathcal{T}})\,dx - a(u_{\mathcal{T}}, w_{\tilde{\mathcal{T}}} - w_{\mathcal{T}})\\
&= \sum_{T\in\mathcal{T}} \Big\{ \int_T f(w_{\tilde{\mathcal{T}}} - w_{\mathcal{T}})\,dx - \int_T \nabla u_{\mathcal{T}} \cdot \nabla(w_{\tilde{\mathcal{T}}} - w_{\mathcal{T}})\Big\}\\
&= \sum_{T\in\mathcal{T}} \Big\{ (\int_T f + \Delta u_{\mathcal{T}})(w_{\tilde{\mathcal{T}}} - w_{\mathcal{T}})\,dx - \int_{\partial T} \nabla u_{\mathcal{T}} \cdot \mathbf{n}(w_{\tilde{\mathcal{T}}} - w_{\mathcal{T}})\Big\} \qquad (8)\\
&\leq \sum_{T\in\mathcal{T}} \|f + \Delta u_{\mathcal{T}}\|_{L_2(T)} \|w_{\tilde{\mathcal{T}}} - w_{\mathcal{T}}\|_{L_2(T)}\\
&\qquad + \sum_{e\in\mathcal{E}(\mathcal{T})} \|[\![\nabla u_{\mathcal{T}}]\!]\|_{L_2(e)} \|w_{\tilde{\mathcal{T}}} - w_{\mathcal{T}}\|_{L_2(e)}.
\end{aligned}$$

Select $w_{\mathcal{T}}$ to be the Scott-Zhang interpolant of $w_{\tilde{\mathcal{T}}}$ as follows: If vertex $\nu \in T \notin R_{\mathcal{T}\to\tilde{\mathcal{T}}}$, select SZ edge on T, so that $w_{\mathcal{T}}(\nu) = w_{\tilde{\mathcal{T}}}(\nu)$. So $w_{\mathcal{T}} = w_{\tilde{\mathcal{T}}}$ on all $T \notin R_{\mathcal{T}\to\tilde{\mathcal{T}}}$, and consequently on all edges of those T.

For the remaining $T \in \mathcal{T}$ and edges $e \in \mathcal{E}(\mathcal{T})$, use that

$$h_T^{-1}\|w_{\tilde{\mathcal{T}}} - w_{\mathcal{T}}\|_{L_2(T)} + |w_{\tilde{\mathcal{T}}} - w_{\mathcal{T}}|_{H^1(T)} \lesssim |w_{\tilde{\mathcal{T}}}|_{H^1(S(\mathcal{T},T))}, \tag{9}$$

with patch $S(\mathcal{T},T) := \{T' \in \mathcal{T} : T \cap T' \neq \emptyset\}$, as well as

$$\|g\|_{L_2(e)} \lesssim h_T^{-1/2}\|g\|_{L_2(T)} + h_T^{1/2}|g|_{H^1(T)}$$

for $T \in \mathcal{T}$ such that e is an edge of T, which yields, using (9) again,

$$\|w_{\tilde{\mathcal{T}}} - w_{\mathcal{T}}\|_{L_2(e)} \lesssim h_T^{1/2} |w_{\tilde{\mathcal{T}}}|_{H^1(S(\mathcal{T},T))}. \tag{10}$$

By combining (8) with (9) and (10), applying Cauchy-Schwarz, the proof is completed by (7). $\square$

Theorem 7.2 (global lower bound provided by error estimator)**.**

$$\eta(u_{\mathcal{T}}, \mathcal{T})^2 \lesssim |u - u_{\mathcal{T}}|^2_{H^1(\Omega)} + \operatorname{osc}(f, \mathcal{T})^2.$$

(Actually holds true for $u_{\mathcal{T}}$ reading as any function in $\mathcal{S}_{\mathcal{T}}$.)

As a consequence of Thm. 7.1 and 7.2, we have that the '*total error*' –defined as the square root of squared error plus squared oscillation– is proportional to the estimator:

Corollary 7.3. $|u - u_{\mathcal{T}}|^2_{H^1(\Omega)} + \operatorname{osc}(f, \mathcal{T})^2 \eqsim \eta(u_{\mathcal{T}}, \mathcal{T})^2$.

Proof of Thm. 7.2. For $v \in H^1_0(\Omega)$, we have

$$a(u - u_{\mathcal{T}}, v) = \sum_{T \in \mathcal{T}} \Big[\int_T (f + \Delta u_{\mathcal{T}}) v - \int_{\partial T} (\nabla u_{\mathcal{T}} \cdot \mathbf{n}) v \Big]. \tag{11}$$

Fixing $T \in \mathcal{T}$, for $v \in H^1_0(T)$, with $\bar{f}_T := P^r_T f$ and using $(P^r_T - I)P^r_T = 0$, we have that

$$\begin{aligned} \Big| \int_T (\bar{f}_T + \Delta u_{\mathcal{T}}) v \Big| &= \Big| a(u - u_{\mathcal{T}}, v) + \int_T (\bar{f}_T - f)(I - P^r_T) v \Big| \\ &\lesssim |u - u_{\mathcal{T}}|_{H^1(T)} |v|_{H^1(T)} + h_T \|f - \bar{f}_T\|_{L_2(T)} |v|_{H^1(T)}, \end{aligned}$$

or

$$\sup_{0 \neq v \in H^1_0(T)} \frac{|\int_T (\bar{f}_T + \Delta u_{\mathcal{T}}) v|}{|v|_{H^1(T)}} \lesssim |u - u_{\mathcal{T}}|_{H^1(T)} + \operatorname{osc}(f, T).$$

From

$$h_T \|p\|_{L_2(T)} \lesssim \sup_{0 \neq v \in H^1_0(T)} \frac{|\int_T p v \, dx|}{|v|_{H^1(T)}} \quad (p \in \mathcal{P}_r(T))$$

([BS08, 9.x.5]), we obtain

$$\begin{aligned} h_T \|f + \Delta u_{\mathcal{T}}\|_{L_2(T)} &\leq h_T \|\bar{f}_T + \Delta u_{\mathcal{T}}\|_{L_2(T)} + \operatorname{osc}(f, T) \\ &\lesssim |u - u_{\mathcal{T}}|_{H^1(T)} + \operatorname{osc}(f, T). \end{aligned} \tag{12}$$

For $e \in \mathcal{E}(\mathcal{T})$, $e = T_1 \cap T_2$, and $v \in V_e := \{w \in H^1_0(T_1 \cup T_2) \colon \int_{T_i} w \mathcal{P}_r = 0\}$, from (11) and $(P^r_T - I)P^r_T = 0$, we infer

$$\begin{aligned} \Big| \int_e [\![\nabla u_{\mathcal{T}}]\!] v \, ds \Big| &= \Big| a(u - u_{\mathcal{T}}, v) + \sum_{i=1}^2 \int_{T_i} (\bar{f}_{T_i} - f)(I - P^r_{T_i}) v \Big| \\ &\lesssim |u - u_{\mathcal{T}}|_{H^1(T_1 \cup T_2)} |v|_{H^1(T_1 \cup T_2)} + \sqrt{\sum_{i=1}^2 h^2_{T_i} \|f - \bar{f}_{T_i}\|^2_{L_2(T_i)}} |v|_{H^1(T_1 \cup T_2)}. \end{aligned}$$

From

$$h_e^{\frac{1}{2}} \|p\|_{L_2(e)} \lesssim \sup_{0 \neq v \in V_e} \frac{|\int_e pv\,dx|}{|v|_{H^1(T_1 \cup T_2)}} \qquad (p \in \mathcal{P}_k)$$

([BS08, 9.x.7], where $h_e := |e|^{1/(n-1)}$), we obtain

$$h_e^{\frac{1}{2}} \| [\![\nabla u_{\mathcal{T}}]\!] \|_{L_2(e)} \lesssim |u - u_{\mathcal{T}}|_{H^1(T_1 \cup T_2)} + \sqrt{\operatorname{osc}(f, T_1)^2 + \operatorname{osc}(f, T_2)^2}. \tag{13}$$

By summing (12) over $T \in \mathcal{T}$, and (13) over $e \in \mathcal{E}(\mathcal{T})$, the proof is completed. □

Literature with this section: [Ver96, Ste07].

8. Newest vertex bisection

The newest vertex bisection algorithm reads as follows:

- In each triangle in an initial, *conforming* partition $\mathcal{T}_0$ of a polygon Ω into triangles, call one of its vertices its *newest vertex*.
- If you want to refine a triangle T in a partition, then connect its newest vertex with the midpoint of opposite edge (the *refinement edge* of T). This midpoint will be the newest vertex of the two triangles being created.

All partitions $\mathcal{T}$ that can be created in this way can be represented as a subtree (being a subset that contains the roots, and for any other element that it contains, it contains its parent and its sibling) of an infinite binary tree (the *master tree*) that has as its roots the triangles from $\mathcal{T}_0$.

For any triangle T in the master tree, $\operatorname{gen}(T)$ is defined as the number of bisections that are needed to create it starting from a root.

The partitions $\mathcal{T}$ that can be created in this way are *uniformly shape regular* (exercise).

To restrict ourselves to the subset of partitions $\mathcal{T}$ that additionally are *conforming*, consider the following procedure to refine a triangle T in a conforming partition $\mathcal{T}$:

`refine`$(T, \mathcal{T})$
% T is triangle in conforming partition $\mathcal{T}$
`if` the neighboring triangle T' at other side of refinement edge of
    T has a different refinement edge
`then` `refine`$(T', \mathcal{T})$
`endif`
simultaneously bisect T and T' in $\mathcal{T}$.

This algorithm may not terminate, see Figure 8. To avoid such a deadlock situation, we impose a *matching condition* on the initial assignment of the newest vertices: If $e = T \cap T'$ is the refinement edge of T, then it is the refinement edge of T'.

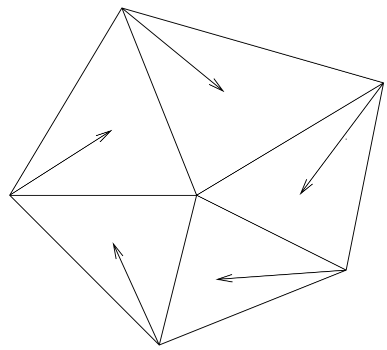

FIGURE 1. Deadlock situation. The arrows indicate the newest vertices

Theorem 8.1 ([BDD04]). *For any conforming triangulation $\mathcal{T}_0$, there exists an assignment of the newest vertices such that the matching condition is satisfied.*

The proof this theorem is not easy, and what is worse, it is not constructive. As an alternative, one may perform an initial refinement of $\mathcal{T}_0$ that yields a triangulation on which a suitable initial assignment of the newest vertices can easily be found, cf. Figure 8.

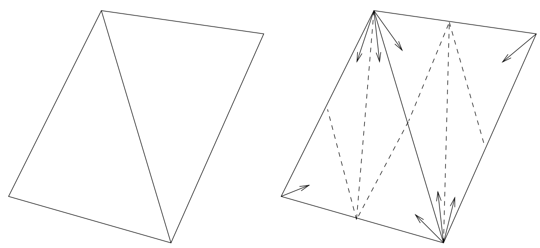

FIGURE 2. A refinement of a given $\mathcal{T}_0$, and a valid assignment of newest vertices in the resulting triangulation.

Theorem 8.2. *Let $\mathcal{T}_0$ be a conforming initial partition that satisfies the matching condition, and let $\mathcal{T}$ denote any partition that is created from $\mathcal{T}_0$ by newest vertex bisection. Then*

(1) *if $\mathcal{T}$ is a uniform refinement of $\mathcal{T}_0$ (meaning that all its triangles have the same generation), then it is conforming.*

(2) *If $\mathcal{T}$ be conforming, $T, T' \in \mathcal{T}$, and T' contains the refinement edge of T, then either*
   - $\mathrm{gen}(T') = \mathrm{gen}(T)$, *and T and T' share their refinement edge, or*
   - $\mathrm{gen}(T') = \mathrm{gen}(T) - 1$, *and T shares its refinement edge with one of both children of T'.*

(3) `refine`$(\mathcal{T}, T)$ *terminates.*

Proof. Exercise. □

From here on, $\mathcal{T}$ will always denote a *conforming* partition that can be created by newest vertex bisection from a conforming initial partition that satisfies the matching condition. The set of all these partitions will be denoted as $\mathbb{T}$.

Lemma 8.3. *If $\mathcal{T}, \mathcal{T}' \in \mathbb{T}$, then the smallest common refinement $\mathcal{T} \oplus \mathcal{T}'$ is in $\mathbb{T}$, and $\#\mathcal{T} \oplus \mathcal{T}' + \#\mathcal{T}_0 \leq \#\mathcal{T} + \#\mathcal{T}'$.*

Proof. Exercise. Hint: do it first for one root, i.e. $\#\mathcal{T}_0 = 1$. □

9. The adaptive finite element method (AFEM)

% Let $\theta \in (0, 1]$ be some parameter

$\texttt{For } k = 0, 1, \ldots, \texttt{ do}$

$$\left.\text{solve } u_k \in \mathcal{S}_{\mathcal{T}_k} \text{ from } a(u_k, v_k) = f(v_k) \ (v_k \in \mathcal{S}_{\mathcal{T}_k})\ \right\} \texttt{SOLVE}$$

$$\left.\begin{array}{l} \text{compute } \{\eta(u_k, T) \colon T \in \mathcal{T}_k\} \\ \texttt{if } \eta(u_k, \mathcal{T}_k) \leq \text{TOL } \texttt{then break endif} \end{array}\right\} \texttt{ESTIMATE}$$

$$\left.\begin{array}{l} \text{select a } \textit{smallest } \mathcal{M}_k \subset \mathcal{T}_k \text{ such that} \\ \qquad\qquad \eta(u_k, \mathcal{M}_k) \geq \theta \eta(u_k, \mathcal{T}_k) \end{array}\right\} \texttt{MARK}$$

$$\left.\begin{array}{l} \texttt{while } \mathcal{T}_k \cap \mathcal{M}_k \neq \emptyset \texttt{ do} \\ \qquad \text{for some } T \in \mathcal{T}_k \cap \mathcal{M}_k,\ \mathcal{T}_k := \texttt{refine}(T, \mathcal{T}_k) \\ \texttt{endwhile} \\ \mathcal{T}_{k+1} := \mathcal{T}_k \end{array}\right\} \texttt{REFINE}$$

`endfor`

The marking strategy is known as *bulk chasing*, and also, after its inventor, as *Dörfler* marking. In `REFINE`, the smallest $\mathbb{T} \ni \mathcal{T}_{k+1} \geq \mathcal{T}_k$ is determined in which all $T \in \mathcal{M}_k$ have been bisected.

10. AFEM is linearly convergent

In this and the next section, let $(\mathcal{T}_k)_{k \geq 0}$, $(u_k)_{k \geq 0}$, and $(\mathcal{M}_k)_{k \geq 0}$ be as produced by `AFEM`.

Theorem 10.1. *$\exists$ constants $\gamma > 0$, $\alpha \in (0, 1)$, such that*

$$|u - u_{k+1}|^2_{H^1(\Omega)} + \gamma \eta(u_{k+1}, \mathcal{T}_{k+1})^2 \leq \alpha(|u - u_k|^2_{H^1(\Omega)} + \gamma \eta(u_k, \mathcal{T}_k)^2).$$

To prove this theorem, first we give two lemmas.

Lemma 10.2. *For $v, w \in \mathcal{S}_{\mathcal{T}}$, $T \in \mathcal{T}$, we have*

$$|\eta(v, T) - \eta(w, T)| \lesssim \|v - w\|_{H^1(S(\mathcal{T}, T))}.$$

Proof. Recall $\eta(z, T)^2 := h_T^2 \|f + \Delta z\|^2_{L_2(T)} + h_T \|[\![\nabla z]\!]\|^2_{L_2(\partial T \setminus \partial\Omega)}$. Now use that $\sqrt{a^2 + b^2} - \sqrt{\tilde{a}^2 + \tilde{b}^2} \leq \sqrt{(a - \tilde{a})^2 + (b - \tilde{b})^2}$, and $|\|\cdot\| - \|\cdot\||^2 \leq \|\cdot - \cdot\|^2$. So

$$|\eta(v, T) - \eta(w, T)|^2 \leq h_T^2 \|\Delta(v - w)\|^2_{L_2(T)} + h_T \|[\![\nabla(v - w)]\!]\|^2_{L_2(\partial T \setminus \partial\Omega)}.$$

Now use that for $z \in \mathcal{P}_k$, $\|\Delta z\|_{L_2(T)} \lesssim h_T^{-1}\|z\|_{H^1(T)}$, and $\|\nabla z\|_{L_2(e)^n} \lesssim h_{T'}^{-\frac{1}{2}}\|z\|^2_{H^1(T')}$, when e is an edge of $T' \in \mathcal{T}$. $\square$

Lemma 10.3. *$\exists$ constant Λ such that for any $\delta > 0$, and with $\lambda := 1 - 2^{-1/n}$,*

$$\eta(u_{k+1}, \mathcal{T}_{k+1})^2 \leq (1+\delta)(\eta(u_k, \mathcal{T}_k)^2 - \lambda\eta(u_k, \mathcal{M}_k)^2) + (1+\delta^{-1})\Lambda|u_{k+1} - u_k|^2_{H^1(\Omega)}.$$

Proof. The previous lemma shows that, for some constant $C > 0$, for $T \in \mathcal{T}_{k+1}$,

$$\eta(u_{k+1}, T) \leq \eta(u_k, T) + C\|u_{k+1} - u_k\|_{H^1(S(\mathcal{T},T))}.$$

We apply Young's inequality $(a + b)^2 \leq (1 + \delta)a^2 + (1 + \delta^{-1})b^2$ (from $(\sqrt{\delta}a + \frac{1}{\sqrt{\delta}}b)^2 \geq 0$), sum over $T \in \mathcal{T}_{k+1}$, use $\|\cdot\|_{H^1(\Omega)} \eqsim |\cdot|_{H^1(\Omega)}$ on $H^1_0(\Omega)$, to arrive at

$$\eta(u_{k+1}, \mathcal{T}_{k+1})^2 \leq (1 + \delta)\eta(u_k, \mathcal{T}_{k+1})^2 + (1 + \delta^{-1})\Lambda|u_{k+1} - u_k|^2_{H^1(\Omega)}$$

for some constant $\Lambda > 0$.

Any $T \in \mathcal{M}_k$ is split into 2 or more triangles. Let us consider the most unfortunate situation that it is split into two triangles, T_1 and T_2. From $h_{T_i} = \frac{1}{2}\sqrt{2}\,h_T$, we have $\sum_{i=1,2}\eta(u_k, T_i)^2 \leq \frac{1}{2}\sqrt{2}\,\eta(u_k, T)^2$. We conclude that

$$\begin{aligned}\eta(u_k, \mathcal{T}_{k+1})^2 &\leq \eta(u_k, \mathcal{T}_k \setminus \mathcal{M}_k)^2 + \frac{1}{2}\sqrt{2}\,\eta(u_k, \mathcal{M}_k)^2 \\ &= \eta(u_k, \mathcal{T}_k)^2 - (1 - \frac{1}{2}\sqrt{2})\eta(u_k, \mathcal{M}_k)^2,\end{aligned}$$

which completes the proof (for $n = 2$). $\square$

Proof of Thm. 10.1. From $u - u_{k+1} \perp_{\langle\nabla\cdot,\nabla\cdot\rangle_{L_2(\Omega)}} \mathcal{S}_{\mathcal{T}_{k+1}}$, and $u_{k+1} - u_k \in \mathcal{S}_{\mathcal{T}_{k+1}}$, we have

$$|u - u_{k+1}|^2_{H^1(\Omega)} = |u - u_k|^2_{H^1(\Omega)} - |u_{k+1} - u_k|^2_{H^1(\Omega)}.$$

From the previous lemma and the marking procedure, which yields $\eta(u_k, \mathcal{M}_k) \geq \theta\eta(u_k, \mathcal{T}_k)$, we have

$$\eta(u_{k+1}, \mathcal{T}_{k+1})^2 \leq (1+\delta)(1-\lambda\theta^2)\eta(u_k, \mathcal{T}_k)^2 + (1+\delta^{-1})\Lambda|u_{k+1} - u_k|^2_{H^1(\Omega)}.$$

By choosing δ such that $(1+\delta)(1-\lambda\theta^2) = 1-\lambda\theta^2/2$, and by multiplying the second estimate with γ, choosing γ such that $\gamma(1 + \delta^{-1})\Lambda = 1$, and by adding both estimates, we infer that

$$\begin{aligned}|u - u_{k+1}|^2_{H^1(\Omega)} + \gamma\eta(u_{k+1}, \mathcal{T}_{k+1})^2 &\leq |u - u_k|^2_{H^1(\Omega)} + \gamma(1 - \lambda\theta^2/2)\eta(u_k, \mathcal{T}_k)^2 \\ &\leq (1 - \frac{\lambda\theta^2/2}{1 + C/\gamma})(|u - u_k|^2_{H^1(\Omega)} + \gamma\eta(u_k, \mathcal{T}_k)^2)\end{aligned}$$

with $C > 0$ such that $|u - u_k|^2_{H^1(\Omega)} \leq C\eta(u_k, \mathcal{T}_k)^2$ (Thm. 7.1). $\square$

Literature with this section: [Dör96, MNS00, MN05].

11. AFEM CONVERGES WITH THE BEST POSSIBLE RATE

Definition 11.1. For $s > 0$, we define the approximation class

$$\mathcal{A}^s := \{u \in H_0^1(\Omega)\colon \Delta u \in L_2(\Omega),$$
$$|u|_{\mathcal{A}^s} := \sup_{N\in\mathbb{N}}(N+1)^s \min_{\{\mathcal{T}\in\mathbb{T}\colon \#\mathcal{T}-\#\mathcal{T}_0\leq N\}} \sqrt{|u-u_{\mathcal{T}}|^2_{H^1(\Omega)} + \operatorname{osc}(f,\mathcal{T})^2} < \infty\}.$$

So $u \in \mathcal{A}^s$ means that for a *best* partition with $N + \#\mathcal{T}_0$ triangles, the total error in the Galerkin approximation is $\leq (N+1)^{-s}|u|_{\mathcal{A}^s}$.

Remark 11.2. If $u \in \mathcal{A}^s$, then for any $\varepsilon > 0$, $\exists\mathcal{T} \in \mathbb{T}$ that realizes a total error $\leq \varepsilon$ where $\#\mathcal{T} - \#\mathcal{T}_0 \leq \varepsilon^{-1/s}|u|_{\mathcal{A}^s}^{1/s}$. Indeed, denoting with $e(N)$ the total error in a best partition with $N + \#\mathcal{T}_0$ triangles, let N be such that $e(N) \leq \varepsilon \leq e(N-1)$. Then $\varepsilon N^s \leq N^s e(N-1) \leq |u|_{\mathcal{A}^s}$.

Example 11.3. If u is smooth –sufficient is $u \in H^{k+1}(\Omega)\cap H_0^1(\Omega)$–, take $\mathcal{T}$ to be a (quasi-) uniform mesh with mesh-size h. Then $|u-u_{\mathcal{T}}|_{H^1(\Omega)} \lesssim h^k|u|_{H^{k+1}(\Omega)}$. Assuming that even $u \in H^{k+2}(\Omega)$, then $f \in H^k(\Omega)$, and by taking $r \geq k-1$, one infers that $\operatorname{osc}(f,\mathcal{T}) \lesssim h^{k+1}|f|_{H^k(\Omega)} \lesssim h^{k+1}|u|_{H^{k+2}(\Omega)}$ (so the oscillation is of higher order). Since $N := \#\mathcal{T} - \#\mathcal{T}_0 \eqsim (h^{-1})^n$, we have that $h^k \eqsim N^{-k/n}$, i.e., $s = k/n$ is the best possible convergence order that generally can be expected. In other words, for $s > k/n$, the class $\mathcal{A}^s$ is basically empty.

On the other hand, for $s \leq k/n$, the class $\mathcal{A}^s$ is *much* bigger than $H_0^1(\Omega)\cap H^{1+sn}(\Omega)$. As shown in [BDDP02], it contains $H_0^1(\Omega)\cap W_p^{1+sn}(\Omega)$ whenever $p > (s+\frac{1}{2})^{-1}$. These spaces $W_p^{1+sn}(\Omega)$ are only just embedded in $H^1(\Omega)$.

For the Poisson problem on a two-dimensional polygon, in [DD97] it was shown that for *any* given $s > 0$, for sufficiently smooth right-hand side f, the solution $u \in H_0^1(\Omega) \cap W_p^{1+sn}(\Omega)$ for some $p > (s+\frac{1}{2})^{-1}$.

The following result about newest vertex bisection will be an essential ingredient in the optimality proof.

Theorem 11.4 ([BDD04]; [Ste08] for a generalization to $n > 2$.)**.** *Let $(\mathcal{T}_i)_i \subset \mathbb{T}$ be such that $\mathcal{T}_{i+1}$ is the smallest refinement in $\mathbb{T}$ of $\mathcal{T}_i$ in which all triangles from some subset $\mathcal{M}_i \subset \mathcal{T}_i$ have been bisected. Then*

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

Note that in contrast, $\frac{\#\mathcal{T}_{i+1}}{\#\mathcal{T}_i+\#\mathcal{M}_i}$ can be arbitrarily large.

Lemma 11.5. *Let $C_1, C_2 > 0$ be constants such that for $\mathcal{T} \leq \tilde{\mathcal{T}} \in \mathbb{T}$,*

$$\eta(u_{\mathcal{T}},\mathcal{T})^2 \leq C_1[|u-u_{\mathcal{T}}|^2_{H^1(\Omega)} + \operatorname{osc}(f,\mathcal{T})^2],$$
$$|u_{\tilde{\mathcal{T}}} - u_{\mathcal{T}}|^2_{H^1(\Omega)} \leq C_2\eta(u_{\mathcal{T}}, R_{\mathcal{T}\to\tilde{\mathcal{T}}})^2,$$

see Thms. 7.2 and 7.1. Let the marking parameter be sufficiently small such that $\theta^2 < (C_1(C_2+1))^{-1}$. Then for $\mathbb{T} \ni \mathcal{T} \geq \mathcal{T}_k$ with

$$|u-u_{\mathcal{T}}|^2_{H^1(\Omega)}+\operatorname{osc}(\mathcal{T},f)^2 \leq [1-\theta^2 C_1(C_2+1)][|u-u_k|^2_{H^1(\Omega)}+\operatorname{osc}(\mathcal{T}_k,f)^2],$$

it holds that

$$\eta(u_k, R_{\mathcal{T}_k\to\mathcal{T}}) \geq \theta\eta(u_k,\mathcal{T}_k)$$

(*and so* $\#\mathcal{M}_k \leq \#R_{\mathcal{T}_k\to\mathcal{T}}$ (!)).

Proof. It holds that

$$\begin{aligned}|u-u_k|^2_{H^1(\Omega)} &= |u-u_{\mathcal{T}}|^2_{H^1(\Omega)} + |u_{\mathcal{T}}-u_k|^2_{H^1(\Omega)},\\ \operatorname{osc}(\mathcal{T}_k,f)^2 &\leq \operatorname{osc}(R_{\mathcal{T}_k\to\mathcal{T}},f)^2+\operatorname{osc}(\mathcal{T},f)^2,\end{aligned}$$

which yields

$$\begin{aligned}\theta^2(C_2+1)\eta(u_k,\mathcal{T}_k)^2 &\leq \theta^2 C_1(C_2+1)(|u-u_k|^2_{H^1(\Omega)}+\operatorname{osc}(\mathcal{T}_k,f)^2)\\ &\leq |u-u_k|^2_{H^1(\Omega)}+\operatorname{osc}(\mathcal{T}_k,f)^2-|u-u_{\mathcal{T}}|^2_{H^1(\Omega)}-\operatorname{osc}(\mathcal{T},f)^2\\ &\leq |u_{\mathcal{T}}-u_k|^2_{H^1(\Omega)}+\operatorname{osc}(R_{\mathcal{T}_k\to\mathcal{T}},f)^2\\ &\leq (C_2+1)\eta(u_k,R_{\mathcal{T}_k\to\mathcal{T}})^2.\end{aligned}$$ □

Corollary 11.6. *Let $\theta^2 < (C_1(C_2+1))^{-1}$. For some $s>0$, let $u \in \mathcal{A}^s$. Then*

$$\#\mathcal{M}_k \lesssim |u|^{1/s}_{\mathcal{A}^s}\Big(\sqrt{|u-u_k|^2_{H^1(\Omega)}+\operatorname{osc}(\mathcal{T}_k,f)^2}\Big)^{-1/s}.$$

Proof. By definition of $\mathcal{A}^s$, there exists a $\tilde{\mathcal{T}} \in \mathbb{T}$ with

$$\#\tilde{\mathcal{T}}-\#\mathcal{T}_0 \leq |u|^{1/s}_{\mathcal{A}^s}\Big(\sqrt{1-\theta^2C_1(C_2+1)}\sqrt{|u-u_k|^2_{H^1(\Omega)}+\operatorname{osc}(\mathcal{T}_k,f)^2}\Big)^{-1/s},$$

and

$$|u-u_{\tilde{\mathcal{T}}}|^2_{H^1(\Omega)}+\operatorname{osc}(\tilde{\mathcal{T}},f)^2 \leq [(1-\theta^2C_1(C_2+1)][|u-u_k|^2_{H^1(\Omega)}+\operatorname{osc}(\mathcal{T}_k,f)^2]$$

(see Remark 11.2). Take $\mathcal{T} = \mathcal{T}_k \oplus \tilde{\mathcal{T}}$. Then

$$|u-u_{\mathcal{T}}|^2_{H^1(\Omega)}+\operatorname{osc}(\mathcal{T},f)^2 \leq [(1-\theta^2C_1(C_2+1)][|u-u_k|^2_{H^1(\Omega)}+\operatorname{osc}(\mathcal{T}_k,f)^2],$$

and so by the previous lemma, the fact that each refined triangle is splitted into at least two, and Lemma 8.3,

$$\begin{aligned}\#\mathcal{M}_k \leq \#R_{\mathcal{T}_k\to\mathcal{T}} &\leq \#\mathcal{T}-\#\mathcal{T}_k \leq \#\tilde{\mathcal{T}}-\#\mathcal{T}_0\\ &\lesssim |u|^{1/s}_{\mathcal{A}^s}\Big(\sqrt{|u-u_k|^2_{H^1(\Omega)}+\operatorname{osc}(\mathcal{T}_k,f)^2}\Big)^{-1/s}.\end{aligned}$$ □

Theorem 11.7. *Let $\theta^2 < (C_1(C_2+1))^{-1}$. For some $s>0$, let $u \in \mathcal{A}^s$. Then it holds that*

$$\#\mathcal{T}_k-\#\mathcal{T}_0 \lesssim |u|^{1/s}_{\mathcal{A}^s}\Big(\sqrt{|u-u_k|^2_{H^1(\Omega)}+\operatorname{osc}(\mathcal{T}_k,f)^2}\Big)^{-1/s}.$$

That is, the total errors of the sequence of Galerkin approximations produced by `AFEM` *decay with the best possible rate s.*

Proof. By applications of Theorem 11.4, the previous corollary, Corollary 7.3, and Thm. 10.1, we have

$$\begin{aligned}
\#\mathcal{T}_k - \#\mathcal{T}_0 &\lesssim \sum_{i=0}^{k-1} \#\mathcal{M}_i \lesssim |u|_{\mathcal{A}^s}^{1/s} \sum_{i=0}^{k-1} \Big(\sqrt{|u-u_i|_{H^1(\Omega)}^2 + \operatorname{osc}(\mathcal{T}_i, f)^2}\Big)^{-1/s} \\
&\eqsim |u|_{\mathcal{A}^s}^{1/s} \sum_{i=0}^{k-1} \big(|u-u_i|_{H^1(\Omega)}^2 + \gamma\eta(u_i, \mathcal{T}_i)^2\big)^{-\frac{1}{2s}} \\
&\lesssim |u|_{\mathcal{A}^s}^{1/s} (\sum_{i=1}^{k} \alpha^{\frac{i}{2s}}) \big(|u-u_k|_{H^1(\Omega)}^2 + \gamma\eta(u_k, \mathcal{T}_k)^2\big)^{-\frac{1}{2s}} \\
&\eqsim |u|_{\mathcal{A}^s}^{1/s} \Big(\sqrt{|u-u_k|_{H^1(\Omega)}^2 + \operatorname{osc}(\mathcal{T}_k, f)^2}\Big)^{-1/s}.
\end{aligned}$$

□

Literature with this section: [Ste07, CKNS08].

References

[BDD04] P. Binev, W. Dahmen, and R. DeVore. Adaptive finite element methods with convergence rates. *Numer. Math.*, 97(2):219 – 268, 2004.

[BDDP02] P. Binev, W. Dahmen, R. DeVore, and P. Petruchev. Approximation classes for adaptive methods. *Serdica Math. J.*, 28:391–416, 2002.

[BS08] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.

[Cia78] P.G. Ciarlet. *The finite element method for elliptic problems*. North-Holland, Amsterdam, 1978.

[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(5):2524–2550, 2008.

[DD97] S. Dahlke and R. DeVore. Besov regularity for elliptic boundary value problems. *Comm. Partial Differential Equations*, 22(1 & 2):1–16, 1997.

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

[MN05] K. Mekchay and R.H. Nochetto. Convergence of adaptive finite element methods for general second order linear elliptic PDEs. *SIAM J. Numer. Anal.*, 43(5):1803–1827 (electronic), 2005.

[MNS00] P. Morin, R. Nochetto, and K. Siebert. Data oscillation and convergence of adaptive FEM. *SIAM J. Numer. Anal.*, 38(2):466–488, 2000.

[Ste07] R.P. Stevenson. Optimality of a standard adaptive finite element method. *Found. Comput. Math.*, 7(2):245–269, 2007.

[Ste08] R.P. Stevenson. The completion of locally refined simplicial partitions created by bisection. *Math. Comp.*, 77:227–241, 2008.

[Ver96] R. Verfürth. *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*. Wiley-Teubner, Chichester, 1996.