AN EXTENSION OF GENERAL QUASI-ORTHOGONALITY

MICHAEL FEISCHL

ABSTRACT. We propose a generalization of quasi-orthogonality which allows the constant to depend on the number of levels involved. The quasi-orthogonality of Galerkin solutions is a key argument in modern proofs of optimal convergence of adaptive mesh refinement algorithms. Our generalization together with other well understood properties of the error estimator still implies linear convergence of the estimator and hence rate optimal convergence.

1. The abstract setting

We consider real Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ as well as a closed subspaces $\mathcal{X}_\mathcal{T} \subseteq \mathcal{X}$, $\mathcal{Y}_\mathcal{T} \subseteq \mathcal{Y}$ which are based on some triangulation $\mathcal{T}$ of some underlying polyhedral domain $\Omega \subseteq \mathbb{R}^d$ and satisfy $\dim(\mathcal{X}_\mathcal{T}) = \dim(\mathcal{Y}_\mathcal{T})$. We denote the set of all admissible triangulations by $\mathbb{T}$ (we will specify this later) and we consider a bilinear form $a(\cdot,\cdot)\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ which is bounded

$$|a(u,v)| \le C_a \|u\|_\mathcal{X} \|v\|_\mathcal{Y} \quad \text{for all } u \in \mathcal{X},\, v \in \mathcal{Y} \tag{1}$$

and inf-sup stable in the following sense: There exists $\gamma > 0$ such that

$$\inf_{u\in\mathcal{X}} \sup_{v\in\mathcal{Y}} \frac{a(u,v)}{\|u\|_\mathcal{X}\|v\|_\mathcal{Y}} \ge \gamma > 0 \tag{2}$$

as well as

$$\inf_{\mathcal{T}\in\mathbb{T}} \inf_{u\in\mathcal{X}_\mathcal{T}} \sup_{v\in\mathcal{Y}_\mathcal{T}} \frac{a(u,v)}{\|u\|_\mathcal{X}\|v\|_\mathcal{Y}} \ge \gamma > 0. \tag{3}$$

Moreover, we assume that for all $v \in \mathcal{Y} \setminus \{0\}$, there exists $u \in \mathcal{X}$ with $a(u,v) \neq 0$. This allows us to consider solutions $u \in \mathcal{X}$ and $u_\mathcal{T} \in \mathcal{X}_\mathcal{T}$ of

$$a(u,v) = f(v) \quad \text{for all } v \in \mathcal{Y} \quad \text{and} \quad a(u_\mathcal{T},v) = f(v) \quad \text{for all } v \in \mathcal{Y}_\mathcal{T} \tag{4}$$

for some $f \in \mathcal{X}^\star$. Moreover, we immediately obtain a Céa-type estimate of the form

$$\|u - u_\mathcal{T}\|_\mathcal{X} \le \Big(1 + \frac{C_a}{\gamma}\Big) \min_{v\in\mathcal{X}_\mathcal{T}} \|u - v\|_\mathcal{X}. \tag{5}$$

1.1. **Adaptive mesh refinement.** We consider an initial regular and shape regular triangulation $\mathcal{T}_0$ of Ω into compact simplices $T \in \mathcal{T}_0$. We assume that $\mathcal{T}$ partitions Ω such that the intersection of two elements $T \neq T' \in \mathcal{T}$ is either a common face, a common node, or empty. By $\mathbb{T}$, we denote the set of all regular triangulations which can be generated by iterated application of newest vertex bisection to $\mathcal{T}_0$ (see, e.g., [5] for details).

Mesh refinement algorithms are steered by an error estimator $\eta(\mathcal{T}) = \eta(\mathcal{T}, u_\mathcal{T}, f) = \sqrt{\sum_{T\in\mathcal{T}} \eta_T(\mathcal{T})^2}$ which satisfies $\eta(\mathcal{T}) \approx \|u - u_\mathcal{T}\|_\mathcal{X}$ and has the following basic structure:

Date: November 18, 2020.

The author is supported by the Deutsche Forschungsgemeinschaft (DFG) through CRC 1173.

Algorithm 1. ***Input:*** *Initial mesh $\mathcal{T}_0$, parameter $0 < \theta < 1$.*
For $\ell = 0, 1, 2, \dots$ do:

(1) *Compute $u_\ell := u_{\mathcal{T}_\ell}$ from (4).*
(2) *Compute error estimate $\eta_T(\mathcal{T})$ for all $T \in \mathcal{T}_\ell$.*
(3) *Find a set $\mathcal{M}_\ell \subseteq \mathcal{T}_\ell$ of minimal cardinality such that*

$$\sum_{T \in \mathcal{M}_\ell} \eta_T(\mathcal{T})^2 \geq \theta \sum_{T \in \mathcal{T}_\ell} \eta_T(\mathcal{T})^2. \tag{6}$$

(4) *Use newest-vertex-bisection to refine at least the elements in $\mathcal{M}_\ell$ and to obtain a new mesh $\mathcal{T}_{\ell+1}$.*

Output: *Sequence of adaptively refined meshes $\mathcal{T}_\ell$ and corresponding approximations $u_\ell \in \mathcal{X}_\ell := \mathcal{X}_{\mathcal{T}_\ell}$.*

Remark. *Newest-vertex-bisection is a method to bisect d-dimensional simplices such that shape regularity is conserved. It is the de facto standard for mesh refinement in FEM as it is the only method for which we currently can prove optimal convergence. Details can be found in [5].* □

We say that Algorithm 1 is rate optimal (or just optimal) if for every possible convergence rate $s > 0$ which satisfies

$$\sup_{N \in \mathbb{N}} \inf_{\substack{\mathcal{T} \in \mathbb{T} \\ \#\mathcal{T} - \#\mathcal{T}_0 \leq N}} \eta(\mathcal{T}) N^s < \infty \tag{7a}$$

the output of Algorithm 1 also satisfies

$$\sup_{\ell \in \mathbb{N}} \eta(\mathcal{T}_\ell)(\#\mathcal{T}_\ell)^s < \infty. \tag{7b}$$

Note that some authors use a different notion of rate optimality, e.g. [2, 4], where they replace the error estimator by a weighted sum of error and data oscillations. We follow the definition in [1] because as long as the error estimator provides a lower bound for the error, the two definitions coincide. For some problems without a known lower bound (such as, e.g., FEM-BEM coupling [3]), the present definition is more general.

2. General quasi-orthogonality and rate optimal convergence

General quasi-orthogonality is a property of the adaptive sequence only. Therefore, we adopt the notation

$$\eta_\ell := \eta(\mathcal{T}_\ell), \quad \mathcal{X}_\ell := \mathcal{X}_{\mathcal{T}_\ell}, \quad \mathcal{Y}_\ell := \mathcal{Y}_{\mathcal{T}_\ell}, \quad u_\ell := u_{\mathcal{T}_\ell}.$$

The only properties we are going to use are reliability in the sense $\|u - u_\ell\|_{\mathcal{X}} \leq C_{\mathrm{rel}} \eta_\ell$ for all $\ell \in \mathbb{N}$, quasi-monotonicity in the sense $\eta_{\ell+k}^2 \leq C_{\mathrm{mon}} \eta_\ell^2$ for all $\ell, k \in \mathbb{N}$, and nestedness in the sense $\mathcal{X}_{\ell+1} \supseteq \mathcal{X}_\ell$ as well as $\mathcal{Y}_{\ell+1} \supseteq \mathcal{Y}_\ell$ for all $\ell \in \mathbb{N}$.

2.1. General quasi-orthogonality.

For the problems we have in mind, general quasi-orthogonality is the key estimate applying the abstract optimality proof of [1]. The main task to establish a version of this quasi-orthogonality and use it to prove linear convergence of the estimator.

We generalize the definition of general quasi-orthogonality from [1] even further by allowing the constant on the right-hand side $C = C(N)$ to depend on the number of levels included in the estimate:

$$\sum_{k=\ell}^{\ell+N} \|u_{k+1} - u_k\|_{\mathcal{X}}^2 \leq C(N) \|u - u_\ell\|_{\mathcal{X}}^2. \tag{8}$$

In Section 2.2 below we show that $C(N) = o(N)$ is (under some additional assumptions such as estimator reduction and reliability) sufficient for linear convergence of the error estimator.

The quasi-orthogonality (8) is a generalization of the Pythagoras identity

$$\|u - u_{k+1}\|_{\mathcal{X}}^2 + \|u_{k+1} - u_k\|_{\mathcal{X}}^2 = \|u - u_k\|_{\mathcal{X}}^2$$

which holds if $a(\cdot, \cdot)$ is the scalar product of $\mathcal{X}$. In this case, (8) follows immediately from a telescoping sum argument even with $C(N) \simeq 1$. If $a(\cdot, \cdot)$ is non-symmetric or indefinite, this argument no longer works and new ideas are required.

The proofs below would allow us to use an even more general version of (8), i.e.,

$$\sum_{k=\ell}^{\ell+N} \|u_{k+1} - u_k\|_{\mathcal{X}}^2 - \varepsilon \eta_k^2 \leq C(N)\eta_\ell^2$$

as long as $\varepsilon > 0$ is sufficiently small (see also [1, Assumption (A3)], where the above estimate is proposed with $C(N) = C$).

2.2. **Linear convergence.** We show that general quasi-orthogonality (8) implies linear convergence. This is a key step in the optimality proof below. Since we use a weaker version of quasi-orthogonality, we have to use stronger arguments to prove linear convergence compared to the proof in [1, Proposition 4.10].

Lemma 2. *Let* $(\eta_\ell)_{\ell \in \mathbb{N}}$ *satisfy* estimator reduction

$$\eta_{\ell+1}^2 \leq \kappa \eta_\ell^2 + C_{\text{est}} \|u_{\ell+1} - u_\ell\|_{\mathcal{X}}^2 \tag{9}$$

for some $0 < \kappa < 1$, $C_{\text{est}} > 0$ *as well as* reliability

$$\|u - u_\ell\|_{\mathcal{X}} \leq C_{\text{rel}} \eta_\ell \tag{10}$$

for all $\ell \in \mathbb{N}$. *Under general quasi-orthogonality* (8), *there holds*

$$\sum_{k=\ell}^{\ell+N} \eta_k^2 \leq D(N)\eta_\ell^2 \tag{11}$$

with $D(N) = C(1 + C(N))$ *with* $C(N)$ *from* (8) *and* $C > 0$ *depending only on* κ, C_{est}, *and* C_a.

Proof. Estimator reduction and general quasi-orthogonality (8) imply

$$\begin{aligned}
\sum_{k=\ell+1}^{\ell+N} \eta_k^2 &\leq \kappa \sum_{k=\ell+1}^{\ell+N} \eta_{k-1}^2 + C_{\text{est}} \sum_{k=\ell+1}^{\ell+N} \|u_k - u_{k-1}\|_{\mathcal{X}}^2 \\
&\leq \kappa \sum_{k=\ell+1}^{\ell+N} \eta_{k-1}^2 + C_{\text{est}} C(N-1) \|u - u_\ell\|_{\mathcal{X}}^2.
\end{aligned}$$

Reliability of η_ℓ shows

$$(1 - \kappa) \sum_{k=\ell+1}^{\ell+N} \eta_k^2 \leq (\kappa + C_{\text{est}} C(N-1) C_{\text{rel}}) \eta_\ell^2$$

and hence (11). This concludes the proof. $\square$

Lemma 3. *Suppose the sequence $(\eta_\ell)_{\ell\in\mathbb{N}} \subset \mathbb{R}$ satisfies (11) with a bound $D(N) > 1$ such that there exists $N_0 \in \mathbb{N}$ with*

$$\log(D(N_0)) - \sum_{j=1}^{N_0} D(j)^{-1} < 0.$$

If $(\eta_\ell)_{\ell\in\mathbb{N}}$ is additionally quasi-monotone in the sense that there exists $C_{\mathrm{mon}} > 0$ such that

$$\eta_{\ell+k}^2 \le C_{\mathrm{mon}}\eta_\ell^2 \quad \text{for all } \ell, k \in \mathbb{N}, \tag{12}$$

then there exists $0 < q < 1$ and $C > 0$ such that

$$\eta_{\ell+k}^2 \le C q^k \eta_\ell^2$$

for all $k, \ell \in \mathbb{N}$.

Remark. *Note that $D(N) \le CN^{1-\delta}$ for any $\delta > 0$ satisfies the assumption in Lemma 3 since*

$$\log(D(N)) - \sum_{j=1}^{N} D(j)^{-1} \le \log(C) + (1-\delta)\log(N) - C^{-1}\sum_{j=1}^{N} j^{-1+\delta} \to -\infty$$

as $N \to \infty$. Even the edge case $D(N) \lesssim N$ works as long as $\limsup_{N\to\infty} D(N)/N$ is sufficiently small. □

Proof of Lemma 3. We prove by mathematical induction on k that

$$\eta_\ell^2 \le \Big(\prod_{j=1}^{k}(1 - D(j)^{-1})\Big) \sum_{j=\ell-k}^{\ell} \eta_j^2 \tag{13}$$

for all $0 \le k \le \ell$. To that end, we fix $\ell \in \mathbb{N}$ and observe that (13) is trivially true for $k = 0$ (we interpret the empty product as 1). Assume that (13) is true for some $1 \le k < \ell$. Then, (11) implies

$$\begin{aligned}
\eta_\ell^2 &\le \prod_{j=1}^{k}(1 - D(j)^{-1})\Big(\sum_{j=\ell-k-1}^{\ell} \eta_j^2 - \eta_{\ell-k-1}^2\Big)\\
&\le \prod_{j=1}^{k}(1 - D(j)^{-1})\Big(\sum_{j=\ell-k-1}^{\ell} \eta_j^2 - D(k+1)^{-1}\sum_{j=\ell-k-1}^{\ell} \eta_j^2\Big)\\
&\le \prod_{j=1}^{k+1}(1 - D(j)^{-1}) \sum_{j=\ell-k-1}^{\ell} \eta_j^2.
\end{aligned}$$

This concludes the induction and proves (13). A final application of (11) to (13) shows

$$\eta_\ell^2 \le \Big(D(k)\prod_{j=1}^{k}(1 - D(j)^{-1})\Big)\eta_{\ell-k}^2$$

and, replacing ℓ with $\ell + k$, we have

$$\eta_{\ell+k}^2 \le \Big(D(k)\prod_{j=1}^{k}(1 - D(j)^{-1})\Big)\eta_\ell^2.$$

It remains to calculate the constants C and q. To then end, we observe

$$\log\Big(D(k)\prod_{j=1}^{k}(1-D(j)^{-1})\Big) = \log(D(k)) + \sum_{j=1}^{k}\log(1-D(j)^{-1})$$
$$\leq \log(D(k)) - \sum_{j=1}^{k} D(j)^{-1}.$$

Under the assumption on $D(\cdot)$, we may choose $k_0 \in \mathbb{N}$ such that $\log(q_0) := \log(D(k_0)) - \sum_{j=1}^{k_0} D(j)^{-1} < 0$. This implies $\eta_{\ell+k_0}^2 \leq q_0\eta_\ell^2$ for all $\ell \in \mathbb{N}$. Quasi-monotonicity of η_ℓ implies

$$\eta_{\ell+k}^2 \leq \frac{1}{q_0} C_{\rm mon} q_0^{k/k_0}\eta_\ell^2$$

and thus concludes the proof with $q := q^{1/k_0}$ and $C = C_{\rm mon}/q_0$. □

2.3. **Proof of optimal convergence.** We follow the proof in [1] and collect all the assumptions we require below: There exist constants $\gamma, C_a, C_{\rm red}, C_{\rm stab}, C_{\rm dlr}, C_{\rm ref} > 0$, and $0 \leq q_{\rm red} < 1$ such that

(A1) Stability on non-refined elements: For all refinements $\widehat{\mathcal{T}} \in \mathbb{T}$ of a triangulation $\mathcal{T} \in \mathbb{T}$, for all subsets $\mathcal{S} \subseteq \mathcal{T} \cap \widehat{\mathcal{T}}$ of non-refined elements, it holds that

$$\Big|\Big(\sum_{T\in\mathcal{S}}\eta_T(\widehat{\mathcal{T}})^2\Big)^{1/2} - \Big(\sum_{T\in\mathcal{S}}\eta_T(\mathcal{T})^2\Big)^{1/2}\Big| \leq C_{\rm stab}\,\|u_{\mathcal{T}} - u_{\widehat{\mathcal{T}}}\|_{\mathcal{X}}.$$

(A2) Reduction property on refined elements: Any refinement $\widehat{\mathcal{T}} \in \mathbb{T}$ of a triangulation $\mathcal{T} \in \mathbb{T}$ satisfies

$$\sum_{T\in\widehat{\mathcal{T}}\setminus\mathcal{T}}\eta_T(\widehat{\mathcal{T}})^2 \leq q_{\rm red}\sum_{T\in\mathcal{T}\setminus\widehat{\mathcal{T}}}\eta_T(\mathcal{T})^2 + C_{\rm red}\|u_{\mathcal{T}} - u_{\widehat{\mathcal{T}}}\|_{\mathcal{X}}^2.$$

(A3) General quasi-orthogonality: There holds (8) with $C(N) = o(N)$.

(A4) Discrete reliability: For all refinements $\widehat{\mathcal{T}} \in \mathbb{T}$ of a triangulation $\mathcal{T} \in \mathbb{T}$, there exists a subset $\mathcal{R}(\mathcal{T},\widehat{\mathcal{T}}) \subseteq \mathcal{T}$ with $\mathcal{T}\setminus\widehat{\mathcal{T}} \subseteq \mathcal{R}(\mathcal{T},\widehat{\mathcal{T}})$ and $|\mathcal{R}(\mathcal{T},\widehat{\mathcal{T}})| \leq C_{\rm ref}|\mathcal{T}\setminus\widehat{\mathcal{T}}|$ such that

$$\|u_{\widehat{\mathcal{T}}} - u_{\mathcal{T}}\|_{\mathcal{X}}^2 \leq C_{\rm dlr}^2\sum_{T\in\mathcal{R}(\mathcal{T},\widehat{\mathcal{T}})}\eta_T(\mathcal{T})^2.$$

Note that we omitted the general quasi-orthogonality (A3) from [1] and replaced it with the new version (8). Hence, to proceed with the optimality proof, we first apply Section 2.2 to prove linear convergence of the estimator. Then, we use the remaining assumptions to complete the optimality proof laid out in [1] (but inspired by [2,4]).

Theorem 4. *Under assumptions (A1)–(A4), Algorithm 1 is rate optimal in the sense* (7).

Proof. Reliability (10) of the estimator sequence follows from (A4) due to [1, Lemma 3.4] with a constant $C_{\rm rel}$ that depends only on $C_{\rm dlr}$. The result [1, Lemma 4.7] shows that (A1)–(A2) imply estimator reduction (9) for some constants κ and $C > 0$ which depend only on the constants in (A1) and (A2). Moreover, [1, Lemma 3.5] shows that (A1), (A2), and (A3) imply quasi-monotonicity (12) for some constant $C_{\rm mon} > 0$ which depends only on the constants in $(A1), (A2), (A4)$.

Hence, Lemma 2 and Lemma 3 prove

$$\eta_{\ell+k}^2 \le Cq^k \eta_\ell^2$$

for all $\ell, k \in \mathbb{N}$ and uniform constants $0 < q < 1$ and $C > 0$. With this, we have all the requirements of [1, Lemma 4.12] to prove the so-called *optimality of Dörfler marking*. Then, the results [1, Lemma 4.14] and [1, Proposition 4.15] prove rate optimality (7) for Algorithm 1. □

References

[1] Carsten Carstensen, Michael Feischl, Marcus Page, and Dirk Praetorius. Axioms of adaptivity. *Comput. Math. Appl.*, 67(6):1195–1253, 2014.

[2] J. Manuel Cascon, Christian Kreuzer, Ricardo H. Nochetto, and Kunibert G. Siebert. Quasi-optimal convergence rate for an adaptive finite element method. *SIAM J. Numer. Anal.*, 46(5):2524–2550, 2008.

[3] Michael Feischl. Optimal adaptivity for non-symmetric fem/bem coupling. *arxiv-preprint 1710.06082.*

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

[5] Rob Stevenson. The completion of locally refined simplicial partitions created by bisection. *Math. Comp.*, 77(261):227–241, 2008.

Institute for Analysis and Scientific Computing, TU Wien, Wiedner Hauptstrasse 8-10, 1040 Vienna.

Email address: michael.feischl@tuwien.ac.at