



Algebraic error and a posteriori error estimation in numerical PDEs

Jan Papež^{1,2}, Zdeněk Strakoš^{2,3}, Martin Vohralík¹

¹ Inria, Paris

² Institute of Computer Science, CAS, Prague

³ Charles University, Prague

Implicitly constituted materials: Modeling, Analysis and Computing,
July 31, Roztoky



- 1 Introduction
- 2 Residual-based error estimator for the total error
- 3 Estimating total and algebraic errors using flux reconstruction
- 4 Conclusion

1. Introduction

Among the challenges of the a posteriori error analysis

Bounding the appropriate norm of the errors.

Estimating the local distribution of the errors.

Specifying all the multiplicative factors in the estimators making them fully computable.

Declaring all the used assumptions and associated restrictions of the results.

Discussing (and possibly reducing) the evaluation cost of the estimators.

1. Setting and notation

Poisson problem: $-\Delta u = f$ in Ω , $u = 0$ on $\partial\Omega$,

Weak solution $u \in V \equiv H_0^1(\Omega)$,

$$(\nabla u, \nabla v) = (f, v) \quad \forall v \in V,$$

FEM discrete approximation $u_h \in V_h \subset V$,

$$(\nabla u_h, \nabla v_h) = (f, v_h) \quad \forall v_h \in V_h.$$

Algebraic problem, using the basis $\Phi = \{\phi_1, \dots, \phi_N\}$ of V_h ,

$$\mathbf{A}U = F, \quad (\mathbf{A})_{j\ell} = (\nabla\phi_\ell, \nabla\phi_j), \quad F_j = (f, \phi_j), \quad u_h = \Phi U.$$

Inexact solution $U^i \approx U$, $u_h^i = \Phi U^i$, residual $R^i = F - \mathbf{A}U^i$.

$$\underbrace{u - u_h^i}_{\text{total error}} = \underbrace{u - u_h}_{\text{discretization error}} + \underbrace{u_h - u_h^i}_{\text{algebraic error}}$$

Section 2

Residual-based error estimator for the total error



Papež, J. and Strakoš, Z. (2016). On a residual-based a posteriori error estimator for the total error. [Preprint MORE/2016/14, Accepted for publication in IMAJNA.](#)

2. Residual-based error estimator – notation

In this section, we consider discretization using the piecewise affine conforming finite elements.

Denote by

- \mathcal{T}_h the triangulation of Ω with the nodes \mathcal{N} and edges \mathcal{E} ,
- φ_z , $z \in \mathcal{N}$, the hat-function with the support ω_z (the patch).

Define the oscillations of the source term $f \in L^2(\Omega)$

$$\text{osc} \equiv \left(\sum_{z \in \mathcal{N}} |\omega_z| \|f - \text{mean}(f, \omega_z)\|_{\omega_z}^2 \right)^{1/2},$$

and for $w_h \in V_h$ the edge residual measuring the jumps of a piecewise constant function ∇w_h over the inner edges

$$J(w_h) \equiv \left(\sum_{E \in \mathcal{E} \setminus \partial\Omega} |E| \|[\nabla w_h \cdot n_E]\|_E^2 \right)^{1/2}.$$

2. Residual-based error estimator

For the Galerkin solution u_h there exists $C > 0$ depending on the minimal angle of the triangulation such that

$$\|\nabla(u - u_h)\|^2 \leq C (J_h^2(u_h) + \text{osc}^2) ;$$

see, e.g., [Carstensen (1999)].

The proof uses the so-called Clément quasi-interpolation operator

$$\mathcal{I} : L^1(\Omega) \rightarrow V_h.$$

2. Bounding the total error

[Becker, Mao (2009), Lemma 3.1]:

$$\|\nabla(u - w_h)\|^2 \leq C (J_h^2(w_h) + \text{osc}^2) + 2 \|\nabla(u_h - w_h)\|^2.$$

Proof: “The upper bound with $w_h = u_h$ has been proven by [Carstensen (1999)] introducing a weighted Clément-type quasi-interpolation operator. The generalization to $w_h \neq u_h$ follows from the triangle inequality.”

2. Bounding the total error

[Becker, Mao (2009), Lemma 3.1]:

$$\|\nabla(u - w_h)\|^2 \leq C (J_h^2(w_h) + \text{osc}^2) + 2 \|\nabla(u_h - w_h)\|^2.$$

Proof: “The upper bound with $w_h = u_h$ has been proven by [Carstensen (1999)] introducing a weighted Clément-type quasi-interpolation operator. The generalization to $w_h \neq u_h$ follows from the triangle inequality.”

[Arioli, Georgoulis, Loghin (2013), proof of Theorem 3.3]:

$$\begin{aligned} \|\nabla(u - w_h)\|^2 &\leq 2C_{2.2} (J_h^2(w_h) + \widetilde{\text{osc}}^2) \\ &\quad + (1 + 2C_{2.2}C_{3.1}) \|\nabla(u_h - w_h)\|^2. \end{aligned}$$

In numerical experiments they empirically set $C_{2.2} := 40$, $C_{3.1} := 10$.

2. Revised bound

Elaborating on [Carstensen (1999)], we can show that

$$\|\nabla(u - w_h)\|^2 \leq C(J_h^2(w_h) + \text{osc}^2) + 2\tilde{C}_{\text{intp}}^2(w_h) \|\nabla(u_h - w_h)\|^2.$$

with

$$\tilde{C}_{\text{intp}}(w_h) \equiv \frac{\|\nabla(\mathcal{I}u - \mathcal{I}w_h)\|}{\|\nabla(u - w_h)\|}.$$

2. Revised bound

Elaborating on [Carstensen (1999)], we can show that

$$\|\nabla(u - w_h)\|^2 \leq C(J_h^2(w_h) + \text{osc}^2) + 2\tilde{C}_{\text{intp}}^2(w_h) \|\nabla(u_h - w_h)\|^2.$$

with

$$\tilde{C}_{\text{intp}}(w_h) \equiv \frac{\|\nabla(\mathcal{I}u - \mathcal{I}w_h)\|}{\|\nabla(u - w_h)\|}.$$

A priori bound [Carstensen (1999), Theorem 3.1]:

There exists $C_{\text{intp}} > 0$ depending only on the triangulation \mathcal{T} such that, for all $w \in H_0^1(\Omega)$,

$$\|\nabla \mathcal{I}w\| \leq C_{\text{intp}} \|\nabla w\|.$$

This gives $C_{\text{intp}} \geq \tilde{C}_{\text{intp}}(w_h)$, for any $w_h \in V_h$.

2. Solution-independent factor and overestimation

The factor C_{intp} represents the worst-case scenario and most likely $C_{\text{intp}} \gg \tilde{C}_{\text{intp}}(w_h)$.

Using the discussion in [Carstensen (2006), Section 2], for a square domain Ω , homogeneous Dirichlet BC and a shape-regular mesh,

$$C_{\text{intp}} \approx 6.$$

In general, “it may be very large for small angles in the triangulation”.

2. Numerical illustration

Poisson problem on the square $\Omega \equiv (-1, 1) \times (-1, 1)$, Delaunay triangulation with 1368 elements and with the minimal angle of the mesh equal to 35.9° (the average of the minimal angles of the elements is 50.3°). We recall, that in this setting $C_{\text{intp}} \approx 6$.

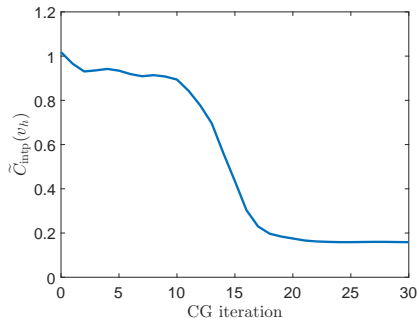
The exact solution is set as

$$u(x, y) = (x - 1)(x + 1)(y - 1)(y + 1),$$

and we plot $\tilde{C}_{\text{intp}}(u_h^i)$ for the approximations u_h^i generated by the conjugate gradient method with zero initial vector for solving the discretized problem.

2. Numerical illustration

Poisson problem on the square $\Omega \equiv (-1, 1) \times (-1, 1)$, Delaunay triangulation with 1368 elements and with the minimal angle of the mesh equal to 35.9° (the average of the minimal angles of the elements is 50.3°). We recall, that in this setting $C_{\text{intp}} \approx 6$.



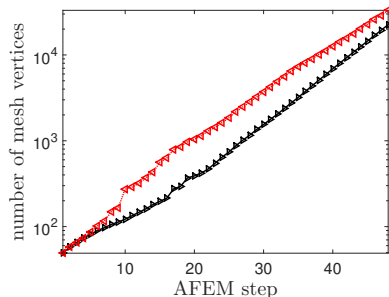
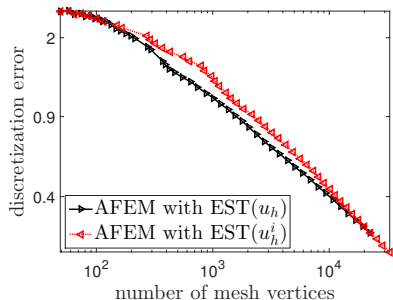
The factor $\tilde{C}_{\text{intp}}(u_h^i)$ for the approximations u_h^i generated in the iterations of the conjugate gradient method.

2. Adaptive mesh refinement

- $\text{EST}(u_h) \equiv (J_h^2(u_h) + \text{osc}^2)^{1/2}$ bounds the *discretization* error and allows its local estimation. The adaptive mesh refinement based on the associated error indicators has been studied and mathematically justified, e.g. in [Morin *et al.* 2002].
- The *efficiency* of adaptive procedures based on $\text{EST}(u_h^i)$ remains an open question. Does $\text{EST}(u_h^i)$ indicate the parts of the computational domain where the discretization error is large?
- $\text{EST}(u_h^i)$ can be evaluated locally. Algebraic error?

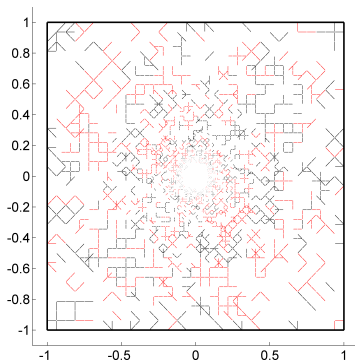
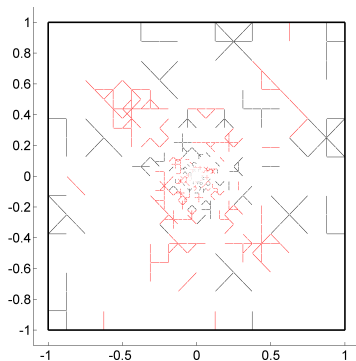
$$\|\nabla(u - u_h^i)\|^2 \leq C \cdot \text{EST}^2(u_h^i) + C_{\text{intp}} \|\nabla(u_h - u_h^i)\|^2,$$

2. Adaptive mesh refinement based on $EST(u_h^i)$



Left: the decrease of the discretization error norm in adaptive FEM that is based on $EST(u_h)$ (black) and $EST(u_h^i)$ (red), respectively.
Right: the corresponding number of degrees of freedom in refinement steps.


2. Adaptive mesh refinement based on $EST(u_h^i)$



The difference of the adaptively refined meshes after 35, respectively 48 refinement steps.

Section 3

Estimating total and algebraic errors using flux reconstruction

 Papež, J., Strakoš, Z., and Vohralík, M. (2016). Estimating and localizing the algebraic and total numerical errors using flux reconstructions. [Preprint MORE/2016/12, Submitted.](#)

3. Estimating total error using flux reconstruction

Goal: derive an estimator in the form

$$\|\nabla(u - u_h^i)\| \leq \eta_{\text{tot}}^i = \eta_{\text{disc}}^i + \eta_{\text{alg}}^i + \eta_{\text{osc}}^i,$$

where

$$\|\nabla(u_h - u_h^i)\| \leq \eta_{\text{alg}}^i$$

and

$$\|\nabla(u - u_h^i)\|_K \approx \eta_{\text{tot},K}^i \quad \|\nabla(u_h - u_h^i)\|_K \approx \eta_{\text{alg},K}^i.$$

The derivation elaborates on [Jiránek *et al.* 2010] and [Ern, Vohralík 2013] and it is based on quasi-equilibrated flux reconstruction.

3. Construction of the estimator

Flux $-\nabla u$ satisfies, in the Poisson problem,

$$-\nabla u \in \mathbf{H}(\operatorname{div}, \Omega), \quad \operatorname{div}(-\nabla u) = f.$$

Given $U^i \approx U$, u_h^i ,

- 1 represent the algebraic residual $R^i = F - \mathbf{A}U^i$ by $r_h^i \in L^2$

Elementwise construction, solving the local problems with mass matrices corresponding to each element.

3. Construction of the estimator

Flux $-\nabla u$ satisfies, in the Poisson problem,

$$-\nabla u \in \mathbf{H}(\operatorname{div}, \Omega), \quad \operatorname{div}(-\nabla u) = f.$$

Given $U^i \approx U$, u_h^i ,

- 1 represent the algebraic residual $R^i = F - \mathbf{A}U^i$ by $r_h^i \in L^2$
- 2 from ∇u_h^i construct the flux reconstruction $\mathbf{d}_h^i \in \mathbf{H}(\operatorname{div}, \Omega)$

$$\operatorname{div} \mathbf{d}_h^i = f - r_h^i$$

Patchwise construction, solving the local problems corresponding to each vertex in the triangulation.

3. Construction of the estimator

Flux $-\nabla u$ satisfies, in the Poisson problem,

$$-\nabla u \in \mathbf{H}(\operatorname{div}, \Omega), \quad \operatorname{div}(-\nabla u) = f.$$

Given $U^i \approx U$, u_h^i ,

- 1 represent the algebraic residual $R^i = F - \mathbf{A}U^i$ by $r_h^i \in L^2$
- 2 from ∇u_h^i construct the flux reconstruction $\mathbf{d}_h^i \in \mathbf{H}(\operatorname{div}, \Omega)$

$$\operatorname{div} \mathbf{d}_h^i = f - r_h^i$$

- 3 estimate the discretization error using $\|\nabla u_h^i + \mathbf{d}_h^i\|$
- 4 bound the algebraic error using $\|r_h^i\|$

The upper bound on the alg.error can be related to the (algebraic) worst-case bound. It can significantly overestimate the algebraic error and therefore the estimators are made more accurate for the price of performing (possibly many) additional algebraic iterations.

3. Construction using additional iterations

Given $U^i \approx U$, u_h^i ,

- 1 represent $R^i = F - \mathbf{A}U^i$ by $r_h^i \in L^2$
- 2 from ∇u_h^i construct $\mathbf{d}_h^i \in \mathbf{H}(\text{div}, \Omega)$

$$\text{div } \mathbf{d}_h^i = f - r_h^i$$

- 3 estimate the discretization error using $\|\nabla u_h^i + \mathbf{d}_h^i\|$
- 4 perform ν additional algebraic iterations giving $U^{i+\nu}$, $u_h^{i+\nu}$
- 5 represent $R^{i+\nu} = F - \mathbf{A}U^{i+\nu}$ by $r_h^{i+\nu} \in L^2$
- 6 from $\nabla u_h^{i+\nu}$ construct $\mathbf{d}_h^{i+\nu} \in \mathbf{H}(\text{div}, \Omega)$

$$\text{div } \mathbf{d}_h^{i+\nu} = f - r_h^{i+\nu}$$

- 7 estimate the algebraic error using $\|r_h^{i+\nu}\|$ and $\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|$

3. Upper bounds

Upper bounds: [Papež, Strakoš, Vohralík (2016)]

$$\|\nabla(u - u_h^i)\| \leq \eta_{\text{osc}} + \|\mathbf{d}_h^i - \mathbf{d}_h^{i+\nu}\| + C_F h_\Omega \|r_h^{i+\nu}\| + \|\nabla u_h^i + \mathbf{d}_h^i\|$$

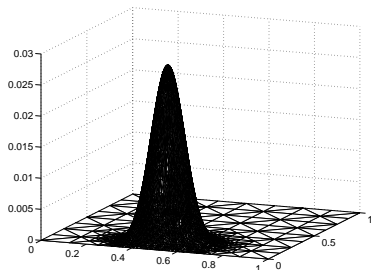
$$\|\nabla(u_h - u_h^i)\| \leq \|\mathbf{d}_h^i - \mathbf{d}_h^{i+\nu}\| + C_F h_\Omega \|r_h^{i+\nu}\|$$

- + the bounds are fully computable
- + the bounds allows for global and local estimation of the error
- + the bounds are independent of the algebraic solver
- evaluation of the estimator can be very costly (flux reconstructions + additional algebraic iterations)

3. Numerical illustration – test problem

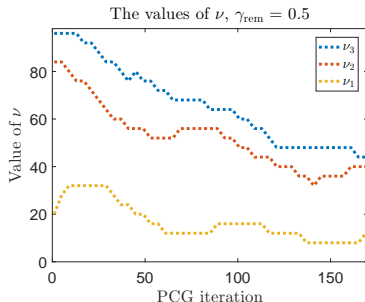
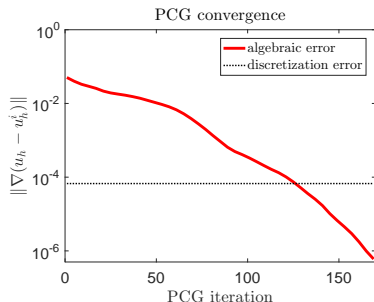
$$u(x, y) = x(x - 1)y(y - 1) \exp\left(-100(x - 0.5)^2 - 100(y - 0.117)^2\right)$$

$\Omega = (0, 1) \times (0, 1)$, FEM discretization with piecewise 2nd order polynomials, algebraic system solved by PCG with `ichol` preconditioner.



Solution u .

3. Cost of the bounds



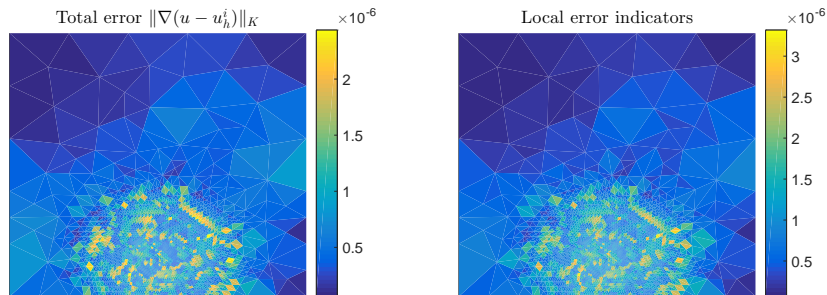
Convergence of PCG with `icho1` preconditioner and the number of additional iterations needed for the evaluation of the error bounds.

ν_1 – theoretical (minimal) value corresponding to no overestimation

ν_2 – value corresponding to algebraic worst-case bound (requires $\lambda_{\min}(\mathbf{A})$)

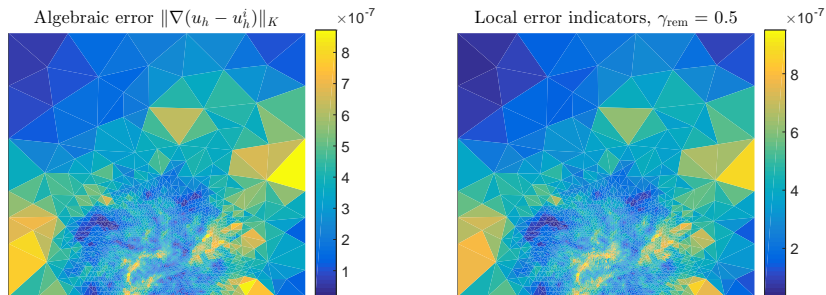
ν_3 – number of iterations needed for our upper bounds

3. Total error and the local indicators



Elementwise distribution of the total error (left) and the local error indicators (right).

3. Algebraic error and the local indicators



Elementwise distribution of the algebraic error (left) and the local error indicators (right).

3. Current work

Avoid additional iteration steps in the setting using a hierarchy of meshes, which naturally fits into the framework of geometric multigrid methods. Idea: represent a (non-zero) algebraic residual by a representer that guarantees a coarsest-level orthogonality.



Papež, J., Růde, U., Vohralík, M., and Wohlmuth, B. Sharp algebraic and total a posteriori error bounds via a multilevel approach. [In preparation.](#)

Herein, we also discuss the cost of flux reconstructions and heuristics how to reduce it.

Section 4

Conclusion

4. Topics of particular importance

- Deriving tight and cost-acceptable bounds on the errors of different origin that allow for accurate estimating their spatial distribution across the domain. Derivations should clearly declare all assumptions that can restrict applicability of the results.
- Studying the influence of the algebraic error on local a posteriori error indicators and the adaptive refinement procedures.
- Deriving mathematically justified stopping criteria that balance (in the appropriate problem-dependent sense) the errors of different origin and that avoid stopping the algebraic iterations prematurely.
- Investigating procedures that would allow to efficiently reduce the algebraic error in some parts of the domain where it is indicated to be large.

4. References I

- Arioli, M., Georgoulis, E. H., and Loghin, D. (2013). Stopping criteria for adaptive finite element solvers. *SIAM J. Sci. Comput.*, 35(3):A1537–A1559.
- Becker, R. and Mao, S. (2009). Convergence and quasi-optimal complexity of a simple adaptive finite element method. *M2AN Math. Model. Numer. Anal.*, 43(6):1203–1219.
- Carstensen, C. (1999). Quasi-interpolation and a posteriori error analysis in finite element methods. *M2AN Math. Model. Numer. Anal.*, 33(6):1187–1202.
- Carstensen, C. (2006). Clément interpolation and its role in adaptive finite element error control. In *Partial differential equations and functional analysis*, volume 168 of *Oper. Theory Adv. Appl.*, pages 27–43. Birkhäuser, Basel.
- Ern, A. and Vohralík, M. (2013). Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs. *SIAM J. Sci. Comput.*, 35(4):A1761–A1791.

4. References II

Golub, G. H. and Strakoš, Z. (1994). Estimates in quadratic formulas. *Numer. Algorithms*, 8(2-4):241–268.

Jiránek, P., Strakoš, Z., and Vohralík, M. (2010). A posteriori error estimates including algebraic error and stopping criteria for iterative solvers. *SIAM J. Sci. Comput.*, 32(3):1567–1590.

Strakoš, Z. and Tichý, P. (2002). On error estimation in the conjugate gradient method and why it works in finite precision computations. *Electron. Trans. Numer. Anal.*, 13:56–80.

Papež, J. (2016). Algebraic error in matrix computations in the context of numerical solution of partial differential equations. *Doctoral thesis, Faculty of Mathematics and Physics, Charles University.*

Thank you for your attention!

jan.papez@inria.fr