On a residual-based a posteriori error estimator for the total error[*]

J. Papež[†] Z. Strakoš[†]

December 28, 2016

Abstract

A posteriori error analysis in numerical PDEs aims at providing sufficiently accurate information about the distance of the numerically computed approximation to the true solution. Besides estimating the total error, a posteriori analysis should also provide information about its discretization and (inexact) algebraic computation parts. This issue has been addressed by many authors using different approaches. Historically probably the first and practically very important approach is based on combination of the classical residual-based bound on the *discretization error* with the adaptive hierarchy of discretizations and computations that allows to incorporate, using various heuristic arguments, the algebraic error. Motivated by some recent publications, this text uses a complementary approach and examines subtleties of the (generalized) residual-based a posteriori error estimator for the *total error* that rigorously accounts for the algebraic part of the error. The aim is to show on the standard Poisson model problem example, which is used here as a case study, that a rigorous incorporation of the algebraic error represents an intriguing problem that is not yet completely resolved. That should be of concern in h-adaptivity approaches where the refinement of the mesh is determined using the residual-based a posteriori error estimator assuming Galerkin orthogonality. The commonly used terminology such as "guaranteed computable upper bounds" should be in the presence of algebraic error cautiously examined.

Keywords: A posteriori error analysis, residual-based estimator, finite element method, Galerkin orthogonality, inexact algebraic solution, total error estimator, guaranteed computable upper bound.

MSC: 65N15, 65N22, 65N30, 65N50.

1 Introduction

Historically, most a posteriori analysis in numerical PDEs focuses on estimating the discretization error, i.e., on the discrepancy between the solution of the original infinite-dimensional formulation of the problem and the *exact solution* of its discretized counterpart. This information is crucial for h-adaptivity, which refines discretization in the parts of the domain where the estimator indicates a large discretization error with the goal of achieving its close-to-uniform spatial distribution over the domain. Estimation of the discretization error, however, has to deal with the principal difficulty: the exact solution of the original problem is unknown, and, unless the algebraic computations providing the coordinates of the discrete solution in the discretization basis are performed exactly or with a negligible algebraic error, the exact solution of the discretized problem is also unknown. Near-to-exact algebraic computations can be prohibitive due to extensive computational cost. When solving practical problems, one typically needs to estimate the error even for the computed approximation far from the exact solution of the discretized problem; see, e.g., (Nordbotten & Bjørstad, 2008, Conclusions), Keilegavlen & Nordbotten (2015), Nissen *et al.* (2015). Reaching exact algebraic results can even be theoretically prohibitive. The matrix eigenvalues, e.g., are (in general) in principle uncomputable by any finite formula as proved by the Abel–Galois theorem, and they can only be approximated iteratively. Moreover, in case of highly non-normal matrices there is no guaranteed forward estimate of the accuracy of the computed eigenvalue approximations and we can guarantee the backward error only. As a consequence, due to the inexactness

[*] This work was supported by the ERC-CZ project MORE LL1202 financed by the MŠMT of the Czech Republic.

[†] Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 186 75 Prague, Czech Republic.

of algebraic computations, the a posteriori error estimates should be based, from their *derivation to their application*, on the available computed approximations to the solution of the discrete problem.

When considering simple model problems, the previous point is seemingly unimportant. The need for solving adaptively large scale problems however requires abandoning the concept of highly accurate algebraic solutions. Incorporating multilevel discretization structure and the associated preconditioned iterative algebraic solvers with reliable stopping criteria is becoming a prerequisite for efficient very large scale numerical PDE solvers. This has been clearly formulated *as a program* by Becker, Johnson, and Rannacher already in the paper Becker *et al.* (1995); see also other relevant references below. Since then, there is a growing number of work in this direction. The recent survey with many references can be found, e.g., in (Arioli *et al.*, 2013a, Section 4); see also the introductory parts of Jiránek *et al.* (2010), Arioli *et al.* (2013b), Papež *et al.* (2016), and (Málek & Strakoš, 2015, Chapter 12). Multilevel discretization structure is often obtained using discretization mesh adaptivity that requires, as mentioned above, local estimation of the discretization error in order to identify mesh elements that need to be refined. For that purpose a standard residual-based a posteriori bound on the discretization error is used; see, e.g., Babuška & Rheinboldt (1978), (Verfürth, 1996, Section 1.2), (Ainsworth & Oden, 2000, Section 2.2), (Brenner & Scott, 2008, Section 9.2). Since the exact solution of the algebraic problem is not available, the computed approximation that does not satisfy Galerkin orthogonality is often used in the bound, which violates the assumptions under which the bound has been derived.

This raises a question to which extent the standard a posteriori error bounds assuming Galerkin orthogonality can be modified in order to estimate the total error and, at the same time, to allow comparison of the discretization and algebraic part of the error, and, consequently, construction of reliable stopping criteria for algebraic iterative solvers. The standard residual-based a posteriori bound is also a key ingredient of the works that use hierarchy of approximation spaces for estimating the total error and its algebraic part; for the early examples we refer, in particular, to Bramble *et al.* (1990), Xu (1992), Oswald (1993), Rüde (1993a,b), Griebel & Oswald (1995), Becker *et al.* (1995).

Therefore we focus in this paper on the residual-based a posteriori bound as a basic building block used elsewhere and investigate its extension for estimating the total error. We do not share the belief that the matter has been fully resolved, apart from simple technicalities. We describe difficulties that still remain open and have to be taken into account in various circumstances.

Using, in particular, the papers by Becker & Mao (2009), Carstensen (1999), and Arioli *et al.* (2013b), we discuss the subtleties one has to deal with while estimating the total and discretization error using a residual-based a posteriori error estimator. We show that removing the standard Galerkin orthogonality assumption, which can not be used in a large scale practical application of the bound, requires a nontrivial revision of the known estimator. Even for the *simple model problem*, the derived extension of the bound contains multiplicative factors that are potentially very large, as shown also in Arioli *et al.* (2013b), and that can not be, in general, easily and accurately determined. Moreover, despite the claims published in literature, there exist no a posteriori estimator for the algebraic part of the error that is cheap, easily computable and that gives in practice a tight guaranteed upper bound. As pointed out below, a rigorous a posteriori analysis that incorporates algebraic errors is for realistic problems substantially more difficult than the analysis that assumes exact algebraic computations.

In Section 2 we set the notation, discuss some methodological questions, and recall several approaches that use a residual-based a posteriori error estimator as a building block. Section 3 recalls the results from Carstensen (1999) on quasi-interpolation. Section 4 presents the revision of the upper bound on the total error from Becker & Mao (2009) and gives its detailed proof that abandons the Galerkin orthogonality assumption. Section 5 comments on the upper bound on the total error presented in Arioli *et al.* (2013b). Numerical illustrations of the difficulties associated with the multiplicative factors are present in Section 6. Section 7 addresses estimates of the algebraic error and explains misunderstandings present in literature. The paper is closed by conclusions.

2 Model problem and comments on methodology

We will use the following standard model problem. Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain (open, bounded and connected set with a polygonal boundary). We consider the Poisson problem with the homogeneous

Dirichlet boundary condition

$$\text{find } u : \Omega \to \mathbb{R} : \qquad -\Delta u = f \quad \text{in} \quad \Omega, \qquad u = 0 \quad \text{on} \quad \partial\Omega\,, \tag{2.1}$$

where $f : \Omega \to \mathbb{R}$ is the source term. Hereafter we use the standard notation for the Sobolev spaces. For $D \subset \Omega$, $L^1(D)$ denotes the space of the (Lebesgue) integrable functions in D, $L^2(D)$ denotes the space of the square integrable functions in D, $(w, v)_D = \int_D v\, w$ denotes the L^2-inner product on $L^2(D)$, and $\|w\|_D = (w, w)_D^{1/2}$ denotes the associated L^2-norm. We omit the subscripts for $D = \Omega$. $H^k(\Omega)$ denotes the Hilbert space of functions in $L^2(\Omega)$ whose weak derivatives up to the order k belong to $L^2(\Omega)$. $H^1_0(\Omega)$ denotes the space of functions in $H^1(\Omega)$ with vanishing trace on the boundary $\partial\Omega$.

Assuming $f \in L^2(\Omega)$, the problem (2.1) can be written in the following weak form

$$\text{find } u \in V \equiv H^1_0(\Omega) : \qquad (\nabla u, \nabla v) = (f, v) \qquad \text{for all } v \in V\,. \tag{2.2}$$

Let $\mathcal{T}$ be a conforming triangulation of the domain Ω, i.e., two distinct and intersecting elements $T_1, T_2 \in \mathcal{T}$ share a common face, edge or vertex. Let $\mathcal{N}$ denote the set of all nodes (i.e. the vertices of the elements of $\mathcal{T}$) while $\mathcal{N}_{\text{int}} \equiv \mathcal{N}\backslash\partial\Omega$ denotes the set of the free nodes. By $\mathcal{E}$ we denote the set of all edges of the elements of $\mathcal{T}$ and, similarly, $\mathcal{E}_{\text{int}} \equiv \mathcal{E}\backslash\partial\Omega$. For any node $z \in \mathcal{N}$, let φ_z be the corresponding hat-function, i.e., the piecewise linear function that takes value 1 at the node z and vanishes at all other nodes. By ω_z we denote the support of φ_z which is equal to the patch $\omega_z = \cup\{T \in \mathcal{T} | z \in T\}$. For an element $T \in \mathcal{T}$ we denote $h_T \equiv \text{diam}(T)$, similarly $h_z \equiv \text{diam}(\omega_z)$ denotes the diameter of ω_z, $z \in \mathcal{N}$. By $V_h \subset V$ we denote the space of the continuous piecewise linear functions on the triangulation $\mathcal{T}$ vanishing on the boundary $\partial\Omega$, i.e. $V_h \equiv \text{span}\,\{\varphi_z | z \in \mathcal{N}_{\text{int}}\}$. The discrete formulation of (2.2) then reads

$$\text{find } u_h \in V_h : \qquad (\nabla u_h, \nabla v_h) = (f, v_h) \qquad \text{for all } v_h \in V_h\,. \tag{2.3}$$

The solution u_h of (2.3) is called the *Galerkin solution.* Subtracting (2.3) from (2.2) and using $V_h \subset V$, we get the *Galerkin orthogonality*

$$(\nabla(u - u_h), \nabla v_h) = 0 \qquad \text{for all } v_h \in V_h\,. \tag{2.4}$$

The difficulty in estimating the discretization error $u - u_h$ mentioned above can be formulated as follows. Consider any estimator $\text{EST}(\cdot)$ that provides an upper bound

$$|||u - u_h||| \le \text{EST}(u_h)\,, \tag{2.5}$$

where $|||\cdot|||$ denotes an appropriate norm (for the model problem (2.1) typically the energy norm $|||w||| = \|\nabla w\| = (\nabla w, \nabla w)^{1/2}$). In order to evaluate the right-hand side of (2.5) we need u_h that is not available. The common practice is then replacing u_h by the computed approximation u_h^C, giving the seemingly easy solution

$$|||u - u_h||| \le \text{EST}(u_h^C)\,.$$

This inequality is, however, not guaranteed to hold without further justification that can be highly nontrivial or even impossible to achieve. Provided that

$$\text{EST}(u_h) = \inf_{v_h \in V_h} \text{EST}(v_h)\,, \tag{2.6}$$

the bound (2.5) does indeed lead to a guaranteed upper bound

$$|||u - u_h||| \le \text{EST}(v_h) \quad \text{for all } v_h \in V_h\,. \tag{2.7}$$

Proving (2.6) can, however, represent a challenge and the authors are unaware of any applicable results of this kind published in literature. Another option is simply writing

$$\text{EST}(u_h) = \text{EST}(u_h^C) + \big(\text{EST}(u_h) - \text{EST}(u_h^C)\big)\,, \tag{2.8}$$

or using a variant of this based on a specific form of the estimator. Then the evaluation of the first term $\text{EST}(u_h^C)$ does not require any assumptions (it should not be confused with estimating the total error). Provided that the second term $\text{EST}(u_h) - \text{EST}(u_h^C)$ could be bounded using the computed quantity u_h^C, the relation (2.8) would give a rigorous bound on the discretization error. This consideration has been used in combination with heuristics in various approaches.

Before focusing on the residual-based error estimator itself, we recall several ideas from the approaches combining this estimator with the hierarchy of discretizations. In the prototype paper Becker *et al.* (1995), the error $u - v_h$ of any approximation $v_h \in H_0^1(\Omega)$ is expressed[1] as

$$\begin{aligned}\|\nabla(u - v_h)\|^2 &= (f, u - v_h) - (\nabla v_h, \nabla(u - v_h)) \\ &=: \langle r(v_h), u - v_h\rangle\end{aligned}$$

where $r(v_h) \in H^{-1}(\Omega)$ is the associated residual and $\langle\cdot,\cdot\rangle : H^{-1}(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ represents the duality pairing. Using the hierarchy of meshes and the associated discrete approximation subspaces $V_0 \subset V_1 \subset \cdots \subset V_L \subset H_0^1(\Omega)$, the paper considers a multigrid algorithm with the Galerkin projection property and the operators I_j, $j = 0, 1, \ldots, L$, where

$$I_j : H_0^1(\Omega) \to V_j\,, \quad j = 0, 1, \ldots, L\,.$$

A straightforward substitution then gives

$$\begin{aligned}\langle r(v_h), u - v_h\rangle &= \langle r(v_h), (u - v_h) - I_L(u - v_h)\rangle && (2.9)\\ &+ \sum_{j=1}^{L}\langle r(v_h), (I_j - I_{j-1})(u - v_h)\rangle && (2.10)\\ &+ \langle r(v_h), I_0(u - v_h)\rangle\,. && (2.11)\end{aligned}$$

The first term (2.9) is then bounded using the standard residual-based a posteriori error estimator with the (non-Galerkin) input quantity v_h. The second term (2.10) is bounded using the algebraic residuals on the individual levels $j = L, L-1, \ldots, 1$. This will bring in nontrivial multiplicative factors analogous to these presented later in our paper. Finally, the last term (2.11) is *assumed to vanish* because of the exact solution of the problem on the coarsest mesh. This assumption is substantial, as demonstrated also by the numerical experiment in Section 7 of the quoted paper. The authors also suggest heuristics for stopping criteria in an adaptive algorithm and for approximation of the unknown multiplicative factors. The derivation does not consider roundoff error.

There is a large amount of work that in principle can be put into the framework of (2.8), where EST$(\cdot)$ is again the residual-based error estimator and the difference (EST(u_h) − EST(u_h^C)) that reflects the algebraic error is estimated using the hierarchy of splittings of the approximation space $H_0^1(\Omega)$ or of its appropriate discretization; see, e.g., Bramble *et al.* (1990), Xu (1992), Oswald (1993), Rüde (1993a,b), Griebel & Oswald (1995), Becker *et al.* (1995), Harbrecht & Schneider (2016), as well as further references in (Arioli *et al.*, 2013a, Section 4), (Málek & Strakoš, 2015, Chapter 12).

The instructive paper Stevenson (2007) extends the approach of another remarkable paper Morin *et al.* (2002) on convergence of adaptive FEM, where the adaptivity is based on the residual-based a posteriori error bound on the discretization error. Stevenson's rigorously presented results account for inexact algebraic computations, but they show that such extension is indeed highly intriguing. The main result in Stevenson (2007) relies on the continuity argument, i.e., it assumes that the algebraic solver deviates from the exact result in a *sufficiently small way.* For the algebraic solver this paper refers to the work of Wu & Chen (2006) on uniform convergence of multigrid V-cycle algorithm. The paper Wu & Chen (2006) assumes exact arithmetic and, in particular, the exact solution of the problem on the coarsest mesh.

The recalled results underline the importance of understanding the extension of the residual-based a posteriori error bounds to inexact algebraic computations (non-Galerkin solutions). In (Becker & Mao, 2009, Lemma 3.1) the bound on the total error is given in the form

$$\|\nabla(u - v_h)\|^2 \le \widetilde{C} \cdot \mathrm{EST}(v_h) + 2\,\|\nabla(u_h - v_h)\|^2\,, \qquad (2.12)$$

where $\widetilde{C}$ is stated to depend only on the minimal angle of the triangulation $\mathcal{T}$, and $v_h \in V_h$ is arbitrary, i.e., it can account for inexact algebraic computations where $v_h = u_h^C$. Here the first term $\widetilde{C} \cdot \mathrm{EST}(v_h)$

[1] Here we use the notation of our paper and consider the estimate for the energy norm of the error.

represents the first term in (2.8) given by the standard residual-based a posteriori error estimator for the discretization error (with replacing the Galerkin solution u_h by v_h). The second term $2\|\nabla(u_h - v_h)\|^2$ does not have the meaning of the second term in (2.8). The proof refers for the case $v_h = u_h$, i.e., for estimating the discretization error, to the paper Carstensen (1999). The proof is completed by arguing, without detailed explanation, that the general case (2.12) follows via the triangle inequality.

In order to be valid and applicable in practical computations as an *upper bound*, the total error estimator of the form (2.12) must resolve two challenges.

1. It must rigorously justify using the arbitrary (non-Galerkin) $v_h \in V_h$ in the first term on the right-hand side of (2.12) and give the necessary information on the value of the factor $\widetilde{C}$.

2. It must be *proved* that in practical computations we can indeed provide a meaningful (i.e. inexpensive and tight) upper bound on the norm $\|\nabla(u_h - u_h^C)\|$ of the algebraic error.

In the present paper we examine the derivation of (2.12) and prove that an estimator of the analogous form can indeed be used also for approximations u_h^C that do not solve the discretized problem exactly. Although the derivation is simple, and in comparison to the standard residual-based a posteriori error estimator assuming Galerkin orthogonality, the resulting bound seemingly only adds the term bounding the algebraic error, the whole matter is, in our opinion, not simple. First, as explained below, there is no proof that the given estimator provides in practice a guaranteed computable upper bound due to subtleties determining the multiplicative factors and due to effects of roundoff on estimating the algebraic error. Second, here we consider a simple model problem and point out difficulties *that are not technical but methodological*. It is not clear at all whether for more complicated problems an extension of the estimator preserves the same form or even whether an estimator based on the related methodology can meaningfully incorporate algebraic errors.

3 Quasi-interpolation results

This section presents results from Carstensen (1999) used further in the text. We include them here for completeness and self-consistency of the text. Denote by ψ a piecewise linear function taking value 1 at the inner nodes $z \in \mathcal{N}_{\rm int}$ and vanishing on the boundary $\partial\Omega$, $\psi \equiv \sum_{z\in\mathcal{N}_{\rm int}} \varphi_z$. Then φ_z/ψ, $z \in \mathcal{N}_{\rm int}$, represents in Ω a partition of unity. Indeed, since φ_z, $z \in \mathcal{N}_{\rm int}$, sum up to ψ, we have $\sum_{z\in\mathcal{N}_{\rm int}} \varphi_z/\psi = 1$ in Ω; see (Carstensen, 1999, Proposition 2.1). The quasi-interpolation operator $\mathcal{I} : L^1(\Omega) \to V_h$ is then defined in the following way. For a given $w \in L^1(\Omega)$,

$$\mathcal{I}w \equiv \sum_{z\in\mathcal{N}_{\rm int}} w_z\varphi_z\,, \quad \text{where} \quad w_z \equiv \frac{(w, \varphi_z/\psi)}{(1, \varphi_z)}\,.$$

The error $w - \mathcal{I}w$ has a vanishing weighted average. Namely, for w, $R \in L^2(\Omega)$ and arbitrary numbers $R_z \in \mathbb{R}$, $z \in \mathcal{N}_{\rm int}$,

$$\int_\Omega R\,(w - \mathcal{I}w) = \sum_{z\in\mathcal{N}_{\rm int}} \int_\Omega (R - R_z)(w - w_z\psi)(\varphi_z/\psi)\,; \tag{3.1}$$

see (Carstensen, 1999, Remark 2.4). Since

$$\int_\Omega (w - w_z\psi)(\varphi_z/\psi) = 0 \qquad \text{for all } z \in \mathcal{N}_{\rm int}\,,$$

the numbers $R_z \in \mathbb{R}$, can be chosen arbitrarily. In particular, R_z can be chosen as the mean value of R on ω_z. Then $\int_{\omega_z} |R - R_z|^2$ is minimal among all $R_z \in \mathbb{R}$.

The following lemmas are stated and proved in Carstensen (1999) for a more general case. Considering the model problem (2.1), we restrict ourselves to the case $w \in H_0^1(\Omega)$. The multiplicative factors in the lemmas then depend on

I. the shape of ω_z ,

II. the shapes of $\omega_{z\partial\Omega} \equiv (\omega_z \cup \omega_\xi \mid z \in \mathcal{N}_{\rm int}, \xi \in \mathcal{N}\backslash\mathcal{N}_{\rm int}, \omega_z \cap \omega_\xi \neq 0)$,

III. the shape coefficients $(\int_{\omega_z} \varphi_z/|\omega_z| \mid z \in \mathcal{N}_{\mathrm{int}})$, where $|\omega_z|$ stands for the Lebesgue measure of ω_z,

IV. the overlap

$$M_1 \equiv \max_{z \in \mathcal{N}_{\mathrm{int}}} \operatorname{card}\{\xi \in \mathcal{N}\backslash\mathcal{N}_{\mathrm{int}} \mid \omega_z \cap \omega_\xi \neq 0\}\,,$$

V. the shape of the elements $T \in \mathcal{T}$,

VI. the value $\max_{z\in\mathcal{N}} h_z\|\nabla\varphi_z\|_\infty$, where $\|\cdot\|_\infty$ denotes the $L^\infty(\Omega)$-norm and $h_z = \operatorname{diam}(\omega_z)$,

VII. the value

$$M_2 \equiv \operatorname*{ess\,sup}_{x\in\Omega} \{h(x)/h_T \mid x \in T \in \mathcal{T}\}\,,$$

where $h(x) \equiv \max\{h_z \mid \varphi_z(x) > 0, z \in \mathcal{N}_{\mathrm{int}}\}$, $h_T = \operatorname{diam}(T)$.

The proofs of the lemmas use the Poincaré inequality on ω_z and the Friedrichs inequality on $\omega_{z\partial\Omega}$ defined in **II.** In order to prove Lemma 3.2, the so-called trace theorem (see, e.g., (Carstensen, 1999, Proposition 4.1)) is used on each element of the triangulation $T \in \mathcal{T}$; the multiplicative factor then depends on the shape of the elements; see **V.** The quantities $\max_{z\in\mathcal{N}} h_z\|\nabla\varphi_z\|_\infty$ and M_2 (see **VI.** and **VII.**) are of the order one on a shape-regular mesh, where $\|\nabla\varphi_z|_T\|_\infty \approx h_T^{-1}$ and $h_z \approx h_T, T \in \omega_z$. They deteriorate on a mesh consisting of triangles with small inner angles, where small and large elements (in the sense of their diameter) adjoint. In order to see the development of the argument, for clarity we present the following two lemmas.

Lemma 3.1 (see (Carstensen, 1999, Theorem 3.1, statement 1.))**.** *There exists a multiplicative factor $C > 0$ depending on the triangulation $\mathcal{T}$ (more precisely on **I.–IV.**), but not on the size of the elements h_T, such that, for all $R \in L^2(\Omega)$, for all $w \in H_0^1(\Omega)$ and for arbitrary numbers $R_z \in \mathbb{R}$, $z \in \mathcal{N}_{\mathrm{int}}$,*

$$\int_\Omega R\,(w - \mathcal{I}w) \leq C\|\nabla w\| \left\{ \sum_{z\in\mathcal{N}_{\mathrm{int}}} h_z^2 \int_{\omega_z} \varphi_z/\psi\ |R - R_z|^2 \right\}^{1/2} .$$

Lemma 3.1 is a consequence of the definition of the quasi-interpolation operator $\mathcal{I}$; see (3.1).

Lemma 3.2 (see (Carstensen, 1999, Theorem 3.2))**.** *Let $S \subset \mathcal{E}$. There exists a multiplicative factor $C > 0$ depending on the triangulation $\mathcal{T}$ (more precisely on **I.–VII.**), but not on the size of the elements h_T, such that for all $J \in L^2(S)$ and for all $w \in H_0^1(\Omega)$,*

$$\int_S J\,(w - \mathcal{I}w) \leq C\|\nabla w\| \left(\sum_{T\in\mathcal{T}} h_T\|J\|^2_{S\cap\partial T} \right)^{1/2} .$$

Combining Lemmas 3.1 and 3.2 we get the final inequality.

Lemma 3.3 (see (Carstensen, 1999, Corollary 3.1))**.** *Let $S \subset \mathcal{E}$. There exists a multiplicative factor $C_1 > 0$ depending on **I.–VII.** such that, for all $J \in L^2(S)$, for all $R \in L^2(\Omega)$, for all $w \in H_0^1(\Omega)$, and for arbitrary numbers $R_z \in \mathbb{R}$, $z \in \mathcal{N}_{\mathrm{int}}$,*

$$\int_\Omega R\,(w - \mathcal{I}w) + \int_S J\,(w - \mathcal{I}w)$$
$$\leq C_1\|\nabla w\| \left\{ \sum_{z\in\mathcal{N}_{\mathrm{int}}} h_z^2\|R - R_z\|^2_{\omega_z} + \sum_{T\in\mathcal{T}} h_T\|J\|^2_{S\cap\partial T} \right\}^{1/2} .$$

The following lemma introduces a positive multiplicative factor C_{intp} that plays a key role in our discussion on incorporating the algebraic error into the a posteriori bound on the total error; see Section 4 and the numerical experiments in Section 6.

Lemma 3.4 (see (Carstensen, 1999, Theorem 3.1, statement 3.))**.** *There exists a multiplicative factor $C_{\mathrm{intp}} > 0$ depending on the triangulation $\mathcal{T}$ (more precisely on **I.–IV.** and **VI.**) such that, for all $w \in H_0^1(\Omega)$,*

$$\|\nabla\mathcal{I}w\| \leq C_{\mathrm{intp}}\|\nabla w\|\,. \tag{3.2}$$

Remark 3.1. The factor C_{intp} satisfies

$$\sup_{w \in H_0^1(\Omega)} \frac{\|\nabla \mathcal{I} w\|}{\|\nabla w\|} \leq C_{\text{intp}} \,.$$

Obtaining a value of C_{intp} such that the above inequality is tight represents a nontrivial issue. Using the proof of (Carstensen, 2006, Theorem 2.4) and the discussion in (Carstensen, 2006, Example 2.3), we can get a better idea about the size of C_{intp} . For a shape-regular mesh with $\max_{z\in\mathcal{N}} h_z \|\nabla \varphi_z\|_\infty \approx 2$ (see **VI.**), there holds $C_{\text{intp}} \approx 6$. In general, as stated in Carstensen (2006), it may be very large for small angles in the triangulation.

For $f \in L^1(\Omega)$ define the mean-value operator $\pi_{\omega_z}(f) \equiv \int_{\omega_z} f/|\omega_z|$. We denote, for $z \in \mathcal{N}$ and for any subset $Z \subset \mathcal{N}$,

$$\text{osc}_z \equiv |\omega_z|^{1/2} \|f - \pi_{\omega_z} f\|_{\omega_z} \,, \qquad \text{osc}(Z) \equiv \left(\sum_{z \in Z} \text{osc}_z^2 \right)^{1/2} \,,$$

measuring the oscillations of f, i.e. the variations of the function f from the mean value $\pi_{\omega_z} f$ on the subdomains ω_z. Given $v_h \in V_h$, define for $E \in \mathcal{E}_{\text{int}}$ and any subset $F \subset \mathcal{E}_{\text{int}}$ the edge residuals

$$J_E(v_h) \equiv |E|^{1/2} \left\| \left[\frac{\partial v_h}{\partial n_E} \right] \right\|_E \,, \qquad J(v_h, F) \equiv \left(\sum_{E \in F} J_E^2(v_h) \right)^{1/2} \,,$$

where $[\cdot]$ denotes the jump of a piecewise continuous function and n_E denotes the unit normal to E (for each $E \in \mathcal{E}_{\text{int}}$ the orientation of the unit normal is set arbitrarily but fixed). The edge residual $J_E(v_h)$, $v_h \in V_h$, measures the jump of the piecewise constant gradient ∇v_h over the inner edge E. We set for brevity $\text{osc} \equiv \text{osc}(\mathcal{N})$ and $J(v_h) \equiv J(v_h, \mathcal{E}_{\text{int}})$. For a given $v_h \in V_h$, we define the jump function $\mathcal{J}(v_h) \in L^2(\mathcal{E}_{\text{int}})$ on the inner edges

$$\mathcal{J}(v_h)|_E \equiv \left[\frac{\partial v_h}{\partial n_E} \right] \,, \quad E \in \mathcal{E}_{\text{int}} \,. \tag{3.3}$$

The Green's formula (see, e.g., (Ciarlet, 2002, p. 14)) gives for a domain D with a Lipschitz continuous boundary ∂D and for $v \in H^2(D)$, $w \in H^1(D)$

$$\int_D \nabla v \cdot \nabla w = -\int_D \Delta v \, w + \int_{\partial D} \left(\frac{\partial v}{\partial n_{\partial D}} \right) w \,, \tag{3.4}$$

where $n_{\partial D}$ denotes the unit normal to ∂D pointing outwards. Let $v_h \in V_h$ and $T \in \mathcal{T}$. Then $v_h|_T$ is a linear function, $v_h|_T \in H^2(T)$ and $\Delta v_h|_T = 0$. Then, applying the Green's formula (3.4) elementwise yields, for any $v_h \in V_h$, $w \in H_0^1(\Omega)$,

$$\begin{aligned} \int_\Omega \nabla v_h \cdot \nabla w &= \sum_{T \in \mathcal{T}} \int_T \nabla v_h \cdot \nabla w = \sum_{T \in \mathcal{T}} \left(-\int_T \Delta v_h \, w + \int_{\partial T} \left(\frac{\partial v_h}{\partial n_{\partial T}} \right) w \right) \\ &= \sum_{E \in \mathcal{E}_{\text{int}}} \int_E \left[\frac{\partial v_h}{\partial n_E} \right] w = \int_{\mathcal{E}_{\text{int}}} \mathcal{J}(v_h) \, w \,. \end{aligned} \tag{3.5}$$

The results recalled in this section are used to prove the upper bound on the total error in the next section.

4 Upper bound on the total error

Now we state the upper bound on the energy norm of the total error using the residual-based a posteriori error estimator.

Theorem 4.1. *There exist triangulation-dependent positive multiplicative factors C_1, $C_{\rm intp}$, and C_2 such that for the solution u of* (2.2)*, the Galerkin solution u_h of (2.3), and an arbitrary $v_h \in V_h$,*

$$\|\nabla(u-v_h)\|^2 \le 2\,C_1^2\,C_2^2\left(J^2(v_h)+\mathrm{osc}^2\right)+2\,C_{\rm intp}^2\|\nabla(u_h-v_h)\|^2\,. \tag{4.1}$$

In particular, C_1 depends on ***I.–VII.*** *(see Lemma 3.3), $C_{\rm intp}$ depends on* ***I.–IV.*** *and* ***VI.*** *(see Lemma 3.4), and the factor C_2 depends on the ratios $h_z^2/|\omega_z|$, $z\in\mathcal{N}_{\rm int}$, and $h_T/|E|$, $T\in\mathcal{T}$, $E\in\partial T\cap\mathcal{E}_{\rm int}$.*

Proof. We will use the standard expression for the norm

$$\|\nabla(u-v_h)\| = \sup_{0\ne w\in H_0^1(\Omega)}\frac{1}{\|\nabla w\|}\int_\Omega \nabla(u-v_h)\cdot\nabla w\,. \tag{4.2}$$

Let $v_h\in V_h$ and $w\in H_0^1(\Omega), w\ne 0$, be arbitrary,

$$\begin{aligned}\int_\Omega \nabla(u-v_h)\cdot\nabla w &= \int_\Omega \nabla(u-v_h)\cdot\nabla(w-\mathcal{I}w)+\int_\Omega\nabla(u-v_h)\cdot\nabla\mathcal{I}w\\ &= \int_\Omega \nabla(u-v_h)\cdot\nabla(w-\mathcal{I}w)+\int_\Omega\nabla(u-u_h)\cdot\nabla\mathcal{I}w\\ &\quad+\int_\Omega\nabla(u_h-v_h)\cdot\nabla\mathcal{I}w\,.\end{aligned}$$

It follows from the definition of the interpolation operator that $\mathcal{I}w\in V_h$. The Galerkin orthogonality (2.4) gives

$$\int_\Omega\nabla(u-u_h)\cdot\nabla\mathcal{I}w=0.$$

Then

$$\begin{aligned}\int_\Omega \nabla(u-v_h)\cdot\nabla w &= \int_\Omega \nabla(u-v_h)\cdot\nabla(w-\mathcal{I}w)+\int_\Omega\nabla(u_h-v_h)\cdot\nabla\mathcal{I}w\\ &= \int_\Omega \nabla u\cdot\nabla(w-\mathcal{I}w)-\int_\Omega\nabla v_h\cdot\nabla(w-\mathcal{I}w)\\ &\quad+\int_\Omega\nabla(u_h-v_h)\cdot\nabla\mathcal{I}w\,.\end{aligned}$$

Using the weak formulation (2.2) and the equality (3.5),

$$\int_\Omega\nabla(u-v_h)\cdot\nabla w=\int_\Omega f(w-\mathcal{I}w)-\int_{\mathcal{E}_{\rm int}}\mathcal{J}(v_h)(w-\mathcal{I}w)+\int_\Omega\nabla(u_h-v_h)\cdot\nabla\mathcal{I}w\,.$$

Then Lemma 3.3 with $S=\mathcal{E}_{\rm int}$, $R=f$, $R_z=\pi_{\omega_z}f$, $z\in\mathcal{N}_{\rm int}$, $J=-\mathcal{J}(v_h)$ gives

$$\begin{aligned}\int_\Omega\nabla(u-v_h)\cdot\nabla w &\le C_1\|\nabla w\|\left\{\sum_{T\in\mathcal{T}}h_T\|\mathcal{J}(v_h)\|^2_{\mathcal{E}_{\rm int}\cap\partial T}+\sum_{z\in\mathcal{N}_{\rm int}}h_z^2\|f-\pi_{\omega_z}f\|^2_{\omega_z}\right\}^{\frac12}\\ &\quad+\int_\Omega\nabla(u_h-v_h)\cdot\nabla\mathcal{I}w\\ &\le C_1C_2\|\nabla w\|\left\{\sum_{E\in\mathcal{E}_{\rm int}}|E|\,\|\mathcal{J}(v_h)\|^2_E+\sum_{z\in\mathcal{N}_{\rm int}}|\omega_z|\,\|f-\pi_{\omega_z}f\|^2_{\omega_z}\right\}^{\frac12}\\ &\quad+\int_\Omega\nabla(u_h-v_h)\cdot\nabla\mathcal{I}w\\ &= C_1C_2\|\nabla w\|\left(J^2(v_h)+\mathrm{osc}^2\right)^{1/2}+\int_\Omega\nabla(u_h-v_h)\cdot\nabla\mathcal{I}w\,.\end{aligned}$$

Using the Cauchy-Schwarz inequality,

$$\int_\Omega\nabla(u-v_h)\cdot\nabla w\le C_1C_2\|\nabla w\|\left(J^2(v_h)+\mathrm{osc}^2\right)^{1/2}+\|\nabla\mathcal{I}w\|\,\|\nabla(u_h-v_h)\|\,. \tag{4.3}$$

Dividing (4.3) by $\|\nabla w\|$ and using Lemma 3.4,

$$\begin{aligned} \frac{1}{\|\nabla w\|}\int_\Omega \nabla(u-v_h)\cdot\nabla w &\leq C_1C_2\left(J^2(v_h)+\mathrm{osc}^2\right)^{1/2}+\frac{\|\nabla\mathcal{I}w\|}{\|\nabla w\|}\|\nabla(u_h-v_h)\| \\ &\leq C_1C_2\left(J^2(v_h)+\mathrm{osc}^2\right)^{1/2}+C_{\mathrm{intp}}\|\nabla(u_h-v_h)\|\,. \end{aligned}$$

Using the representation (4.2) of the energy norm, recall that $w\in H_0^1(\Omega)$ was chosen arbitrarily,

$$\|\nabla(u-v_h)\|\leq C_1C_2\left(J^2(v_h)+\mathrm{osc}^2\right)^{1/2}+C_{\mathrm{intp}}\|\nabla(u_h-v_h)\|\,. \tag{4.4}$$

Finally, we deduce (4.1) using the inequality $(a+b)^2\leq 2a^2+2b^2$. □

The multiplicative factor C_{intp} is independent of the solution $u\in H_0^1$ of (2.2) [2] and of its approximation $v_h\in V_h$. We will now give another bound on the total error that will contain information about the exact solution u of (2.2). It is useful for illustration of the role of the factor C_{intp} in (4.1). Setting $w\equiv u-v_h\in H_0^1(\Omega)$ in (4.3), we get

$$\|\nabla(u-v_h)\|^2\leq C_1C_2\|\nabla(u-v_h)\|\left(J^2(v_h)+\mathrm{osc}^2\right)^{1/2}+\|\nabla(\mathcal{I}u-\mathcal{I}v_h)\|\,\|\nabla(u_h-v_h)\|\,.$$

Dividing both sides by $\|\nabla(u-v_h)\|$ gives

$$\|\nabla(u-v_h)\|\leq C_1C_2\left(J^2(v_h)+\mathrm{osc}^2\right)^{1/2}+\frac{\|\nabla(\mathcal{I}u-\mathcal{I}v_h)\|}{\|\nabla(u-v_h)\|}\,\|\nabla(u_h-v_h)\|\,, \tag{4.5}$$

which is formulated as the following corollary.

Corollary 4.1. Using the notation of Theorem 4.1,

$$\|\nabla(u-v_h)\|^2\leq 2\,C_1^2\,C_2^2\left(J^2(v_h)+\mathrm{osc}^2\right)+2\,\widetilde{C}_{\mathrm{intp}}^2(u,v_h)\,\|\nabla(u_h-v_h)\|^2\,, \tag{4.6}$$

where

$$\widetilde{C}_{\mathrm{intp}}(u,v_h)\equiv\frac{\|\nabla(\mathcal{I}u-\mathcal{I}v_h)\|}{\|\nabla(u-v_h)\|}\leq C_{\mathrm{intp}}\,. \tag{4.7}$$

Since C_{intp} is independent of $u\in H_0^1$ (independent of the source term f) and of $v_h\in V_h$, we must have

$$\sup_{u\in H_0^1,\ u \text{ solves } (2.2),\ f\in L^2(\Omega)}\ \sup_{v_h\in V_h}\frac{\|\nabla(\mathcal{I}u-\mathcal{I}v_h)\|}{\|\nabla(u-v_h)\|}\leq C_{\mathrm{intp}}\,. \tag{4.8}$$

Therefore C_{intp} represents a worst-case scenario factor and one may expect that most likely $\widetilde{C}_{\mathrm{intp}}(u,v_h)\ll C_{\mathrm{intp}}$.

5 A related result

In Arioli *et al.* (2013b) the authors consider elliptic self-adjoint problems and they use a residual-based error estimator for setting the stopping criterion for the conjugate gradient method (CG). Following their approach, one can easily get a theoretical upper bound on the total error. Although the bound is not stated explicitly in Arioli *et al.* (2013b), it appears in the proof of Theorem 3.3; see the inequality (Arioli *et al.*, 2013b, (3.22)). The derivation proceeds differently from the proof of the bound (4.1); it again demonstrates that the price to be paid for removing the assumption on exact algebraic computations in terms of including unknown and possibly large multiplicative factors can be high. Moreover, in contrast to the statements in Arioli *et al.* (2013b), the resulting estimator cannot be considered a guaranteed practical upper bound due to the difficulties in estimating the algebraic part of the error; see Section 7 below.

[2] In the setting of this paper it means independent of the source term $f\in L^2(\Omega)$.

First, (Arioli *et al.*, 2013b, Theorem 2.2) recalls the bound on the discretization error: there exists a multiplicative factor $C_{2.2} > 0$ that is dependent on the minimal angle of the triangulation $\mathcal{T}$ but independent of h, u, and u_h, such that

$$\|\nabla(u-u_h)\|^2 \leq C_{2.2}\,\eta^2(u_h)\,, \qquad \eta(u_h) \equiv \left(\sum_{T\in\mathcal{T}} |T|\,\|f+\Delta u_h\|_T^2 + (J(u_h))^2\right)^{1/2}; \tag{5.1}$$

see, e.g., (Verfürth, 1996, Section 1.2) and (Ainsworth & Oden, 2000, Section 2.2). Using inverse estimates for piecewise polynomial functions and Young's inequality, (Arioli *et al.*, 2013b, Lemma 3.1) yields the inequality

$$\eta^2(w_h) \leq (1+\gamma)\,\eta^2(v_h) + C_{3.1}(1+\gamma^{-1})\|\nabla(v_h-w_h)\|^2\,, \quad \text{for all } w_h, v_h \in V_h,\ \gamma > 0\,,$$

where the positive factor $C_{3.1}$ depends on the minimal angle of the triangulation $\mathcal{T}$. Combining these bounds and the equality

$$\|\nabla(u-v_h)\|^2 = \|\nabla(u-u_h)\|^2 + \|\nabla(u_h-v_h)\|^2$$

that follows from the Galerkin orthogonality (2.4), we get the upper bound on the total error

$$\begin{aligned}\|\nabla(u-v_h)\|^2 &\leq C_{2.2}\,\eta^2(u_h) + \|\nabla(u_h-v_h)\|^2\\ &\leq C_{2.2}\,(1+\gamma)\,\eta^2(v_h) + \big(1 + C_{2.2}\,C_{3.1}\,(1+\gamma^{-1})\big)\,\|\nabla(u_h-v_h)\|^2\end{aligned}$$

Finally, by setting $\gamma \equiv 1$,

$$\|\nabla(u-v_h)\|^2 \leq 2\,C_{2.2}\,\eta^2(v_h) + (1+2\,C_{2.2}\,C_{3.1})\,\|\nabla(u_h-v_h)\|^2\,. \tag{5.2}$$

Comparing (5.1) with (5.2) we see that replacing the Galerkin solution u_h by an arbitrary $v_h \in V_h$ results in the additional term $\|\nabla(u_h-v_h)\|^2$, which is, however, multiplied by an unknown and potentially very large multiplicative factor $(1+2\,C_{2.2}\,C_{3.1})$. In the criteria proposed in (Arioli *et al.*, 2013b, Section 5) for numerical experiments the factors are empirically set to $C_{2.2} \equiv 40$, $C_{3.1} \equiv 10$, giving $(1+2\,C_{2.2}\,C_{3.1}) = 801$; cf. (2.12). This nicely underlines the subtleties of the residual-based bounds discussed above. In addition, as mentioned above and explained in detail in Section 7 below, getting a tight practical upper bound on $\|\nabla(u_h-v_h)\|^2$ represents an unresolved challenge.

6 Numerical illustrations

We use, *on purpose*, very simple problems to illustrate the possible difference in the values of $\widetilde{C}_{\text{intp}}(u,v_h)$ and C_{intp}. While $\widetilde{C}_{\text{intp}}(u,v_h)$ can be, assuming the knowledge of the exact solution u, evaluated up to a negligible quadrature error, for the factor C_{intp} we present a lower bound given by plugging a chosen function into (3.2). The derivation of a more accurate estimate for C_{intp} (see also the discussion in Remark 3.1) is beyond the scope of this paper.

6.1 Numerical illustration in one dimension

We first consider a one-dimensional analogue of the Clément-type quasi-interpolation operator $\mathcal{I}$ to illustrate that C_{intp} can be significantly larger than one.

Consider the domain $\Omega = (0,1)$ with the (non-uniform) partition

$$[0, \beta, 1/3 \pm \beta, 2/3 \pm \beta, 1-\beta, 1]; \quad \beta = 0.01.$$

This partition is adapted to the 1D Laplace problem with the solution

$$\begin{aligned}u(x) \;&=\; \tan^{-1}(cx) - \tan^{-1}(c(x-1/3)) + \tan^{-1}(c(x-2/3))\\ &\quad - \tan^{-1}(c(x-1)) - \tan^{-1}(c/3) + \tan^{-1}(2c/3) - \tan^{-1}(c)\,,\end{aligned} \tag{6.1}$$

with $c = 1000$. The left part of Figure 1 depicts the solution u and the Clément-type quasi-interpolant $\mathcal{I}u$. For a zero approximate vector, we have

$$\widetilde{C}_{\mathrm{intp}}(u, 0) = \frac{\|(\mathcal{I}u)'\|}{\|u'\|} = 0.77. \tag{6.2}$$

For a quadratic function $w(x) = x(1 - x)$, $w \in H_0^1(\Omega)$, we have

$$\frac{\|(\mathcal{I}w)'\|}{\|w'\|} = 3.70;$$

see the right part of Figure 1 for the plot of the function w and the interpolant $\mathcal{I}w$. Consequently, $C_{\mathrm{intp}} \geq 3.70$.

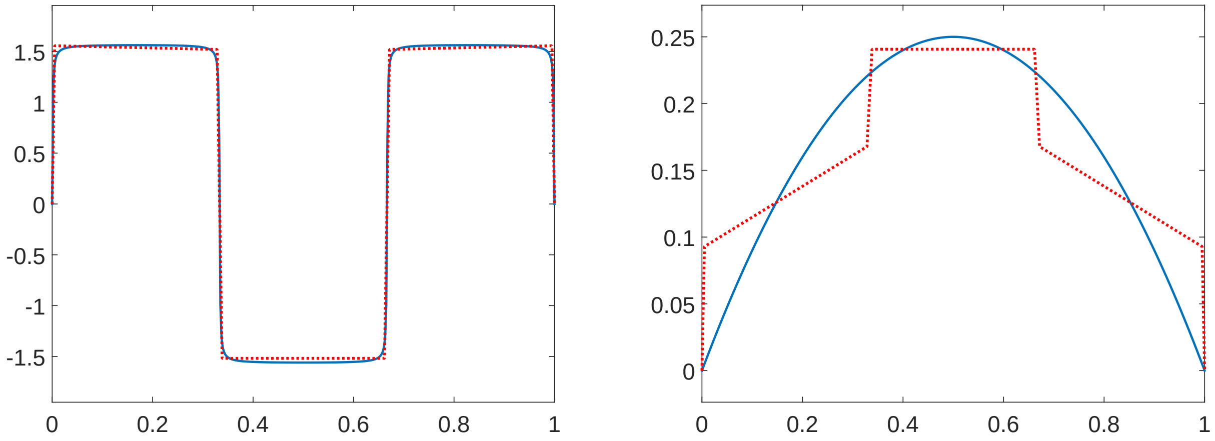

Figure 1: Left: the solution u (6.1) (solid line) and the interpolant $\mathcal{I}u$ (dotted line). Right: the function $w(x) = x(1 - x)$ (solid line) and $\mathcal{I}w$ (dotted line).

6.2 Two-dimensional numerical illustration

For two-dimensional numerical illustration we consider the square domain $\Omega \equiv (-1, 1) \times (-1, 1)$ and the triangulation $\mathcal{T}$ generated by MATLAB[3] command `initmesh('squareg', 'Hmax', 0.1)` that provides a Delaunay triangulation consisting of 1 368 elements with the maximal diameter less than or equal to 0.1. The minimal angle of the mesh is 35.9° and the average of the minimal angles of the elements is 50.3°.

Consider the solution of problem (2.1):

$$u^{(1)}(x, y) = (x - 1)(x + 1)(y - 1)(y + 1)\,. \tag{6.3}$$

For the zero approximate solution and the Galerkin solution $u_h^{(1)}$ corresponding to $u^{(1)}$, we have

$$\widetilde{C}_{\mathrm{intp}}(u^{(1)}, 0) = 1.02\,, \qquad \widetilde{C}_{\mathrm{intp}}(u^{(1)}, u_h^{(1)}) = 0.16\,.$$

Similarly, for the exact solution

$$u^{(2)}(x, y) = (x - 1)(x + 1)(y - 1)(y + 1)\exp(-100(x^2 + y^2))\,, \tag{6.4}$$

we have

$$\widetilde{C}_{\mathrm{intp}}(u^{(2)}, 0) = 0.76\,, \qquad \widetilde{C}_{\mathrm{intp}}(u^{(2)}, u_h^{(2)}) = 0.28\,.$$

In Figure 2 we show the values of $\widetilde{C}_{\mathrm{intp}}(u^{(j)}, v_h)$ for v_h generated in CG iterations with zero initial vector for solving the linear algebraic systems corresponding to the discretization of (2.2) with the solutions $u^{(1)}$, $u^{(2)}$ defined above.

[3] using the Partial Differential Equation Toolbox

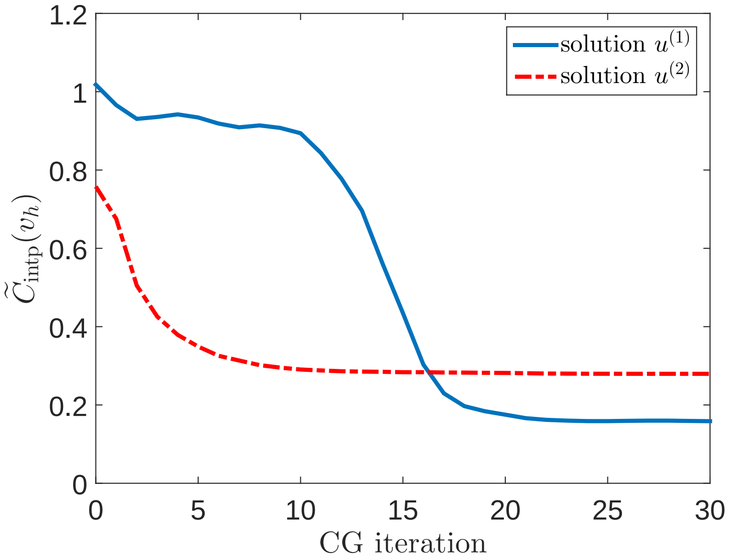

Figure 2: The values of $\widetilde{C}_{\text{intp}}(u^{(j)}, v_h)$ for v_h generated in conjugate gradient iterations with zero initial vector for solving the linear algebraic systems corresponding to the discretization of (2.2) with the solutions $u^{(1)}$, $u^{(2)}$; see (6.3) and (6.4) respectively.

To bound the constant C_{intp} from below we consider $w_h \in V_h$ such that

$$w_h(z) = 1, \quad z \in \mathcal{N}_{\text{int}}, \qquad w_h = 0 \quad \text{on } \partial\Omega \,. \tag{6.5}$$

For this function

$$1.10 = \frac{\|\nabla \mathcal{I} w_h\|}{\|\nabla w_h\|} \leq C_{\text{intp}} \,.$$

Figure 3 gives the difference $w_h - \mathcal{I} w_h$. This is a piecewise linear function that is on the machine precision level in most of the domain except patches around the inner nodes adjacent to the boundary $\partial\Omega$. We recall that for this simple problem and a shape-regular mesh $C_{\text{intp}} \approx 6$; see Remark 3.1 and the original paper Carstensen (2006). It can therefore indeed hold $C_{\text{intp}} \gg \widetilde{C}_{\text{intp}}(u, v_h)$.

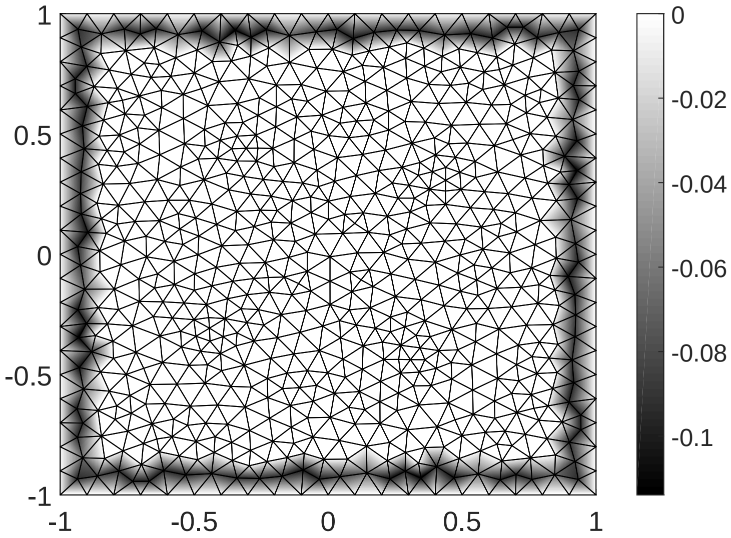

Figure 3: The difference $w_h - \mathcal{I} w_h$ for w_h given by (6.5).

7 Algebraic error estimates

Various published papers that include algebraic errors in a posteriori error analysis consider preconditioned iterative methods (such as the conjugate gradient method) for solving the associate symmetric positive definite algebraic problem. Estimating the algebraic error $\|\nabla(u_h - u_h^C)\|$ is then often considered to be resolved, apart from some seemingly technical issues such as the choice of a number of extra CG iterations or the approximation of the smallest eigenvalue of the system matrix; see, e.g., (Arioli *et al.*, 2013b, Section 4), where the presented Gauss–Radau quadrature-based bound is considered "the only guaranteed upper bound for the A-norm of the CG error." Here we present mathematical arguments that challenge this opinion. In short, as above, theoretical results with assumptions that do not hold in practical computations can not be applied without an appropriate *theoretical justification*, here the *numerical stability analysis*. In practice the Gauss–Radau quadrature-based estimate does not give, in general, a guaranteed and tight upper bound.

The Gauss–Radau quadrature requires an approximation of the smallest eigenvalue of the system matrix *from below*. A close approximation to the smallest eigenvalue *from above* can be in theory obtained in the process of (exact arithmetic) CG computations. A closer investigation reveals, however, a serious difficulty. If the approximation to the smallest eigenvalue is not accurate enough, the estimate for the norm of the algebraic error can be inaccurate. If, on the other hand, the eigenvalue approximation approaches the smallest eigenvalue, then the computation of the Gauss–Radau estimator must be cautiously done in a numerically stable way since it (implicitly) inverts a matrix that becomes in such case close to numerically singular. Moreover, if a second approximation of the smallest eigenvalue is formed due to loss of linear independence of the computed vectors generating the associated Krylov subspaces, then the Gauss–Radau quadrature estimates exhibit instabilities that are not yet understood. We can see again the generally valid principle: *In evaluation of a posteriori error estimates all sources of errors, including roundoff, must be taken into consideration.* This is also illustrated by the following point, which is valid even if the difficulty with a tight approximation of the smallest eigenvalue from below is resolved. In numerical computations we can not guarantee that Gauss–Radau quadrature estimates give an upper bound due to effects of roundoff errors to the underlying short recurrences in CG computations. This is a *principal issue* as explained next.

Due to the loss of orthogonality caused by roundoff errors, the formulas that express the error norms in CG iterations under the assumption of exact arithmetic and, consequently, orthogonality among the computed basis vectors *are no longer valid.* In practical computations we do not have orthogonal bases nor we have Krylov subspaces in the strict mathematical meaning. Unless the problem is really computationally simple (as in the Poisson model problem)[4], orthogonality and also linear independence among the vectors computed using short recurrences is, in general, rather quickly lost and convergence of the computed iterations is substantially delayed. In terms of the computed Jacobi matrices that are substantial in the derivation of the Gauss–Radau quadrature bounds, their entries will quickly become far away (*orders of magnitude*) from their theoretical counterparts. Recent description of the related issues can be found in the paper Gergelits & Strakoš (2014) with the references to many earlier papers; see, e.g., Paige (1980), Greenbaum (1989), Strakoš (1991), Greenbaum & Strakoš (1992), Notay (1993), Strakoš & Tichý (2002), Strakoš & Tichý (2005), Meurant & Strakoš (2006), O'Leary *et al.* (2007). The matter has been comprehensively discussed already in the survey paper Strakoš & Liesen (2005) and it is also covered in the recent monograph (Liesen & Strakoš, 2013, Section 5.9).

In summary, the following principal question is omitted in works assuming exact arithmetic: Considering the effects of roundoff recalled above, *how do we know that the formulas derived under the assumptions that are so drastically violated give anything meaningful in practical computations?* This fundamental issue can not be resolved by heuristic arguments; it requires thorough analysis.

The question on the effects of roundoff errors to error estimation in iterative methods such as CG has already been raised in the seminal paper Dahlquist *et al.* (1979). It has been investigated in Section 5 (called "Rounding error analysis") of the paper Golub & Strakoš (1994), and again very thoroughly (in the context of different estimates) in the paper Strakoš & Tichý (2002) with the title "On error estimation in the conjugate gradient method and why it works in finite precision computations"; see also the survey

[4] Simple model problems can not be used for verification of computational efficiency and numerical behaviour of methods and algorithms. Extrapolation of the results observed on simple model problems to difficult practical problems can lead to false conclusions, because some phenomena (such as significant loss of orthogonality and delay of convergence in CG) are on model problems (such as the Poisson boundary value problem) simply not observable.

paper Meurant & Strakoš (2006). This underlines the point.

Without thorough rounding error analysis, we may say that we have *observed* some behaviour on some examples, and *nothing more*. With thorough rounding error analysis, we can explain why and under which conditions some estimates work and prove them numerically safe. Perhaps even more important, we can prove that other (mathematically equivalent) estimates can behave in a numerically unstable way and they should not be used. Estimates that has been proved numerically unstable (see, e.g., Strakoš & Tichý (2002)) are unfortunately indeed frequently used in practice. The facts presented, e.g., in (Golub & Strakoš, 1994, Section 5), (Strakoš & Tichý, 2002, Section 7), (Meurant, 2006, Chapter 7), (Meurant & Strakoš, 2006, Section 5), O'Leary *et al.* (2007), and (Liesen & Strakoš, 2013, Section 5.9) show that without a thorough and rigorous analysis, application of the error bounds, derived under the assumption of exact computation, to the results of finite precision computations, is not only methodologically wrong but it can indeed lead to false conclusions.

A partial theoretical justification of the Gauss and Gauss–Radau quadrature estimates is provided (based on the earlier works of Paige, Greenbaum and others) in Section 5 of the paper Golub & Strakoš (1994). In short, justification of the quadrature-based bounds must be based on the Riemann–Stieltjes integration using the distribution function with possibly many more points of increase than the original distribution function determined by the data of the problem; see also the associated arguments in the paper O'Leary *et al.* (2007) on sensitivity of the Gauss quadrature. The paper Golub & Strakoš (1994) mentioned above justifies using Gauss and Gauss–Radau quadrature estimates in finite precision computations (with limitations specified in the paper). The estimates based on the Gauss–Radau quadrature can be useful but they *can not be proved to give a tight guaranteed upper bound on the error norms.*

Summarizing, accurate and computationally efficient estimation of the algebraic error $\|\nabla(u_h - u_h^C)\|$ still represents a challenge. This challenge is not resolved by deriving estimators assuming exact computations and then plugging in the computed quantities. The fact that some estimators can be used *in the same form* for input entries computed using finite precision arithmetic is not at all obvious and it can not be guessed a priori by any heuristics. This fact can only result from a rigorous analysis that is (in these cases) highly nontrivial.

8 Conclusion

The presented paper investigates changes in derivation and application of the standard residual-based a posteriori error bound on the discretization error needed in order to get an estimator (not necessarily a practically applicable guaranteed upper bound) for the total error. Technically, this paper provides a detailed proof of the residual-based upper bound on the total approximation error that requires knowledge of the associated multiplicative factors. As published previously in Becker & Mao (2009) and Arioli *et al.* (2013b), abandoning the Galerkin orthogonality assumption in the derivation leads naturally to an additional term accounting for the algebraic part of the error.

We show that the contribution of the algebraic error is scaled by a nontrivial multiplicative factor; see (4.1) and (4.5)–(4.7). This multiplicative factor $\widetilde{C}_{\text{intp}}(u, v_h)$ depends, besides the computed approximation v_h, also on the unknown infinite-dimensional solution u of (2.2). Therefore it is generally uncomputable. It can be bounded, using the *a priori* information, by the factor C_{intp} independent of u and v_h given by (3.2). The value of C_{intp} can therefore overestimate the value of $\widetilde{C}_{\text{intp}}(u, v_h)$. Moreover, as another substantial point, a guaranteed computationally efficient upper bound on the algebraic part of the error that provides, in general, sufficiently tight results is not available.

The main message of this paper does not concern technical details about multiplicative factors in the residual-type a posteriori error estimators for the total error developed for the given model problem. It shows that handling inexact algebraic computations still needs, despite many remarkable results, further work. In practical computations we can not avoid using various heuristics. This paper supports using heuristics (they are used in many papers coauthored by the authors of this paper). It points out, however, a need for supporting analysis. It shows that even for the simple model problem and the standard residual-based a posteriori error estimator the matter is not easy and the unresolved questions can be practically important. As, e.g., numerically illustrated in (Papež, 2016, Section 4.2) and as mentioned in the Introduction, application of the residual-based error estimator for the mesh refinement adaptivity remains in the presence of algebraic errors an open problem. When the standard estimator for the discretization error is evaluated using computed quantities, there is no guaranty that its local contributions provide a meaningful indication of the spatial distribution of the *discretization* error over the domain. The matter is practically very important because it can affect efficiency of h-adaptive computations. Extrapolation of the observations obtained for simple model problems can not be considered the final and generally valid justification. The fact that in (4.1) and (4.6) the algebraic error is estimated globally, and the multiplicative factors C_{intp} and $\widetilde{C}_{\mathrm{intp}}(u, v_h)$ are not easy to estimate, suggests that the matter is intriguing and it requires further work.

This paper can not survey the previously published and recently developed results towards robust stopping criteria that balance the algebraic and discretization error. This is addressed, e.g., in the works using the hierarchy of subspace splittings recalled in Section 2 (see also Huber *et al.* (2017)) or in (Arioli *et al.*, 2013a, Section 4), Jiránek *et al.* (2010), Papež *et al.* (2014), (Carstensen *et al.*, 2014, Section 7.1), and Papež *et al.* (2016). In Papež *et al.* (2016) it is shown, using the methodology based on flux reconstruction, that such stopping criteria could indeed be constructed and rigorously supported by analysis. It also shows, however, that in practical applications there is a substantial computational cost to be paid, and this cost can even become in some cases excessive. There is still a work to be done. As shown at the example of the residual-based a posteriori error estimator in the presented paper, such work should go hand in hand with analysis. Practical importance of Theorem 4.1 is not in application of the bound (4.1). It is in the warning that such application can be difficult and unjustified heuristics can be misleading. In particular, any use of the adjective “guaranteed” should carefully examine the assumptions under which this adjective holds in practice.

Acknowledgement: The authors are grateful to David Silvester, Endre Süli and an anonymous referee for valuable comments, which improved the text.

References

Ainsworth, M. & Oden, J. T. (2000) *A posteriori error estimation in finite element analysis*. Pure and Applied Mathematics (New York). Wiley-Interscience [John Wiley & Sons], New York, pp. xx+240.

Arioli, M., Liesen, J., Międlar, A. & Strakoš, Z. (2013a) Interplay between discretization and algebraic computation in adaptive numerical solution of elliptic PDE problems. *GAMM-Mitt.*, **36**, 102–129.

Arioli, M., Georgoulis, E. H. & Loghin, D. (2013b) Stopping criteria for adaptive finite element solvers. *SIAM J. Sci. Comput.*, **35**, A1537–A1559.

Babuška, I. & Rheinboldt, W. C. (1978) Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.*, **15**, 736–754.

Becker, R., Johnson, C. & Rannacher, R. (1995) Adaptive error control for multigrid finite element methods. *Computing*, **55**, 271–288.

Becker, R. & Mao, S. (2009) Convergence and quasi-optimal complexity of a simple adaptive finite element method. *M2AN Math. Model. Numer. Anal.*, **43**, 1203–1219.

Bramble, J. H., Pasciak, J. E. & Xu, J. (1990) Parallel multilevel preconditioners. *Numerical analysis 1989 (Dundee, 1989)*. Pitman Res. Notes Math. Ser., vol. 228. Longman Sci. Tech., Harlow, pp. 23–39.

Brenner, S. C. & Scott, L. R. (2008) *The Mathematical Theory of Finite Element Methods*. Texts in Applied Mathematics, vol. 15, third edn. New York: Springer-Verlag, pp. xviii+397.

Carstensen, C. (1999) Quasi-interpolation and a posteriori error analysis in finite element methods. *M2AN Math. Model. Numer. Anal.*, **33**, 1187–1202.

Carstensen, C. (2006) Clément interpolation and its role in adaptive finite element error control. *Partial differential equations and functional analysis*. Oper. Theory Adv. Appl., vol. 168. Birkhäuser, Basel, pp. 27–43.

Carstensen, C., Feischl, M., Page, M. & Praetorius, D. (2014) Axioms of adaptivity. *Comput. Math. Appl.*, **67**, 1195–1253.

Ciarlet, P. G. (2002) *The finite element method for elliptic problems*. Classics in Applied Mathematics, vol. 40. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, pp. xxviii+530. Reprint of the 1978 original [North-Holland, Amsterdam].

Dahlquist, G., Golub, G. H. & Nash, S. G. (1979) Bounds for the error in linear systems. *Semi-infinite programming (Proc. Workshop, Bad Honnef, 1978)*. Lecture Notes in Control and Information Sci., vol. 15. Berlin: Springer, pp. 154–172.

Gergelits, T. & Strakoš, Z. (2014) Composite convergence bounds based on Chebyshev polynomials and finite precision conjugate gradient computations. *Numer. Algorithms*, **65**, 759–782.

Golub, G. H. & Strakoš, Z. (1994) Estimates in quadratic formulas. *Numer. Algorithms*, **8**, 241–268.

Greenbaum, A. (1989) Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences. *Linear Algebra Appl.*, **113**, 7–63.

Greenbaum, A. & Strakoš, Z. (1992) Predicting the behavior of finite precision Lanczos and conjugate gradient computations. *SIAM J. Matrix Anal. Appl.*, **13**, 121–137.

Griebel, M. & Oswald, P. (1995) On the abstract theory of additive and multiplicative Schwarz algorithms. *Numer. Math.*, **70**, 163–180.

Harbrecht, H. & Schneider, R. (2016) A note on multilevel based error estimation. *Comput. Methods Appl. Math.*, **16**, 447–458.

Huber, M., Rüde, U., Strakoš, Z. & Wohlmuth, B. (2017) Error estimators for highly parallel multigrid solver. In preparation.

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

Keilegavlen, E. & Nordbotten, J. M. (2015) Inexact linear solvers for control volume discretizations in porous media. *Comput. Geosci.*, **19**, 159–176.

Liesen, J. & Strakoš, Z. (2013) *Krylov subspace methods: principles and analysis*. Numerical Mathematics and Scientific Computation. Oxford: Oxford University Press.

Málek, J. & Strakoš, Z. (2015) *Preconditioning and the conjugate gradient method in the context of solving PDEs*. SIAM Spotlights, vol. 1. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, pp. x+104.

Meurant, G. (2006) *The Lanczos and Conjugate Gradient Algorithms. From Theory to Finite Precision Computations*. Software, Environments, and Tools, vol. 19. Philadelphia, PA: SIAM, pp. xvi+365.

Meurant, G. & Strakoš, Z. (2006) The Lanczos and conjugate gradient algorithms in finite precision arithmetic. *Acta Numer.*, **15**, 471–542.

Morin, P., Nochetto, R. H. & Siebert, K. G. (2002) Convergence of adaptive finite element methods. *SIAM Rev.*, **44**, 631–658 (2003). Revised reprint of "Data oscillation and convergence of adaptive FEM" [SIAM J. Numer. Anal. **38** (2000), no. 2, 466–488].

Nissen, A., Pettersson, P., Keilegavlen, E. & Nordbotten, J. M. (2015) Incorporating geological uncertainty in error control for linear solvers. *SPE Reservoir Simulation Symposium*. Society of Petroleum Engineers.

Nordbotten, J. M. & Bjørstad, P. E. (2008) On the relationship between the multiscale finite-volume method and domain decomposition preconditioners. *Comput. Geosci.*, **12**, 367–376.

Notay, Y. (1993) On the convergence rate of the conjugate gradients in presence of rounding errors. *Numer. Math.*, **65**, 301–317.

O'Leary, D. P., Strakoš, Z. & Tichý, P. (2007) On sensitivity of Gauss-Christoffel quadrature. *Numer. Math.*, **107**, 147–174.

Oswald, P. (1993) On a BPX-preconditioner for P1 elements. *Computing*, **51**, 125–133.

Paige, C. C. (1980) Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem. *Linear Algebra Appl.*, **34**, 235–258.

Papež, J. (2016) Algebraic error in matrix computations in the context of numerical solution of partial differential equations. *Ph.D. thesis*, Charles University, Prague.

Papež, J., Liesen, J. & Strakoš, Z. (2014) Distribution of the discretization and algebraic error in numerical solution of partial differential equations. *Linear Algebra Appl.*, **449**, 89–114.

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

Rüde, U. (1993a) Fully adaptive multigrid methods. *SIAM J. Numer. Anal.*, **30**, 230–248.

Rüde, U. (1993b) *Mathematical and computational techniques for multilevel adaptive methods*. Frontiers in Applied Mathematics, vol. 13. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, pp. xii+140.

Stevenson, R. (2007) Optimality of a standard adaptive finite element method. *Found. Comput. Math.*, **7**, 245–269.

Strakoš, Z. (1991) On the real convergence rate of the conjugate gradient method. *Linear Algebra Appl.*, **154–156**, 535–549.

Strakoš, Z. & Liesen, J. (2005) On numerical stability in large scale linear algebraic computations. *ZAMM Z. Angew. Math. Mech.*, **85**, 307–325.

Strakoš, Z. & 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.

Strakoš, Z. & Tichý, P. (2005) Error estimation in preconditioned conjugate gradients. *BIT*, **45**, 789–817.

Verfürth, R. (1996) *A Review of a Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*. Advances in Numerical Mathematics Series. WileyTeubner.

Wu, H. & Chen, Z. (2006) Uniform convergence of multigrid V-cycle on adaptively refined finite element meshes for second order elliptic problems. *Sci. China Ser. A*, **49**, 1405–1429.

Xu, J. (1992) Iterative methods by space decomposition and subspace correction. *SIAM Rev.*, **34**, 581–613.