Estimating and localizing the algebraic and total numerical errors using flux reconstructions

Jan Papež, Zdeněk Strakoš, Martin Vohralík

Estimating and localizing the algebraic and total numerical errors using flux reconstructions*

J. Papež[†‡] Z. Strakoš[†] M. Vohralík[§]

April 20, 2018

Abstract

This paper presents a methodology for computing upper and lower bounds for both the algebraic and total errors in the context of the conforming finite element discretization of the Poisson model problem and an arbitrary iterative algebraic solver. The derived bounds do not contain any unspecified constants and allow estimating the local distribution of both errors over the computational domain. Combining these bounds, we also obtain guaranteed upper and lower bounds on the discretization error. This allows to propose novel mathematically justified stopping criteria for iterative algebraic solvers ensuring that the algebraic error will lie below the discretization one. Our upper algebraic and total error bounds are based on locally reconstructed fluxes in $\mathbf{H}(\mathrm{div},\Omega)$, whereas the lower algebraic and total error bounds rely on locally constructed $H^1_0(\Omega)$-liftings of the algebraic and total residuals. We prove global and local efficiency of the upper bound on the total error and its robustness with respect to the approximation polynomial degree. Relationships to the previously published estimates on the algebraic error are discussed. Theoretical results are illustrated on numerical experiments for higher-order finite element approximations and the preconditioned conjugate gradient method. They in particular witness that the proposed methodology yields a tight estimate on the local distribution of the algebraic and total errors over the computational domain and illustrate the associate cost.

Keywords: Numerical solution of partial differential equations, finite element method, a posteriori error estimation, algebraic error, discretization error, stopping criteria, spatial distribution of the error

MSC: 65N15, 65N30, 76M10, 65N22, 65F10.

*This work was supported by the ERC-CZ project LL1202 financed by the MŠMT of the Czech Republic, and by the project 13-06684S of the Grant Agency of the Czech Republic. It has also received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 647134 GATIPOR).

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

[‡]Institute of Computer Science, Czech Academy of Sciences, Pod Vodárenskou věží 2, 182 07 Prague, Czech Republic.

[§]Inria Paris, 2 rue Simone Iff, 75589 Paris, France & Université Paris-Est, CERMICS (ENPC), 77455 Marne-la-Vallée, France.

1 Introduction

Most a posteriori error analyses of numerical approximations of partial differential equations still assume that the discretized algebraic problem is solved *exactly*. This is an unrealistic assumption that cannot be satisfied in large scale numerical computations. There is, fortunately, a growing body of work avoiding it, based on different approaches, see, e.g., [19, 5, 8, 49, 60, 47, 54, 11, 31, 34, 7, 53, 2, 24], the references given in the survey [3, Section 4], and in the monograph [38, Chapter 12]. Despite this development, a rigorous, mathematically justified, cheap, and accurate estimation of the discretization and algebraic errors that would allow for their comparison in *practical computations* is not, in our opinion, a fully solved problem. On the algebraic side, such comparison should include *localization* of the algebraic error. Since the algebraic computation aims at approximating the inverse of the discrete operator with respect to the given right-hand side, the algebraic error is of global nature and its distribution over the computational domain can be very different from the distribution of the discretization error; see, e.g., [45] and the references therein. To point out challenges that *any* approach that aims at mathematically rigorous incorporation of the algebraic error into a posteriori error analysis must consider, we now discuss several ways of how the algebraic error in numerical PDEs is estimated.

The conjugate gradient (CG) method minimizes the energy norm of the algebraic error over the Krylov subspaces associated with a symmetric positive definite matrix $\mathbf{A}$ and the initial residual; see, e.g., [32], [36, Section 2.2]. The estimates for the error of the CG approximations are widely studied; see, e.g., [28, 12, 55, 41], and the references given there. The estimates can be associated with the relationship of CG to the Gauss quadrature; see, e.g., [36, Section 3.5]. We will briefly discuss the upper bound based on the Gauss–Radau quadrature; see [17, 28, 30, 42] and called in [2, p. A1548] "[t]he only guaranteed upper bound for the $\mathbf{A}$-norm of the CG error". Considering a preassigned node λ, $0 < \lambda < \lambda_{\min}(\mathbf{A})$, where $\lambda_{\min}(\mathbf{A})$ is the smallest eigenvalue of the matrix $\mathbf{A}$, the Gauss–Radau quadrature gives indeed, assuming *exact arithmetic*, an upper bound on the energy norm of the algebraic error. In [2, Section 4.2] the Poincaré inequality adaptive approach for bounding $\lambda_{\min}(\mathbf{A})$ from below and setting the value of λ is proposed.

Numerically, however, the situation is very subtle. In short, if $0 < \lambda \ll \lambda_{\min}(\mathbf{A})$, then the Gauss–Radau quadrature bound may largely overestimate the actual error. On the other hand, for λ very close to $\lambda_{\min}(\mathbf{A})$, which can make the upper bound tight, it might be impossible to compute the upper bound to a sufficient accuracy because of numerical instabilities. The *derivation* of the estimate includes (implicitly or explicitly) inversion of the matrix $\lambda\mathbf{I} - \mathbf{T}_i$, where $\mathbf{I}$ stands for the identity matrix and $\mathbf{T}_i$ is the Jacobi matrix associated with the ith CG iteration. For λ very close to $\lambda_{\min}(\mathbf{A}) \leq \lambda_{\min}(\mathbf{T}_i)$, and, at the same time, $\lambda_{\min}(\mathbf{T}_i)$ very close to $\lambda_{\min}(\mathbf{A})$, the matrix $\lambda\mathbf{I} - \mathbf{T}_i$ may become close to numerically singular. It should be emphasized that the numerical difficulty may not be immediately visible from the final formulas giving the bound; see, e.g., [42]. The numerical stability analysis provided in [30] explained that although the estimates based on the relationship of CG with the Gauss–Radau quadrature can be very useful, they cannot be considered generally applicable guaranteed and computable upper bounds for the energy norm of the algebraic error. The meaning of the terms *guaranteed* and *computable* is within numerical

linear algebra restricted only to the cases where the results are justified for all possible input data by a rigorous numerical stability analysis.

Multigrid or, more general, multilevel computations can serve as a second example. Here a standard assumption for a posteriori bounds on the algebraic error, which might require further substantial analysis, is that the algebraic problem on the *coarsest grid* is solved *exactly*; see, e.g., [5, 54]. Moreover, the literature known to the authors does not provide computable upper bounds on the algebraic and the total errors. This topic has recently been addressed in [46]. Alternatively, in the multilevel context the a priori arguments are often used; see the discussion in Section 3.3.

A remarkable early concept relating the algebraic and discretization errors is represented by the Cascadic Conjugate Gradient method; see [19, 52]. In [19], the algebraic error is estimated assuming the superlinear convergence behavior of the CG method in the subsequent iterations, and using several heuristics and empirically chosen parameters. The analysis of [52] relies on the upper bound for the CG method based on Chebyshev polynomials that is typically not descriptive, and its refined version based on composite polynomials may not hold in finite precision computations; see [27]. The CG iterations can exhibit locally the so-called staircase behavior (see [36, Chapter 5]) that makes the analysis difficult.

The general a posteriori error estimation framework of [51] provides a guaranteed upper bound on the total error independent of the algebraic solver. However, the estimates do not generally allow to distinguish and compare the parts of the error corresponding to different sources and seem not suitable for constructing stopping criteria for iterative solvers.

The widely used residual-based error estimators (see, e.g., [54, 6, 2] and the references in [58]) provide upper bounds on the total error (and possibly on its components) with unspecified generic constants that can be of large value. The proposed practical stopping criteria and algorithms then require an empirical choice of these constants. A review of these and other approaches can be found in the survey [3]; see also the discussion in the Introduction of [34].

The presented paper elaborates further on the ideas used in [34] for finite volume discretizations, and a more general framework in [24]; see also their application to discontinuous Galerkin finite element discretizations in [21]. Here we consider the conforming finite element setting and derive an upper bound on the total error that will be proved locally efficient and *polynomial-degree-robust* in the spirit of [9, 25]. All results account for the presence of the algebraic error of an arbitrary iterative solver. The paper newly presents a guaranteed upper bound on the *algebraic error* and thoroughly discusses its relationship to formulas derived purely algebraically. Fast and reliable numerical computations using iterative algebraic solvers rely on meaningful stopping criteria. The stopping criteria from [34, 24] are modified here in order to avoid a possible early stopping that could invalidate the computed results.

The paper is organized as follows. The diffusion model problem considered in the paper and the notation are described in Section 2. In Section 3 we discuss known results on estimating the algebraic error using algebraic worst-case bounds, a priori arguments, and techniques using additional iteration steps of the algebraic solver. Section 4 gives, following previously published results, an upper and a lower bound on the *total error*. In Section 5 we derive new upper and lower bounds on the *algebraic error* and discuss the relationship of

the upper bound to the bounds presented in Section 3. Section 6 is devoted to estimates of the *discretization error* and to discussion of the stopping criteria. We also derive there new mathematically justified stopping criteria balancing the algebraic and discretization errors. We finally illustrate the obtained results numerically in Section 7 and give a concluding discussion in Section 8. We provide the details on the quasi-equilibrated flux reconstruction in Appendix A. The proofs of the global and local efficiency of the presented upper bound on the total error are given in Appendix B.

2 Setting and notation

Let $\Omega \subset \mathbb{R}^d, d = 2, 3$, be a polygonal (polyhedral) domain (open, bounded, and connected set). We consider the Poisson model problem: find $u : \Omega \to \mathbb{R}$ such that

$$-\nabla\cdot(\nabla u) = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega, \tag{2.1}$$

that can be equivalently written as the system of two first order equations for the scalar-valued *potential* u and the vector-valued function called *flux* $\boldsymbol{\sigma} \equiv -\nabla u$,

$$\begin{bmatrix} \nabla & I \\ 0 & \nabla\cdot \end{bmatrix} \begin{bmatrix} u \\ \boldsymbol{\sigma} \end{bmatrix} = \begin{bmatrix} 0 \\ f \end{bmatrix} \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega.$$

Assuming $f \in L^2(\Omega)$, the weak form of the model problem (2.1) is as follows: find $u \in V \equiv H^1_0(\Omega)$ such that

$$(\nabla u, \nabla v) = (f, v) \qquad \forall v \in V, \tag{2.2}$$

where $H^1_0(\Omega)$ denotes the standard Hilbert space of $L^2(\Omega)$ functions whose weak derivatives are in $L^2(\Omega)$ and with trace vanishing on $\partial\Omega$. For $v, w \in L^2(\Omega)$, (v, w) stands for $\int_\Omega v(\mathbf{x})w(\mathbf{x})\,\mathrm{d}\mathbf{x}$ (and similarly in the vector-valued case). Hereafter $\|\cdot\|$ denotes the L^2 norm, $\|w\| \equiv (w, w)^{1/2}$, $w \in L^2(\Omega)$. Owing to (2.2), the flux $\boldsymbol{\sigma}$ is in the space $\mathbf{H}(\mathrm{div}, \Omega)$ of the functions in $[L^2(\Omega)]^d$ with the weak divergence in $L^2(\Omega)$; see, e.g., [16, Section 6.13].

Let $\mathcal{T}_h$ be a simplicial mesh of Ω. We suppose that the mesh is conforming in the sense that, for two distinct elements of $\mathcal{T}_h$, their intersection is either an empty set or a common ℓ-dimensional face, $0 \le \ell \le d-1$. We denote a generic element of $\mathcal{T}_h$ by K and its diameter by h_K. We denote by $\mathbb{P}_p(K)$, $p \ge 0$, the space of pth order polynomial functions on an element K and by $\mathbb{P}_p(\mathcal{T}_h)$ the broken polynomial space spanned by $v_h|_K \in \mathbb{P}_p(K)$ for all $K \in \mathcal{T}_h$.

Let

$$V_h \equiv \left\{ v_h \in \mathbb{P}_p(\mathcal{T}_h) \cap C(\overline{\Omega}) \mid v_h = 0 \text{ on } \partial\Omega \right\} \subset H^1_0(\Omega)$$

be the usual finite element space of continuous, piecewise pth order polynomial functions, $p \ge 1$. The discrete formulation corresponding to the problem (2.2) reads: find $u_h \in V_h$ such that

$$(\nabla u_h, \nabla v_h) = (f, v_h) \qquad \forall v_h \in V_h. \tag{2.3}$$

The (exact) solution u_h of (2.3) satisfies the Galerkin orthogonality

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

Let $\psi_j \in V_h,\ j = 1, \ldots, N$, denote a basis of V_h, $\Psi = \{\psi_1, \ldots, \psi_N\}$. Employing these functions in (2.3) gives rise to the system of linear algebraic equations

$$\mathbf{A}\mathsf{U} = \mathsf{F}, \tag{2.5}$$

where $u_h = \sum_{j=1}^{N} \mathsf{U}_j \psi_j = \Psi\mathsf{U}$, $\mathsf{U} = [\mathsf{U}_j]$ is the vector of unknowns, the system matrix $\mathbf{A} = [\mathbf{A}_{j\ell}]$ is symmetric and positive definite, $\mathbf{A}_{j\ell} = (\nabla\psi_\ell, \nabla\psi_j)$, $j, \ell = 1, \ldots, N$, and the right-hand side vector $\mathsf{F} = [\mathsf{F}_j]$ is given by $\mathsf{F}_j = (f, \psi_j)$, $j = 1, \ldots, N$. Within this model problem setting, we consider an *iterative* algebraic solver approximating the exact solution U of (2.5). At the i-th step, $i = 0, 1, 2, \ldots$, we obtain the approximation $\mathsf{U}^i = [\mathsf{U}^i_j]$ and the algebraic residual vector $\mathsf{R}^i = [\mathsf{R}^i_j]$ with

$$\mathsf{R}^i \equiv \mathsf{F} - \mathbf{A}\mathsf{U}^i. \tag{2.6}$$

By u^i_h we denote the approximation to the solution u of (2.2) determined by the coefficient vector U^i, $u^i_h \equiv \sum_{j=1}^{N} \mathsf{U}^i_j \psi_j = \Psi\mathsf{U}^i$. We also rewrite (2.6) in a functional setting. For this purpose, let a function $r^i_h \in L^2(\Omega)$ be a representation of the algebraic residual vector R^i satisfying

$$(r^i_h, \psi_j) = \mathsf{R}^i_j, \qquad j = 1, \ldots, N. \tag{2.7}$$

Two examples are given in Section 5.1 below. Then (2.6) can be rewritten as

$$(r^i_h, \psi_j) = (f, \psi_j) - (\nabla u^i_h, \nabla\psi_j) \qquad \forall j = 1, \ldots, N \tag{2.8}$$

and, together with (2.3), it also implies

$$(r^i_h, v_h) = (f, v_h) - (\nabla u^i_h, \nabla v_h) = (\nabla(u_h - u^i_h), \nabla v_h) \qquad \forall v_h \in V_h. \tag{2.9}$$

This representation will play the key role in the construction of the estimators below as it allows to bound from above the energy norm of the algebraic error. A function satisfying (2.7) was used for error estimation also in [5]. The construction proposed in Section 5.1 below is different and computationally less costly.

The total error between the exact solution u and the approximate solution u^i_h is measured in the energy norm $\|\nabla(u - u^i_h)\|$. Analogously, the algebraic energy norm of the error $u_h - u^i_h$ is

$$\begin{aligned} \|\nabla(u_h - u^i_h)\| = \|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}} &= \big((\mathsf{U} - \mathsf{U}^i), \mathbf{A}(\mathsf{U} - \mathsf{U}^i)\big)^{1/2} \\ &= (\mathbf{A}^{-1}\mathsf{R}^i, \mathsf{R}^i)^{1/2} = \|\mathsf{R}^i\|_{\mathbf{A}^{-1}}, \end{aligned}$$

where (V, U) denotes the standard inner product of the vectors U and V, $\|\mathsf{V}\| \equiv (\mathsf{V}, \mathsf{V})^{1/2}$ stands for the Euclidean norm of the vector V, and $\|\mathbf{A}\|$ is the induced spectral norm of the matrix $\mathbf{A}$.

3 Algebraic bounds

This section presents some well-known algebraic bounds, with a few comments towards the conjugate gradient method and multilevel methods.

3.1 The L^2 (Euclidean) norm residual bound

The simplest algebraic error upper bound consists in

$$\|\nabla(u_h - u_h^i)\| = \|\mathsf{R}^i\|_{\mathbf{A}^{-1}} \leq \|\mathbf{A}^{-1}\|^{1/2} \cdot \|\mathsf{R}^i\|. \tag{3.1}$$

For a symmetric positive definite matrix, the norm $\|\mathbf{A}^{-1}\|$ is given by the reciprocal of the smallest eigenvalue of the matrix $\mathbf{A}$. It is clear that for $\mathbf{A}$ ill-conditioned, the bound (3.1) can significantly overestimate the algebraic error. Note that equality is attained for a vector R^i collinear with the eigenvector corresponding to the smallest eigenvalue of $\mathbf{A}$.

Even this simplest worst-case bound may not be easy to compute. The smallest eigenvalue of $\mathbf{A}$ is typically not available, and, if it is close to zero, then the cost of its reliable and accurate approximation may not be negligible; see, e.g., [39, 40]. We derive easily computable L^2 norm residual bounds in Section 5.2 below, based on the residual representation r_h^i in (2.7); see the estimates (5.3), (5.4), and (5.8).

3.2 Bounds using additional algebraic iterations

The following simple idea was to our knowledge first presented for algebraic error estimates in [30, pp. 262–263] for the CG method; see also [55, 41]. For estimating the total error it was then used in [34] and in [24], where an arbitrary algebraic solver was considered.

The triangle inequality gives, at the cost of $\nu > 0$ additional iterations,

$$\|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}} \leq \|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}} + \|\mathsf{U} - \mathsf{U}^{i+\nu}\|_{\mathbf{A}} = \|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}} + \|\mathsf{R}^{i+\nu}\|_{\mathbf{A}^{-1}}. \tag{3.2}$$

Assuming that for a given parameter $\gamma > 0$, the choice of ν ensures

$$\|\mathbf{A}^{-1}\|^{1/2} \cdot \|\mathsf{R}^{i+\nu}\| \leq \gamma \|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}}, \tag{3.3}$$

we have, using (3.1), an easily computable *upper bound*

$$\|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}} \leq (1+\gamma)\|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}}. \tag{3.4}$$

Moreover,

$$\|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}} \leq \|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}} + \|\mathsf{U} - \mathsf{U}^{i+\nu}\|_{\mathbf{A}} \leq \|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}} + \gamma\|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}},$$

so that, assuming that $0 < \gamma < 1$, we get the *lower bound*

$$(1-\gamma)\|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}} \leq \|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}}. \tag{3.5}$$

Here (3.4) and (3.5) show that the accuracy of the estimate $\|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}}$ is controlled by the user-specified parameter γ.

We must, however, take into account the following principal issue. If

$$\|\mathsf{U} - \mathsf{U}^{i+\nu}\|_{\mathbf{A}} = \|\mathsf{R}^{i+\nu}\|_{\mathbf{A}^{-1}} \ll \|\mathbf{A}^{-1}\|^{1/2} \cdot \|\mathsf{R}^{i+\nu}\|,$$

the value of ν satisfying (3.3) can be very large. In the worst case, the value of ν can be even comparable with the size of the problem. Such situation is highly improbable in practical problems where preconditioning is used in order to get

a reasonable convergence behavior. Still, for a given parameter γ, the smallest ν_1, respectively ν_2, satisfying

$$\|\mathsf{R}^{i+\nu_1}\|_{\mathbf{A}^{-1}} \leq \gamma \|\mathsf{U}^{i+\nu_1} - \mathsf{U}^i\|_{\mathbf{A}} \quad \text{resp.} \quad \|\mathbf{A}^{-1}\|^{1/2} \cdot \|\mathsf{R}^{i+\nu_2}\| \leq \gamma \|\mathsf{U}^{i+\nu_2} - \mathsf{U}^i\|_{\mathbf{A}}, \tag{3.6}$$

where both sides of the inequalities depend on ν_1 respectively ν_2, can significantly differ with $\nu_1 \ll \nu_2$. Section 7.1 below presents a numerical illustration.

Estimating the algebraic error in the CG method in [30, pp. 262-263] considered performing ν additional iterations and using the relation

$$\|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}}^2 = \|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}}^2 + \|\mathsf{U} - \mathsf{U}^{i+\nu}\|_{\mathbf{A}}^2 = \|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}}^2 + \|\mathsf{R}^{i+\nu}\|_{\mathbf{A}^{-1}}^2 \tag{3.7}$$

that is based on the *global* $\mathbf{A}$*-orthogonality of the CG direction vectors.* The detailed rounding error analysis (see [55, (4.9)], [56, (3.7)] with the reference to the original paper [32]) leads to the following mathematical (exact arithmetic) equivalent of (3.7)

$$\|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}}^2 = (\mu_{\text{alg}}^{\text{CG},i,\nu})^2 + \|\mathsf{R}^{i+\nu}\|_{\mathbf{A}^{-1}}^2. \tag{3.8}$$

This relation can be derived assuming only *local orthogonality* that is well-preserved also in finite precision CG computations as a consequence of enforcing numerically the orthogonality among the consecutive direction vectors and residuals. Therefore (3.8) holds, apart from a small inaccuracy proportional to machine precision, also for the *computed quantities*. The same, however, has not been proved for (3.7).

In [55, 56], it was shown how to compute $\mu_{\text{alg}}^{\text{CG},i,\nu}$ at a negligible cost directly from the coefficients in the CG recurrences; see also [29], [41, Section 5.3]. The resulting lower bound

$$\mu_{\text{alg}}^{\text{CG},i,\nu} \leq \|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}} \tag{3.9}$$

holds until the ratio $\|\mathsf{U}-\mathsf{U}^i\|_{\mathbf{A}}/\|\mathsf{U}-\mathsf{U}^0\|_{\mathbf{A}}$ becomes close to the machine precision (for details see [55, Section 10]), and it is tight providing that the actual energy norm of the error decreases reasonably fast. Analogously to (3.3), assuming (nontrivially) that for a given parameter $\gamma > 0$, the number $\nu > 0$ of additional iteration steps is such that

$$\|\mathbf{A}^{-1}\| \cdot \|\mathsf{R}^{i+\nu}\|^2 \leq \gamma^2 (\mu_{\text{alg}}^{\text{CG},i,\nu})^2,$$

then $\mu_{\text{alg}}^{\text{CG},i,\nu}$ gives (neglecting the terms proportional to machine precision)

$$(\mu_{\text{alg}}^{\text{CG},i,\nu})^2 \leq \|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}}^2 \leq (1+\gamma^2)\,(\mu_{\text{alg}}^{\text{CG},i,\nu})^2. \tag{3.10}$$

In conclusion, the general bounds in (3.4) and (3.5) do not require any additional assumptions. Their value can be determined directly from the computed quantities $\mathsf{U}^i, \mathsf{U}^{i+\nu}$. The bounds for the CG method in (3.10) can be evaluated at almost no cost, but their validity for numerically computed approximations $\mathsf{U}^i, \mathsf{U}^{i+\nu}$ had to be proved using a careful numerical stability analysis. As a reward, which is based on the particular properties of the CG method, we get an improved accuracy of the bounds, with the factor characterizing the gap between the lower and the upper bound reduced from $(1+\gamma)/(1-\gamma)$ in (3.4)–(3.5) to $\sqrt{1+\gamma^2}$ in (3.10).

3.3 A priori arguments in multilevel methods

Convergence of multilevel methods is typically proved using the *a priori* contraction argument

$$\|\mathsf{U} - \mathsf{U}^{i+1}\|_{\mathsf{A}} \leq \gamma \|\mathsf{U} - \mathsf{U}^{i}\|_{\mathsf{A}},$$

where $0 < \gamma < 1$. Then the triangle inequality immediately gives the algebraic error bound

$$\|\mathsf{U} - \mathsf{U}^{i}\|_{\mathsf{A}} \leq \frac{1}{1-\gamma} \|\mathsf{U}^{i+1} - \mathsf{U}^{i}\|_{\mathsf{A}}.$$

Though such bounds with a priori determined constant γ can be useful (see, e.g., [7, (2.17)–(2.18)] and the references therein), we believe, as discussed in the introduction, that *a posteriori* bounds such as that of [5] or its unknown-constant-free improvement in [46] are preferable.

4 Estimating the total error

We give in this section computable upper and lower bounds on the total error. The upper bound based on flux reconstruction following [18, 10, 34, 24, 25] is derived in a form where the component associated with the algebraic error actually turns out to give its upper bound; see Section 5. The lower bound on the total error is given in Section 4.4 using conforming residual reconstruction. We will frequently use the following representation of the energy norm of the total error

$$\|\nabla(u - u_h^i)\| = \sup_{v \in V,\, \|\nabla v\| = 1} (\nabla(u - u_h^i), \nabla v). \tag{4.1}$$

4.1 Concept of the flux reconstructions

The motivation for our approach is to mimic the continuous world, where (using (4.1), (2.2), the Green theorem, and the Cauchy–Schwarz inequality),

$$\begin{aligned}\|\nabla(u - u_h^i)\| &= \inf_{\mathbf{d} \in \mathbf{H}(\mathrm{div},\Omega),\, \nabla\cdot\mathbf{d} = f} \sup_{v \in V,\, \|\nabla v\| = 1} \{(f - \nabla\cdot\mathbf{d}, v) - (\nabla u_h^i + \mathbf{d}, \nabla v)\} \\ &= \inf_{\mathbf{d} \in \mathbf{H}(\mathrm{div},\Omega),\, \nabla\cdot\mathbf{d} = f} \|\nabla u_h^i + \mathbf{d}\|;\end{aligned}$$

the equality occurs for $\mathbf{d} = \boldsymbol{\sigma} = -\nabla u$. We also wish to use an upper bound on the algebraic error based on the representation r_h^i. This allows to relate the algebraic and discretization error components.

Practically, a *reconstructed flux* is a piecewise polynomial function in the Raviart–Thomas–Nédélec subspace $\mathbf{V}_h$ of the infinite-dimensional space $\mathbf{H}(\mathrm{div},\Omega)$. It is constructed in an inexpensive *local* way, around each node of the mesh $\mathcal{T}_h$, and it satisfies, on each iteration step $i \geq 1$,

$$\nabla\cdot\mathbf{d}_h^i = f_h - r_h^i. \tag{4.2}$$

Here f_h is a piecewise polynomial approximation of the source term f satisfying

$$(f - f_h, 1)_K = 0 \qquad \forall K \in \mathcal{T}_h. \tag{4.3}$$

The precise definition of the space $\mathbf{V}_h$ and the detailed construction of $\mathbf{d}_h^i$ following [24, Section 6.2.4] are given in Appendix A.

4.2 Upper bound using the L^2 norm of the algebraic residual representation

Similarly to Section 3.1, to illustrate the ideas, we first present a simple upper bound on the total error following [34, Section 7.1]. It typically yields a large overestimation. It follows from (4.1), the weak formulation (2.2), the construction (4.2), and the Green theorem that

$$\|\nabla(u-u_h^i)\| = \sup_{v\in V,\,\|\nabla v\|=1} \left\{(f-f_h,v)+(r_h^i,v)-(\nabla u_h^i+\mathbf{d}_h^i,\nabla v)\right\}. \tag{4.4}$$

Using (4.3) and the Poincaré inequality on the mesh elements,

$$(f-f_h,v)\le \eta_{\mathrm{osc}}\|\nabla v\|,\quad \eta_{\mathrm{osc}}\equiv\left(\sum_{K\in\mathcal{T}_h}\eta_{\mathrm{osc},K}^2\right)^{1/2},\quad \eta_{\mathrm{osc},K}\equiv\frac{h_K}{\pi}\|f-f_h\|_K; \tag{4.5}$$

see, e.g., [24, p. A1767]. The Friedrichs inequality states that there exists a generic constant $0<C_{\mathrm{F}}\le 1$ such that

$$\|v\|\le C_{\mathrm{F}}h_\Omega\|\nabla v\| \qquad \forall v\in V, \tag{4.6}$$

where h_Ω denotes the diameter of the domain Ω. The value of C_{F} can be bounded[1] using, e.g., [50, Chapter 18]. Thus, from the Cauchy–Schwarz inequality and from (4.6),

$$(r_h^i,v)\le\|r_h^i\|\|v\|\le\|r_h^i\|C_{\mathrm{F}}h_\Omega\|\nabla v\|, \tag{4.7}$$

$$(\nabla u_h^i+\mathbf{d}_h^i,\nabla v)\le\|\nabla u_h^i+\mathbf{d}_h^i\|\|\nabla v\|. \tag{4.8}$$

Then (4.4) immediately gives the upper bound on the total error

$$\|\nabla(u-u_h^i)\|\le\eta_{\mathrm{osc}}+C_{\mathrm{F}}h_\Omega\|r_h^i\|+\|\nabla u_h^i+\mathbf{d}_h^i\|. \tag{4.9}$$

The part η_{osc} measures the oscillations in the right-hand side f and it is often negligible in comparison to the discretization error. The part $C_{\mathrm{F}}h_\Omega\|r_h^i\|$ in (4.9) bounds the algebraic error; see (5.3) below. Finally, we will associate the last term $\|\nabla u_h^i+\mathbf{d}_h^i\|$ with estimating the discretization error as in [24].

4.3 Upper bound using additional algebraic iterations

Following [24], the idea of using $\nu>0$ additional iterations described in Section 3.2 can be analogously applied here to substantially improve the bound (4.9).

Given the computed approximation u_h^i, we construct the algebraic residual representation r_h^i satisfying (2.7) and a reconstructed flux $\mathbf{d}_h^i\in\mathbf{V}_h$ satisfying (4.2). After $\nu>0$ additional iterations of the algebraic solver, giving the approximation $u_h^{i+\nu}$, we construct $r_h^{i+\nu}$ satisfying (2.7) with i replaced by $i+\nu$ and a reconstructed flux $\mathbf{d}_h^{i+\nu}\in\mathbf{V}_h$ satisfying $\nabla\cdot\mathbf{d}_h^{i+\nu}=f_h-r_h^{i+\nu}$. Thus,

$$r_h^i=-\nabla\cdot\mathbf{d}_h^i+f_h=-\nabla\cdot\mathbf{d}_h^i+\nabla\cdot\mathbf{d}_h^{i+\nu}+r_h^{i+\nu} \tag{4.10}$$

[1] For example, for a square domain $\Omega\subset\mathbb{R}^2$ we can take $C_{\mathrm{F}}=1/(2\pi)$, corresponding to the smallest eigenvalue of the Laplace operator; see, e.g., [50, relation (18.48) on p. 196]

and we have as above

$$\|\nabla(u-u_h^i)\| = \sup_{v\in V,\,\|\nabla v\|=1} \{(f-f_h,v)+(\mathbf{d}_h^i-\mathbf{d}_h^{i+\nu},\nabla v) \\ +(r_h^{i+\nu},v)-(\nabla u_h^i+\mathbf{d}_h^i,\nabla v)\},$$

which immediately leads to, cf. [24, Theorem 3.6]:

Theorem 1 (Upper bound on the total error)**.** *Let u be the weak solution given by (2.2) and let $u_h^i \in V_h$ be its approximation given at the ith algebraic solver iteration with the corresponding algebraic residual representation r_h^i given by (2.8). Let a reconstructed flux $\mathbf{d}_h^i \in \mathbf{V}_h$ satisfy (4.2). Consider $\nu > 0$ additional algebraic iterations, resulting in $r_h^{i+\nu}$ and $\mathbf{d}_h^{i+\nu}$. Then*

$$\|\nabla(u-u_h^i)\| \le \eta_{\mathrm{total}}^{i,\nu} \equiv \eta_{\mathrm{osc}} + \|\mathbf{d}_h^{i+\nu}-\mathbf{d}_h^i\| + C_{\mathrm{F}}h_\Omega\|r_h^{i+\nu}\| + \|\nabla u_h^i+\mathbf{d}_h^i\|,$$

where the data oscillation term η_{osc} is given by (4.5) and $C_{\mathrm{F}}h_\Omega$ is the constant from the Friedrichs inequality (4.6).

Remark 1. The statement of Theorem 1 deserves several comments that point out to the results presented later in the text. We typically choose ν in concordance with the theoretical justification (global efficiency) of Theorem 7 below; see also (7.3c) in the numerical experiments. Local efficiency of $\eta_{\mathrm{total}}^{i,\nu}$ is proved in Appendix B for i and ν based on local stopping criteria. Note that the sum $\|\mathbf{d}_h^{i+\nu}-\mathbf{d}_h^i\| + C_{\mathrm{F}}h_\Omega\|r_h^{i+\nu}\|$ gives an upper bound on the algebraic error (see Theorem 3 below), whereas the term $\|\nabla u_h^i+\mathbf{d}_h^i\|$ can be associated, at least in the case of a small algebraic error, with the discretization error; see the further results in Section 6 and Section 7.4.

4.4 Lower bound

Following the ideas in [4, Section 5.1], [51, Section 4.1.1], or [25, Section 3.3], we bound the total error $\|\nabla(u-u_h^i)\|$ from below using the solution of local conforming finite element problems.

Let $\mathcal{V}_h$ denote the set of mesh vertices with subsets $\mathcal{V}_h^{\mathrm{int}}$ for interior vertices and $\mathcal{V}_h^{\mathrm{ext}}$ for boundary ones. Let $\psi_{\mathtt{a}} \in \mathbb{P}_1(\mathcal{T}_h)\cap H^1(\Omega)$ stand for the hat function associated with a vertex $\mathtt{a} \in \mathcal{V}_h$ (i.e., $\psi_{\mathtt{a}}(\mathtt{a})=1$, $\psi_{\mathtt{a}}(\mathtt{a}')=0$ for $\mathtt{a}\neq\mathtt{a}' \in \mathcal{V}_h$). We denote by $\mathcal{T}_{\mathtt{a}}$ the union of elements sharing the vertex $\mathtt{a}\in\mathcal{V}_h$ and by $\omega_{\mathtt{a}}$ the corresponding open subdomain.

For each vertex $\mathtt{a}\in\mathcal{V}_h$, consider the infinite-dimensional space $H^1_*(\omega_{\mathtt{a}})$

$$H^1_*(\omega_{\mathtt{a}}) \equiv \begin{cases} v\in H^1(\omega_{\mathtt{a}});\,(v,1)_{\omega_{\mathtt{a}}}=0 & \mathtt{a}\in\mathcal{V}_h^{\mathrm{int}},\\ v\in H^1(\omega_{\mathtt{a}});\,v=0 \text{ on } \partial\omega_{\mathtt{a}}\cap\partial\Omega & \mathtt{a}\in\mathcal{V}_h^{\mathrm{ext}}. \end{cases} \tag{4.11}$$

For the functions from the space $H^1_*(\omega_{\mathtt{a}})$ the following Poincaré–Friedrichs-type inequalities hold: there exists a positive constant $C_{\mathrm{PF},\omega_{\mathtt{a}}}$, depending on the shape of the elements of the patch $\mathcal{T}_{\mathtt{a}}$ but not on their diameters, and a positive constant $C_{\mathrm{cont,PF},\omega_{\mathtt{a}}} \equiv 1+C_{\mathrm{PF},\omega_{\mathtt{a}}}h_{\omega_{\mathtt{a}}}\|\nabla\psi_{\mathtt{a}}\|_{\infty,\omega_{\mathtt{a}}}$ (see, e.g., [25, inequality (3.29)]) such that

$$\|v\|_{\omega_{\mathtt{a}}} \le C_{\mathrm{PF},\omega_{\mathtt{a}}}h_{\omega_{\mathtt{a}}}\|\nabla v\|_{\omega_{\mathtt{a}}} \qquad \forall v\in H^1_*(\omega_{\mathtt{a}}), \tag{4.12}$$

$$\|\nabla(\psi_{\mathtt{a}}v)\| \le C_{\mathrm{cont,PF},\omega_{\mathtt{a}}}\|\nabla v\|_{\omega_{\mathtt{a}}} \qquad \forall v\in H^1_*(\omega_{\mathtt{a}}). \tag{4.13}$$

For convex patches $\mathcal{T}_\mathtt{a}$ around the interior vertices $\mathtt{a}$ we have $C_{\mathrm{PF},\omega_\mathtt{a}} = 1/\pi$; see, e.g., [48]. For nonconvex patches we refer to [25, 57] and the references therein. For a shape-regular mesh $h_{\omega_\mathtt{a}}\|\nabla\psi_\mathtt{a}\|_{\infty,\omega_\mathtt{a}} = O(1)$ (see, e.g., [15, relation (3.1.43) on p. 124]), giving $C_{\mathrm{cont,PF},\omega_\mathtt{a}} = O(1)$; see the discussion in [25, Remark 3.24].

For each vertex $\mathtt{a} \in \mathcal{V}_h$, let $W_h^\mathtt{a}$ be a finite-dimensional subspace of $H^1_*(\omega_\mathtt{a})$. The simplest choice, which we use in numerical experiments in Section 7.4, is $W_h^\mathtt{a} \equiv \mathbb{P}_p(\mathcal{T}_\mathtt{a}) \cap H^1_*(\omega_\mathtt{a})$. We then have the following bound:

Theorem 2 (Lower bound on the total error)**.** *Let u be the weak solution given by (2.2) and let $u_h^i \in V_h$ be its approximation given at the ith algebraic solver iteration with the corresponding algebraic residual representation r_h^i given by (2.8). For each vertex $\mathtt{a} \in \mathcal{V}_h$, let $m_{h,\mathtt{a}} \in W_h^\mathtt{a}$ be the solution of*

$$(\nabla m_{h,\mathtt{a}}, \nabla v_h)_{\omega_\mathtt{a}} = (f, \psi_\mathtt{a} v_h)_{\omega_\mathtt{a}} - (\nabla u_h^i, \nabla(\psi_\mathtt{a} v_h))_{\omega_\mathtt{a}} \qquad \forall v_h \in W_h^\mathtt{a}.$$

Set $m_h \equiv \sum_{\mathtt{a}\in\mathcal{V}_h} \psi_\mathtt{a} m_{h,\mathtt{a}} \in V$. Then

$$\|\nabla(u - u_h^i)\| \geq \mu_{\mathrm{total}}^i \equiv \frac{\sum_{\mathtt{a}\in\mathcal{V}_h} \|\nabla m_{h,\mathtt{a}}\|^2_{\omega_\mathtt{a}}}{\|\nabla m_h\|}.$$

Proof. Since $m_h \in V$ by construction, we have from (4.1)

$$\begin{aligned}
\|\nabla(u - u_h^i)\| &= \sup_{v\in V,\, \|\nabla v\|=1} (\nabla(u - u_h^i), \nabla v) \\
&\geq \frac{1}{\|\nabla m_h\|}(\nabla(u - u_h^i), \nabla m_h) \\
&= \frac{1}{\|\nabla m_h\|} \sum_{\mathtt{a}\in\mathcal{V}_h} (\nabla(u - u_h^i), \nabla(\psi_\mathtt{a} m_{h,\mathtt{a}}))_{\omega_\mathtt{a}} \\
&= \frac{1}{\|\nabla m_h\|} \sum_{\mathtt{a}\in\mathcal{V}_h} \{(f, \psi_\mathtt{a} m_{h,\mathtt{a}})_{\omega_\mathtt{a}} - (\nabla u_h^i, \nabla(\psi_\mathtt{a} m_{h,\mathtt{a}}))_{\omega_\mathtt{a}}\} \\
&= \frac{1}{\|\nabla m_h\|} \sum_{\mathtt{a}\in\mathcal{V}_h} \|\nabla m_{h,\mathtt{a}}\|^2_{\omega_\mathtt{a}},
\end{aligned}$$

where we have used the fact that $\psi_\mathtt{a} m_{h,\mathtt{a}} \in H^1_0(\omega_\mathtt{a})$ for all vertices $\mathtt{a} \in \mathcal{V}_h$ and the definition of $m_{h,\mathtt{a}}$. □
□

Remark 2. The bound μ_{total}^i can further be localized using (4.13) as

$$\mu_{\mathrm{total}}^i \geq \frac{\left\{\sum_{\mathtt{a}\in\mathcal{V}_h} \|\nabla m_{h,\mathtt{a}}\|^2_{\omega_\mathtt{a}}\right\}^{1/2}}{(d+1)^{1/2} C_{\mathrm{cont,PF}}},$$

where $C_{\mathrm{cont,PF}} \equiv \max_{\mathtt{a}\in\mathcal{V}_h} C_{\mathrm{cont,PF},\omega_\mathtt{a}}$. Denoting by $\mathcal{V}_K$ the vertices of an element K and using the fact that each simplex has $(d+1)$ vertices, this can be seen from

$$\begin{aligned}
\|\nabla m_h\|^2 &= \sum_{K\in\mathcal{T}_h} \left\| \sum_{\mathtt{a}\in\mathcal{V}_K} (\nabla(\psi_\mathtt{a} m_{h,\mathtt{a}}))|_K \right\|^2_K \leq (d+1) \sum_{K\in\mathcal{T}_h} \sum_{\mathtt{a}\in\mathcal{V}_K} \|\nabla(\psi_\mathtt{a} m_{h,\mathtt{a}})\|^2_K \\
&= (d+1) \sum_{\mathtt{a}\in\mathcal{V}_h} \|\nabla(\psi_\mathtt{a} m_{h,\mathtt{a}})\|^2_{\omega_\mathtt{a}} \leq (d+1) C^2_{\mathrm{cont,PF}} \sum_{\mathtt{a}\in\mathcal{V}_h} \|\nabla m_{h,\mathtt{a}}\|^2_{\omega_\mathtt{a}}.
\end{aligned}$$

5 Estimating the algebraic error

We will now derive upper bounds on the algebraic error with the help of the representation of the algebraic residual r_h^i satisfying (2.7) and of the flux reconstruction $\mathbf{d}_h^i$ of Section A. We will make links to the bounds of Section 3 derived purely algebraically and to the total error bounds of the previous section. Section 5.4 recalls the lower bounds on the algebraic error of Section 3 and proposes a (function-based) construction of a lower bound analogously to Section 4.4.

5.1 Representation of the algebraic residual

We first propose two piecewise polynomial representations of the algebraic residual r_h^i satisfying (2.7).

The choice $r_h^i \in V_h = \mathbb{P}_p(\mathcal{T}_h) \cap H_0^1(\Omega)$ given by (2.7) requires solving the linear algebraic system with the *global mass matrix*

$$\mathsf{G}\mathsf{C}^i = \mathsf{R}^i, \qquad \mathsf{G}_{j\ell} \equiv (\psi_\ell, \psi_j), \quad j, \ell = 1, \dots, N. \tag{5.1}$$

Then $r_h^i = \Psi \mathsf{C}^i = \Psi \mathsf{G}^{-1} \mathsf{R}^i$. This representation of the algebraic residual has been considered in [5, Section 4], where it is called *the discrete residual*.

Equation (5.1) represents a global problem of the same size as (2.5). In order to avoid performing a global solve, we introduce a piecewise *discontinuous* polynomial representation $r_h^i \in \mathbb{P}_p(\mathcal{T}_h)$ using mutually independent local problems. For the ease of notation, the construction below is described for the Lagrangian basis of V_h. Denote by n_j the number of mesh elements forming the support of the basis function ψ_j, $j = 1, \dots, N$. Then, for each element $K \in \mathcal{T}_h$, define $r_h^i|_K \in \mathbb{P}_p(K)$, $r_h^i|_{\partial\Omega} = 0$, such that

$$(r_h^i, \psi_j)_K = \mathsf{R}_j^i / n_j \qquad \text{for } \psi_j \text{ nonvanishing on } K. \tag{5.2}$$

Summing (5.2) over all elements $K \in \mathcal{T}_h$, we see that (2.7) indeed holds. Denoting by R_K^i the vector on the right-hand side of (5.2) and by $\mathbf{G}_K$ the *local mass matrix*

$$(\mathbf{G}_K)_{j\ell} \equiv (\psi_\ell, \psi_j)_K \qquad \text{for } \psi_\ell, \psi_j \text{ nonvanishing on } K,$$

we have

$$r_h^i|_K = \Psi|_K (\mathbf{G}_K^{-1} \mathsf{R}_K^i) \qquad \forall K \in \mathcal{T}_h.$$

Construction (5.2) requires solving the system of the size $\frac{1}{2}(p+1)(p+2)$ separately on each element $K \in \mathcal{T}_h$.

5.2 Bound using the L^2 norm of the residual representation

Similarly to (4.1), using (2.9) and (4.7), the energy norm of the algebraic error satisfies

$$\begin{aligned} \|\nabla(u_h - u_h^i)\| &= \sup_{v_h \in V_h, \|\nabla v_h\| = 1} (\nabla(u_h - u_h^i), \nabla v_h) = \sup_{v_h \in V_h,\, \|\nabla v_h\| = 1} (r_h^i, v_h) \\ &\le C_{\mathrm{F}} h_\Omega \|r_h^i\|. \end{aligned} \tag{5.3}$$

We first discuss the bound (5.3) for the representation r_h^i constructed globally using (5.1). The discussion shows the relationship of (5.3) to the algebraic worst-case bounds of Section 3.1 and the role of the Friedrichs inequality constant $C_{\mathrm{F}} h_\Omega$. In the case (5.1),

$$\|r_h^i\|^2 = (\Psi \mathbf{G}^{-1} \mathsf{R}^i, \Psi \mathbf{G}^{-1} \mathsf{R}^i) = (\mathbf{G}^{-1} \mathsf{R}^i)^T \mathbf{G} (\mathbf{G}^{-1} \mathsf{R}^i) = (\mathsf{R}^i)^T \mathbf{G}^{-1} \mathsf{R}^i = \|\mathsf{R}^i\|_{\mathbf{G}^{-1}}^2,$$

and therefore

$$\|\nabla(u_h - u_h^i)\| = \|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}} = \|\mathsf{R}^i\|_{\mathbf{A}^{-1}} \leq C_{\mathrm{F}} h_\Omega \|\mathsf{R}^i\|_{\mathbf{G}^{-1}}. \tag{5.4}$$

An analogous estimate for the finite volume method is given in [34, Section 7.1], where it was observed in numerical experiments that this estimate can significantly overestimate the algebraic error. We note that

$$\begin{aligned} \|\mathsf{R}^i\|_{\mathbf{A}^{-1}}^2 &= (\mathsf{R}^i, \mathbf{A}^{-1} \mathsf{R}^i) = (\mathbf{G}^{-1/2} \mathsf{R}^i, \mathbf{G}^{1/2} \mathbf{A}^{-1} \mathbf{G}^{1/2} \mathbf{G}^{-1/2} \mathsf{R}^i) \\ &\leq \|\mathbf{G}^{1/2} \mathbf{A}^{-1} \mathbf{G}^{1/2}\| \cdot \|\mathbf{G}^{-1/2} \mathsf{R}^i\|^2 = \|\mathbf{G}^{1/2} \mathbf{A}^{-1} \mathbf{G}^{1/2}\| \cdot \|\mathsf{R}^i\|_{\mathbf{G}^{-1}}^2. \end{aligned} \tag{5.5}$$

Because (5.4) holds also for the special choice of R^i giving the equality in (5.5) (when $\mathbf{G}^{-1/2} \mathsf{R}^i$ is collinear with the eigenvector of $\mathbf{G}^{1/2} \mathbf{A}^{-1} \mathbf{G}^{1/2}$ corresponding to its largest eigenvalue), we have

$$\|\mathbf{G}^{1/2} \mathbf{A}^{-1} \mathbf{G}^{1/2}\| \leq (C_{\mathrm{F}} h_\Omega)^2. \tag{5.6}$$

This means that the reciprocal of the squared Friedrichs inequality constant $(C_{\mathrm{F}} h_\Omega)^{-2}$ (and through that the related smallest eigenvalue of the continuous operator; see, e.g., [50, Section 18]) gives a computable lower bound on the smallest eigenvalue of the (preconditioned) matrix $\mathbf{G}^{-1/2} \mathbf{A} \mathbf{G}^{-1/2}$ (cf. also [33], [2, Section 4.2]),

$$\frac{1}{(C_{\mathrm{F}} h_\Omega)^2} \leq \min_{\lambda \in \mathrm{sp}(\mathbf{G}^{-1/2} \mathbf{A} \mathbf{G}^{-1/2})} \lambda. \tag{5.7}$$

The local construction (5.2) leads to

$$\|\nabla(u_h - u_h^i)\| \leq C_{\mathrm{F}} h_\Omega \left(\sum_{K \in \mathcal{T}_h} \|r_h^i\|_K^2 \right)^{1/2} = C_{\mathrm{F}} h_\Omega \left(\sum_{K \in \mathcal{T}_h} \|\mathsf{R}_K^i\|_{\mathbf{G}_K^{-1}}^2 \right)^{1/2}. \tag{5.8}$$

There holds

$$\|\nabla(u_h - u_h^i)\| \leq C_{\mathrm{F}} h_\Omega \|\mathsf{R}^i\|_{\mathbf{G}^{-1}} \leq C_{\mathrm{F}} h_\Omega \left(\sum_{K \in \mathcal{T}_h} \|\mathsf{R}_K^i\|_{\mathbf{G}_K^{-1}}^2 \right)^{1/2}, \tag{5.9}$$

i.e., the bound (5.8) is weaker than the bound (5.4). The second inequality in (5.9) can be proved, e.g., using the results well-established in the domain decomposition methods; see, e.g., [20, Section 7.8]. A purely algebraic proof is given in [44, Section 5.2].

5.3 Upper bound using additional algebraic iterations

Analogously to Sections 3.2 and 4.3, we can bound the algebraic error using ν additional iteration steps. From (2.9), (4.10), and the Green theorem, for $v_h \in V_h$,

$$(\nabla(u_h - u_h^i), \nabla v_h) = (r_h^i, v_h) = (\mathbf{d}_h^i - \mathbf{d}_h^{i+\nu}, \nabla v_h) + (r_h^{i+\nu}, v_h). \tag{5.10}$$

Thus the following upper bound on the algebraic error immediately follows from (5.3):

Theorem 3 (Upper bound on the algebraic error)**.** *Let the assumptions of Theorem 1 be satisfied. Then*

$$\|\nabla(u_h - u_h^i)\| \le \eta_{\rm alg}^{i,\nu} \equiv \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\| + C_{\rm F} h_\Omega \|r_h^{i+\nu}\|.$$

Remark 3. The upper bound of Theorem 3 on the algebraic error allows evaluation of the *local indicators* $\eta_{{\rm alg},K}^{i,\nu} \equiv \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_K + C_{\rm F} h_\Omega \|r_h^{i+\nu}\|_K$ for the mesh elements $K \in \mathcal{T}_h$, with subsequently using these indicators for estimating the *local distribution* of the algebraic error $\|\nabla(u_h - u_h^i)\|_K$. This can indeed be very useful in localization of the significant components of the algebraic error over the discretization domain Ω, which represents an important problem; see [45] and the numerical illustrations in Section 7.2.

In order to show the relationship between (5.10) and (3.2), we note that, using (2.9),

$$(\mathbf{d}_h^i - \mathbf{d}_h^{i+\nu}, \nabla v_h) = (r_h^i - r_h^{i+\nu}, v_h) = (\nabla(u_h^{i+\nu} - u_h^i), \nabla v_h),$$

so that

$$\|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathsf{A}} = \sup_{v_h \in V_h,\, \|\nabla v_h\|=1} (\nabla(u_h^{i+\nu} - u_h^i), \nabla v_h) \le \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|.$$

Employing also (5.3) for $i+\nu$ in place of i, the upper bound of Theorem 3 appears weaker than the algebraic bound (3.2),

$$\|\mathsf{U} - \mathsf{U}^i\|_{\mathsf{A}} \le \|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathsf{A}} + \|\mathsf{R}^{i+\nu}\|_{\mathsf{A}^{-1}} \le \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\| + \|\mathsf{R}^{i+\nu}\|_{\mathsf{A}^{-1}}.$$

The fluxes $\mathbf{d}_h^i$ (and $\mathbf{d}_h^{i+\nu}$) are, however, essential for bounding the total error in Theorem 1 and, importantly, the algebraic estimator in Theorem 1 indeed bounds the algebraic error as we see from Theorem 3.

5.4 Lower bound

As seen in Section 3.2 (see (3.3)–(3.5)), a lower bound on the algebraic error is given by

$$(1-\gamma)\|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathsf{A}} \le \|\mathsf{U} - \mathsf{U}^i\|_{\mathsf{A}}$$

whenever $C_{\rm F} h_\Omega \|r_h^{i+\nu}\| \le \gamma \|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathsf{A}}$ with a parameter $0 < \gamma < 1$. For the CG method, the estimator $\mu_{\rm alg}^{{\rm CG},i,\nu}$ of (3.9) should be used instead. Alternatively, we can construct (cf. [46, Theorem 5.2]) a lower bound using homogeneous Dirichlet problems on patches ω_{a}, $\mathsf{a} \in \mathcal{V}_h$, (or larger subdomains of Ω):

Theorem 4 (Lower bound on the algebraic error)**.** *Let the assumptions of Theorem 2 be satisfied. For each vertex $\mathsf{a} \in \mathcal{V}_h$, let $m_{h,\mathsf{a}} \in V_h \cap H_0^1(\omega_{\mathsf{a}})$ be the solution of*

$$(\nabla m_{h,\mathsf{a}}, \nabla v_h)_{\omega_{\mathsf{a}}} = (f, v_h)_{\omega_{\mathsf{a}}} - (\nabla u_h^i, \nabla v_h)_{\omega_{\mathsf{a}}} \qquad \forall v_h \in V_h \cap H_0^1(\omega_{\mathsf{a}}).$$

Set $m_h \equiv \sum_{\mathsf{a} \in \mathcal{V}_h} m_{h,\mathsf{a}} \in V_h$. Then

$$\|\nabla(u_h - u_h^i)\| \ge \mu_{\rm alg}^i \equiv \frac{\sum_{\mathsf{a} \in \mathcal{V}_h} \|\nabla m_{h,\mathsf{a}}\|_{\omega_{\mathsf{a}}}^2}{\|\nabla m_h\|}.$$

Proof. Using (5.3) and the fact that $m_h \in V_h$,

$$\|\nabla(u_h - u_h^i)\| \geq \frac{1}{\|\nabla m_h\|}(\nabla(u_h - u_h^i), \nabla m_h) = \frac{\sum_{\mathtt{a}\in\mathcal{V}_h} \|\nabla m_{h,\mathtt{a}}\|^2_{\omega_\mathtt{a}}}{\|\nabla m_h\|}. \qquad \square$$

$\square$

6 Estimating the discretization error and construction of stopping criteria

A posteriori estimation of the discretization error $\|\nabla(u - u_h)\|$ is rather complicated as both u and u_h are unknown. The standard approaches proposed in literature are based on additional assumptions or properly justified heuristics on the algebraic error. Using

$$\|\nabla(u - u_h^i)\|^2 = \|\nabla(u - u_h)\|^2 + \|\nabla(u_h - u_h^i)\|^2 \tag{6.1}$$

that follows from the Galerkin orthogonality (2.4) and the results of the two previous sections, we give upper and lower bounds on the discretization error. We then propose global and local stopping criteria for a linear algebraic solver. In distinction with the previous works [34, Section 6.1] or [24, Section 3.3], the new stopping criteria guarantee that the iterations will not be stopped prematurely.

6.1 Lower bound

The first result follows easily from (6.1) and from the bounds of Theorems 2 and 3:

Theorem 5 (Lower bound on the discretization error)**.** *Let the assumptions of Theorems 2 and 3 hold. Let* $\mu^i_{\mathrm{total}} > \eta^{i,\nu}_{\mathrm{alg}}$. *Then*

$$\|\nabla(u - u_h)\| \geq \mu^{i,\nu}_{\mathrm{discr}} \equiv \left[\left(\mu^i_{\mathrm{total}}\right)^2 - \left(\eta^{i,\nu}_{\mathrm{alg}}\right)^2\right]^{1/2}.$$

In practice the assumption $\mu^i_{\mathrm{total}} > \eta^{i,\nu}_{\mathrm{alg}}$ may not be satisfied in the iterations where $\|\nabla(u_h - u_h^i)\| \approx \|\nabla(u - u_h^i)\|$. The accuracy of the bound in Theorem 5 becomes good from the point where $\eta^{i,\nu}_{\mathrm{alg}}$ gets small enough; see Section 7.4 for numerical illustrations.

6.2 Upper bound

One can similarly combine the upper bound on the total error of Theorem 1 and the lower bound on the algebraic error of Theorem 4 (note that $\eta^{i,\nu}_{\mathrm{total}} \geq \mu^i_{\mathrm{alg}}$):

Theorem 6 (Upper bound on the discretization error)**.** *Let the assumptions of Theorems 1 and 4 hold. Then*

$$\|\nabla(u - u_h)\| \leq \eta^{i,\nu}_{\mathrm{discr}} \equiv \left[\left(\eta^{i,\nu}_{\mathrm{total}}\right)^2 - \left(\mu^i_{\mathrm{alg}}\right)^2\right]^{1/2}.$$

When the CG method is used for solving the algebraic system (2.5), $\mu^{\mathrm{CG},i,\nu}_{\mathrm{alg}}$ of (3.9) is suggested to be used instead of μ^i_{alg} above.

6.3 Stopping criteria balancing the error components

Stopping criteria for algebraic iterative solvers typically aim at stopping the iterations when the algebraic error does not substantially contribute to the total error. Using the (global) energy norm, it seems natural to require that

$$\|\nabla(u_h - u_h^i)\| \leq \gamma_{\mathrm{alg}} \|\nabla(u - u_h)\|\,, \tag{6.2a}$$

where $\gamma_{\mathrm{alg}} > 0$ is a prescribed tolerance. As mentioned above, the spatial distribution of the discretization error and of the algebraic error can be very different from each other and the criterion (6.2a) may not be descriptive; see [45]. Therefore one may rather require that

$$\|\nabla(u_h - u_h^i)\|_{\omega_{\mathsf{a}}} \leq \gamma_{\mathrm{alg},\omega_{\mathsf{a}}} \|\nabla(u - u_h)\|_{\omega_{\mathsf{a}}} \qquad \forall \mathsf{a} \in \mathcal{V}_h. \tag{6.2b}$$

The stopping criteria proposed in [34, Section 6.1] or [24, Section 3.3] replaced $\|\nabla(u_h - u_h^i)\|$ and $\|\nabla(u - u_h)\|$ above by their computable estimates of the form (in the present setting) $\eta_{\mathrm{alg}}^{i,\nu}$ and $\|\nabla u_h^i + \mathbf{d}_h^i\|$. Such criteria seem to work well in practice and allow to prove efficiency of the total error bound (see also Theorem 7 below), but they do not guarantee (6.2a) and there is a danger that the algebraic iterations can be stopped prematurely.

Using the upper bound on the algebraic error $\eta_{\mathrm{alg}}^{i,\nu}$ of Theorem 3 and the lower bound on the discretization error $\mu_{\mathrm{discr}}^{i,\nu}$ of Theorem 5, we propose the stopping criterion

$$\eta_{\mathrm{alg}}^{i,\nu} \leq \gamma_{\mathrm{alg}} \mu_{\mathrm{discr}}^{i,\nu} \tag{6.3}$$

that *guarantees* balancing the error components while implying the validity of (6.2a). Note that (6.3) is equivalent to requesting

$$\eta_{\mathrm{alg}}^{i,\nu} \leq \widetilde{\gamma}_{\mathrm{alg}} \mu_{\mathrm{total}}^{i} \qquad \text{with } \widetilde{\gamma}_{\mathrm{alg}} \equiv \gamma_{\mathrm{alg}}/(1+\gamma_{\mathrm{alg}}^2)^{1/2} < 1.$$

Following [34, equation (6.3)] or [24, equations (3.13)–(3.15)] a *local stopping criterion* that mimics (6.2b) can be set as

$$\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_{\omega_{\mathsf{a}}} + C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\|_{\omega_{\mathsf{a}}} \leq \widetilde{\gamma}_{\mathrm{alg},\omega_{\mathsf{a}}} \frac{\|\nabla m_{h,\mathsf{a}}\|_{\omega_{\mathsf{a}}}}{C_{\mathrm{cont,PF},\omega_{\mathsf{a}}}} \qquad \forall \mathsf{a} \in \mathcal{V}_h. \tag{6.4}$$

Unfortunately, the error estimator of Theorem 3 is not guaranteed to locally bound the algebraic error from above, so that (6.2b) may not be, in general, satisfied. Nevertheless, the criterion (6.4) is sufficient to prove the local efficiency of the total error estimator $\eta_{\mathrm{total}}^{i,\nu}$ (see Theorem 8 in Appendix B below) and it seems to ensure the local balance of the algebraic and discretization errors; see numerical experiments in Section 7.5.

7 Numerical illustrations

For numerical illustration we use two Poisson test problems that were considered, e.g., in [37, 1].

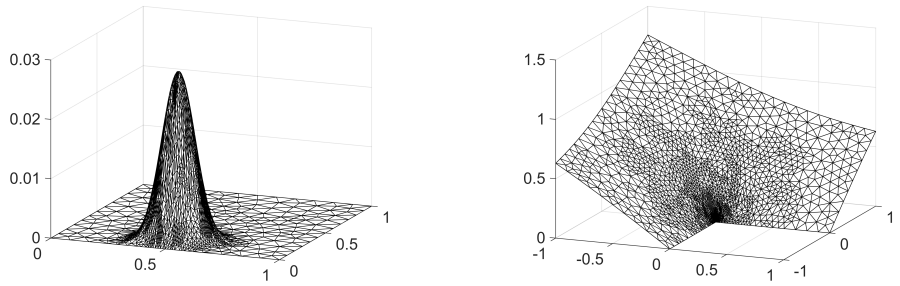

Figure 1: Left: solution (7.1) of the peak problem. Right: solution (7.2) of the L-shape problem.

Peak problem The model problem (2.1) with the square domain $\Omega \equiv (0,1) \times (0,1)$ and the right-hand side f chosen so that the solution u is given by

$$u(x,y) = x(x-1)y(y-1)\exp\left(-100\,(x-0.5)^2 - 100\,(y-0.117)^2\right), \qquad (7.1)$$

illustrated in Figure 1 (left). In the experiments, we discretize the problem on an adaptively refined mesh with 3 463 nodes using the piecewise quadratic polynomials. The corresponding algebraic system has 13 633 unknowns.

L-shape problem We take $\Omega \equiv (-1,1) \times (-1,1) \setminus [0,1] \times [-1,0]$ and solve

$$-\Delta u = 0 \quad \text{in } \Omega, \qquad u = u_D \quad \text{on } \partial\Omega,$$

where the (inhomogeneous) Dirichlet boundary condition u_D is chosen so that the solution u is in polar coordinates (r,θ) given by

$$u(r,\theta) = r^{2/3}\sin\left(\frac{2}{3}\theta\right), \qquad (7.2)$$

illustrated in Figure 1 (right). The extension of our estimates to $u_D \neq 0$ is possible following [22]. In particular, the flux reconstruction of Appendix A and the upper bound of Theorem 3 for the algebraic error remain unchanged. In the upper bound (4.9) and in Theorem 1, an additional term corresponding to the approximation of u_D by a piecewise polynomial function is added. This term is neglected in the experiments. We discretize the problem on an adaptively refined mesh with 628 nodes using the piecewise cubic polynomials. The corresponding algebraic system has here 5 098 unknowns.

The experiments are performed in Matlab R2014b with Partial Differential Equation Toolbox. We use our implementation of arbitrary degree conforming finite element method and of Raviart–Thomas–Nédélec spaces. We set $p' = p$, i.e., the reconstructed fluxes $\mathbf{d}_h^i$ are of the same order as the FEM approximation u_h^i. The algebraic system (2.5) is solved using the CG method preconditioned by the incomplete Cholesky decomposition with zero fill-in (Matlab `ichol` command) and starting with the zero initial guess. The exact solutions of the algebraic systems are approximated using the build-in Matlab "backslash" direct solver; in the performed numerical experiments, the algebraic error in this

approximate solution is negligible. We point out that the experiments do not aim at the preconditioning tuned to the problem, but at demonstrating fairly the issues that might be encountered in practical use of the presented bounds.

The initial (uniform) meshes are generated using the Matlab Delaunay triangulation (`initmesh` command). For generating the sequence of adaptively refined meshes we, for the reproducibility of the results, refine according to the actual distribution of the *discretization error*, i.e., we compute (up to a quadrature error that is in the given experiments negligible) the discretization error $\|\nabla(u-u_h)\|_K$ on each element of the triangulation (recall that u_h is for the purpose of the experiments sufficiently accurately approximated using the direct solution of the algebraic system). We mark the smallest subset of elements that contributes to the squared energy norm of the discretization error by at least 25%. This requires ordering the elements according to the error size, which is in practice usually avoided, e.g., by proceeding as in [23, Section 5.2] or [54, pp. 10–11]. The refinement of the mesh uses the newest-vertex-bisection algorithm implemented in the Matlab `refinemesh` command.

7.1 Algebraic error: the cost of the additional iterations

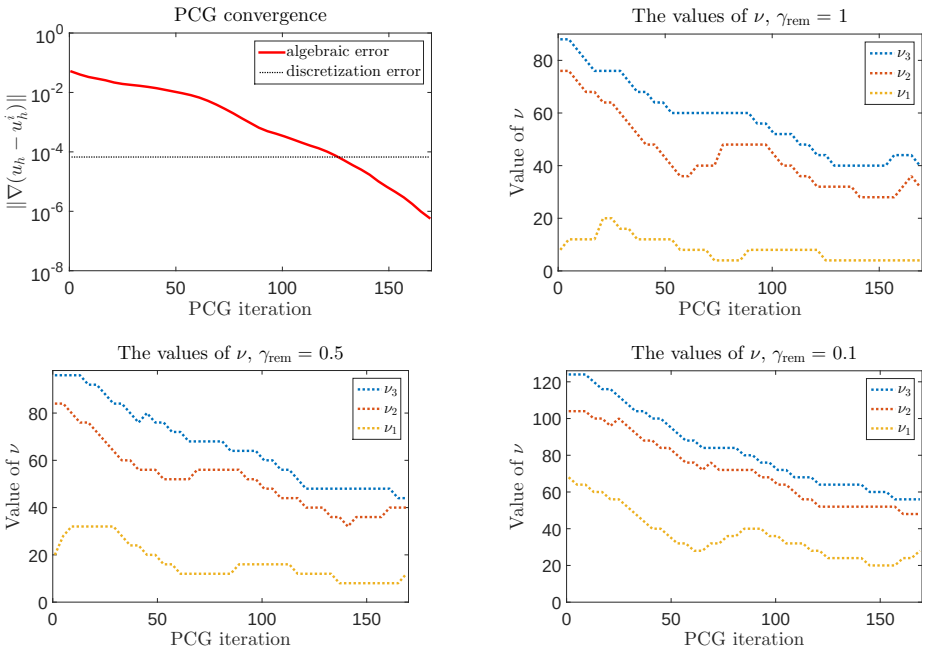

Figure 2: Peak problem: PCG convergence and the values of ν_1, ν_2, ν_3 determined by (7.3) for different choices of $\gamma_{\rm rem}$.

We first compare the cost of the upper bounds on the algebraic error of Sections 3.2 and 5.3 in terms of the number ν of the additional algebraic iterations. For the given tolerance $\gamma_{\rm rem} = 1, 0.5, 0.1$, we identify ν_1, ν_2, and ν_3 as

the smallest values satisfying

$$\|\mathsf{R}^{i+\nu_1}\|_{\mathbf{A}^{-1}} \le \gamma_{\rm rem}\|\mathsf{U}^{i+\nu_1}-\mathsf{U}^{i}\|_{\mathbf{A}}, \tag{7.3a}$$

$$\|\mathbf{A}^{-1}\|^{1/2}\cdot\|\mathsf{R}^{i+\nu_2}\| \le \gamma_{\rm rem}\|\mathsf{U}^{i+\nu_2}-\mathsf{U}^{i}\|_{\mathbf{A}}, \tag{7.3b}$$

$$C_{\rm F}h_\Omega\|r_h^{i+\nu_3}\| \le \gamma_{\rm rem}\|\mathbf{d}_h^{i+\nu_3}-\mathbf{d}_h^{i}\|, \tag{7.3c}$$

for each iteration step i. The number of additional iterations ν_1 of (7.3a) is always smaller than ν_2, ν_3. We recall, however, that $\|\mathsf{R}^{i+\nu_1}\|_{\mathbf{A}^{-1}} = \|\mathsf{U}-\mathsf{U}^{i+\nu_1}\|_{\mathbf{A}}$ is not available in practice. The criterion (7.3b) corresponds to the worst-case algebraic bound for $\|\mathsf{R}^{i+\nu_2}\|_{\mathbf{A}^{-1}}$ described in Section 3.1; see (3.6). For the purpose of the present study we (tightly) approximate the norm $\|\mathbf{A}^{-1}\|$ using the Matlab `eigs` command estimating the smallest eigenvalue of $\mathbf{A}$. Finally, the criterion (7.3c) corresponds to the computable upper bound of Theorem 3 on the algebraic error based on the flux reconstruction.

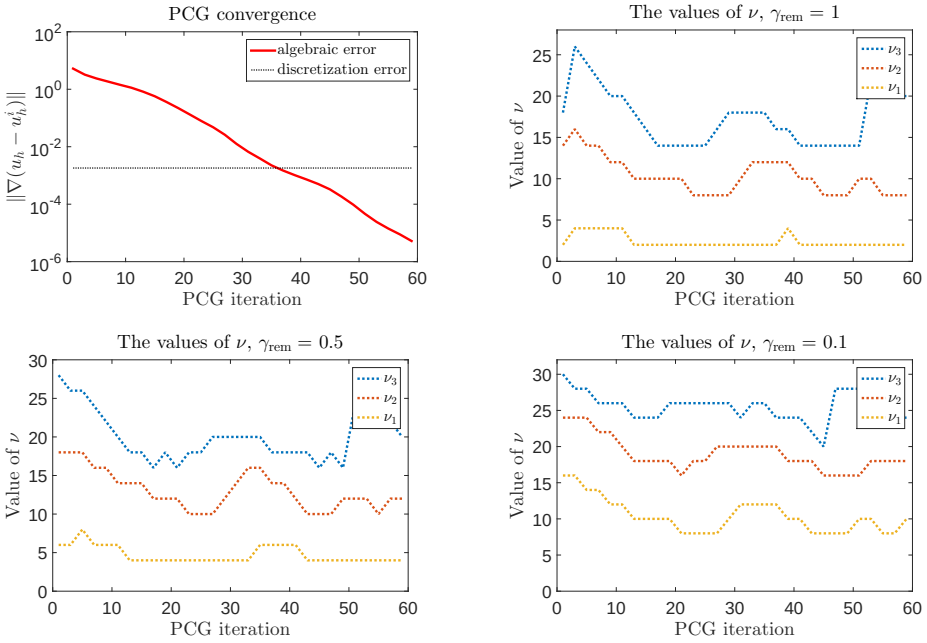

Figure 3: L-shape problem: PCG convergence and the values of ν_1, ν_2, ν_3 determined by (7.3) for different choices of $\gamma_{\rm rem}$.

In the experiments (see Figures 2 and 3) we observe relatively large values of ν_2 and ν_3, with $\nu_2 \le \nu_3$. The large value of ν_3 indicates a possible nonnegligible cost of the upper bound of Theorem 3 (and also of the upper bound of Theorem 1 on the total error). The comparison with ν_1 reveals that there may be a room for further improvements. However, as demonstrated below, for the cost of the additional ν_3 iterations, we get in our experiments upper bounds for the total and algebraic errors with very favorable effectivity indices and, in particular, a remarkably accurate information on the *local distribution of these errors.*

We also comment on the difference between the upper bound on the algebraic error (5.8) corresponding to the locally constructed representation of the

algebraic residual and the bound (5.4) corresponding to the global construction of r_h^i; see the inequality (5.9). In our numerical experiments, the relative overestimation

$$\frac{\left(\sum_{K\in\mathcal{T}_h}\|\mathsf{R}_K^i\|^2_{\mathsf{G}_K^{-1}}\right)^{1/2}-\|\mathsf{R}^i\|_{\mathsf{G}^{-1}}}{\|\mathsf{R}^i\|_{\mathsf{G}^{-1}}}$$

is below 18% (peak problem), respectively below 12% (L-shape problem).

7.2 Algebraic error: effectivity indices and localization

In this section we study how far the upper bounds on the algebraic error are from the actual error. For the ease of notation, let, corresponding to the bounds of Sections 3.2 and 5.3,

$$\eta^{i,\nu_1}_{\mathrm{alg},1}\equiv\|\mathsf{U}^{i+\nu_1}-\mathsf{U}^i\|_{\mathsf{A}}+\|\mathsf{R}^{i+\nu_1}\|_{\mathsf{A}^{-1}}, \tag{7.4a}$$

$$\eta^{i,\nu_2}_{\mathrm{alg},2}\equiv\|\mathsf{U}^{i+\nu_2}-\mathsf{U}^i\|_{\mathsf{A}}+\|\mathsf{A}^{-1}\|^{1/2}\cdot\|\mathsf{R}^{i+\nu_2}\|, \tag{7.4b}$$

$$\eta^{i,\nu_3}_{\mathrm{alg},3}\equiv\|\mathbf{d}_h^{i+\nu_3}-\mathbf{d}_h^i\|+C_{\mathrm{F}}h_\Omega\|r_h^{i+\nu_3}\|. \tag{7.4c}$$

Here ν_1, ν_2, and ν_3 are determined by (7.3). For these bounds, the effectivity indices

$$I^i_{\mathrm{eff}}(\eta^{i,\nu_\bullet}_{\mathrm{alg},\bullet})\equiv\frac{\eta^{i,\nu_\bullet}_{\mathrm{alg},\bullet}}{\|\mathsf{U}-\mathsf{U}^i\|_{\mathsf{A}}} \tag{7.5}$$

are given in Figures 4–6. They confirm our expectation (see (3.4) and (3.5)) that $I^i_{\mathrm{eff}}(\eta^{i,\nu_\bullet}_{\mathrm{alg},\bullet})\approx 1+\gamma_{\mathrm{rem}}$, so that, for the cost of $\nu_\bullet$ additional iterations, we get the estimates with the efficiency controlled by the parameter γ_{rem}. In Figure 5, we give additionally the effectivity index

$$I^i_{\mathrm{eff}}(\mu^{\mathrm{CG},i,\nu}_{\mathrm{alg}})\equiv\frac{\mu^{\mathrm{CG},i,\nu}_{\mathrm{alg}}}{\|\mathsf{U}-\mathsf{U}^i\|_{\mathsf{A}}}$$

that illustrates the efficiency of the lower bound $\mu^{\mathrm{CG},i,\nu}_{\mathrm{alg}}$ (see (3.9)) from [55, 56], with the values of ν fixed for the peak and the L-shape problems to $\nu=5$, 10 and 2, 5 respectively. We note that $I^i_{\mathrm{eff}}(\mu^{\mathrm{CG},i,\nu}_{\mathrm{alg}})$ strongly depends on the decrease of the energy norm of the algebraic error between the iteration steps i and $i+\nu$. With a more powerful preconditioner resulting in a faster PCG convergence, analogous results will be achieved for much smaller number of additional algebraic iterations.

As discussed in Remark 3, the flux-reconstruction-based upper bound of Theorem 3 allows evaluating the local indicators $\|\mathbf{d}_h^{i+\nu}-\mathbf{d}_h^i\|_K+C_{\mathrm{F}}h_\Omega\|r_h^{i+\nu}\|_K$ and estimating the *local distribution of the algebraic error* $\|\nabla(u_h-u_h^i)\|_K$. As we can see in Figures 7 and 8, the local indicators provide a remarkably accurate description of the local distribution of the algebraic error. We observed similarly good results also in other iteration steps, choices of $\gamma_{\mathrm{rem}}=0.1,1$, and other test problems. Please note that the algebraic error can be localized in parts of the discretization domain Ω where the discretization error can be small, see [45] and Figures 10 and 11 below. We point out that the algebraic error does not equilibrate over the domain using the adaptive mesh refinement.

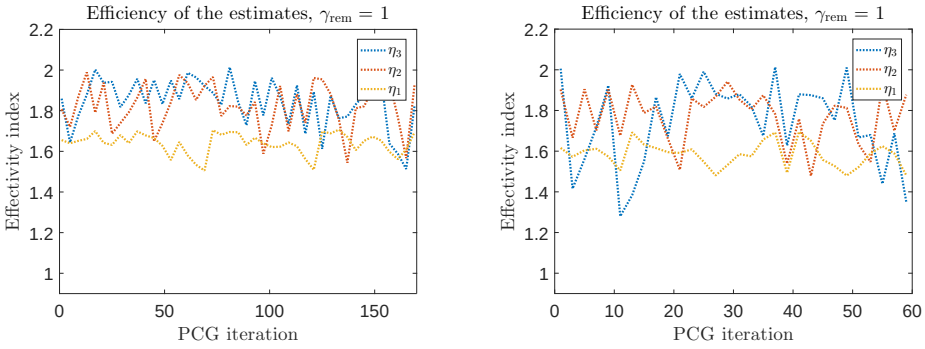

Figure 4: Effectivity indices $I^i_{\rm eff}(\eta^{i,\nu_\bullet}_{{\rm alg},\bullet})$ (7.5) of the algebraic error upper bounds (7.4) in the peak (left) and L-shape problems (right). The values of ν_1, ν_2, ν_3 are determined by (7.3) with $\gamma_{\rm rem} = 1$. Here $\eta^{i,\nu_k}_{{\rm alg},k}$ is simply denoted as η_k.

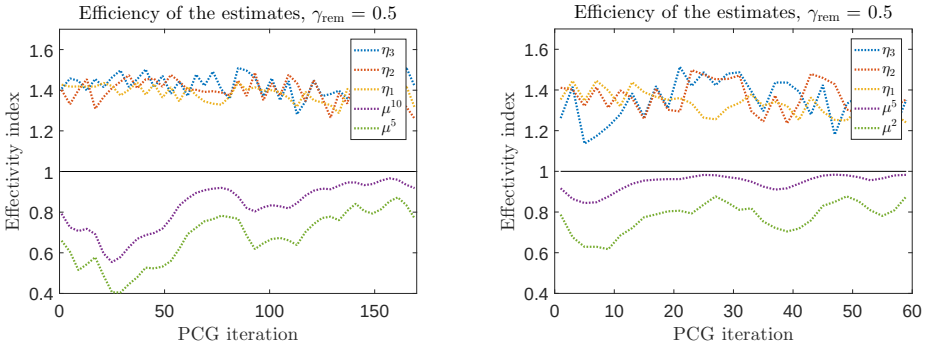

Figure 5: Effectivity indices $I^i_{\rm eff}(\eta^{i,\nu_\bullet}_{{\rm alg},\bullet})$ (7.5) of the algebraic error upper bounds (7.4) and the effectivity index $I^i_{\rm eff}(\mu^{{\rm CG},i,\nu}_{\rm alg})$ of the lower bound $\mu^{{\rm CG},i,\nu}_{\rm alg}$ with the fixed values of ν in the peak (left) and L-shape problems (right). The values of ν_1, ν_2, ν_3 are determined by (7.3) with $\gamma_{\rm rem} = 0.5$. Here $\eta^{i,\nu_k}_{{\rm alg},k}$ and $\mu^{{\rm CG},i,\nu}_{\rm alg}$ are simply denoted as η_k and μ^ν, respectively.

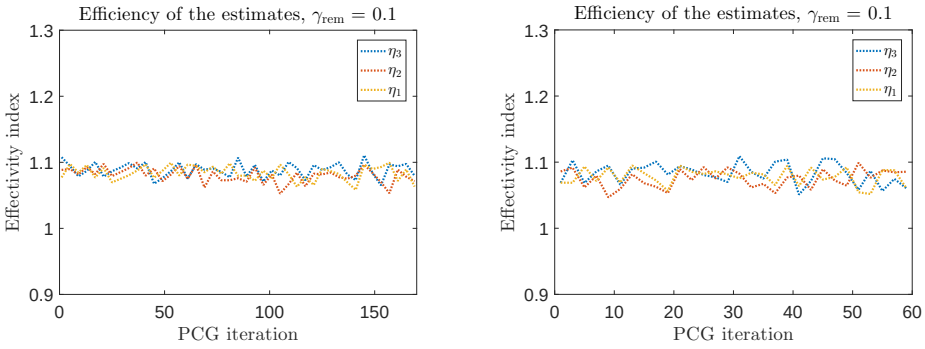

Figure 6: Effectivity indices $I^i_{\rm eff}(\eta^{i,\nu_\bullet}_{{\rm alg},\bullet})$ (7.5) of the algebraic error upper bounds (7.4) in the peak (left) and L-shape problems (right). The values of ν_1, ν_2, ν_3 are determined by (7.3) with $\gamma_{\rm rem} = 0.1$. Here $\eta^{i,\nu_k}_{{\rm alg},k}$ is simply denoted as η_k.

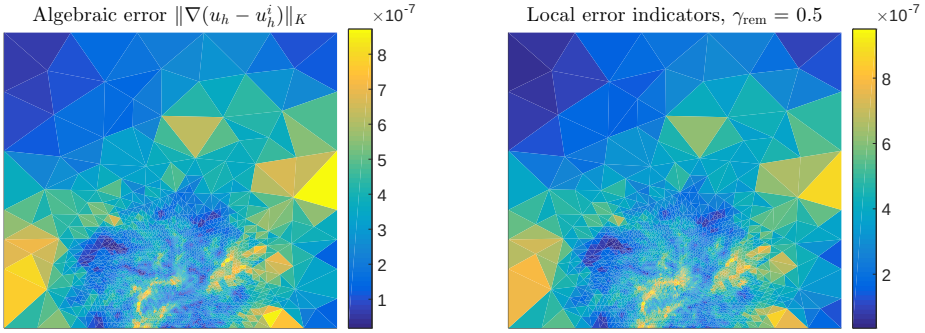

Figure 7: Peak problem, iteration $i = 137$: elementwise distribution of the algebraic error $\|\nabla(u_h - u_h^i)\|_K$ and the local algebraic error indicators $\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_K + C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\|_K$. The value of ν, $\nu = 48$, is determined by (7.3c) with $\gamma_{\text{rem}} = 0.5$.

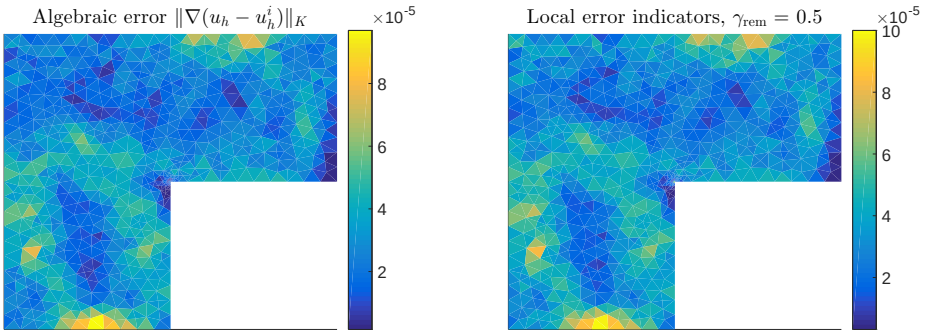

Figure 8: L-shape problem, iteration $i = 39$: elementwise distribution of the algebraic error $\|\nabla(u_h - u_h^i)\|_K$ and the local algebraic error indicators $\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_K + C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\|_K$. The value of ν, $\nu = 18$, is determined by (7.3c) with $\gamma_{\text{rem}} = 0.5$.

7.3 Bounding and localizing the total error

We now illustrate the upper bound $\eta^{i,\nu}_{\text{total}}$ of Theorem 1. Figure 9 depicts the total error $\|\nabla(u-u^i_h)\|$, the upper bound, and the error indicators $\|\nabla u^i_h + \mathbf{d}^i_h\|$, $\|\mathbf{d}^{i+\nu}_h - \mathbf{d}^i_h\|$, and $C_{\mathrm{F}} h_\Omega \|r^{i+\nu}_h\|$. We observe that $\eta^{i,\nu}_{\text{total}}$ tightly follows the actual value of the error. The parameter γ_{rem} in (7.3c) is set to 0.5.

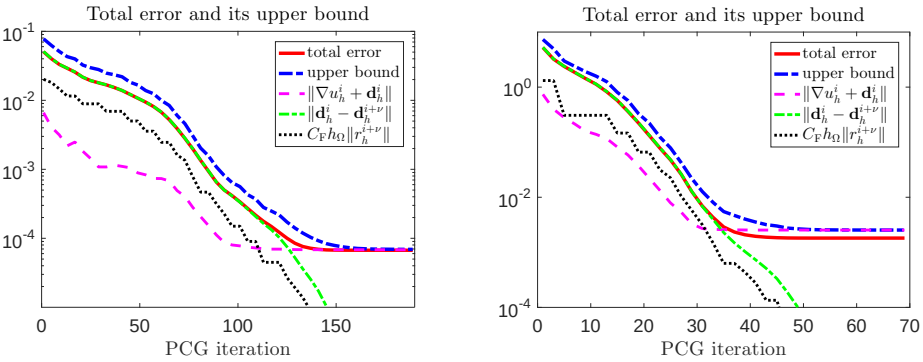

Figure 9: Total error $\|\nabla(u-u^i_h)\|$, the upper bound of Theorem 1, and the error indicators $\|\nabla u^i_h + \mathbf{d}^i_h\|$, $\|\mathbf{d}^{i+\nu}_h - \mathbf{d}^i_h\|$, and $C_{\mathrm{F}} h_\Omega \|r^{i+\nu}_h\|$ in the peak (left) and L-shape problems (right). The value of ν is determined by (7.3c) with $\gamma_{\text{rem}} = 0.5$.

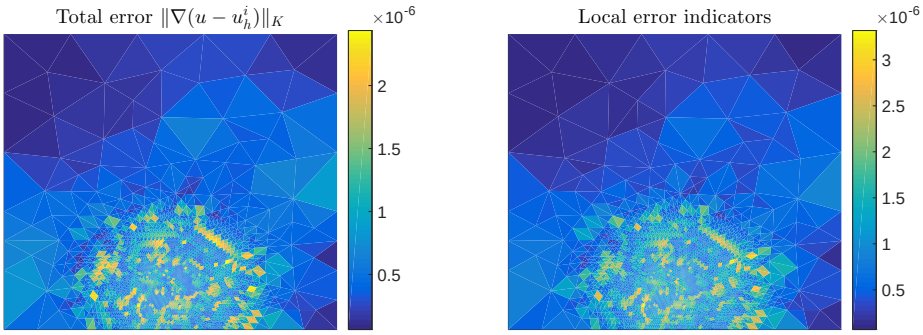

Figure 10: Peak problem: elementwise distribution of the total error $\|\nabla(u-u^i_h)\|_K$ and the local error indicators $\eta_{\text{osc},K} + \|\mathbf{d}^{i+\nu}_h - \mathbf{d}^i_h\|_K + C_{\mathrm{F}} h_\Omega \|r^{i+\nu}_h\|_K + \|\nabla u^i_h + \mathbf{d}^i_h\|_K$ in the iteration $i = 137$ with $\nu = 48$.

In Figures 10 and 11 we give the comparison of the local distribution of the total error $\|\nabla(u-u^i_h)\|_K$ and the sum $\eta_{\text{osc},K} + \|\mathbf{d}^{i+\nu}_h - \mathbf{d}^i_h\|_K + C_{\mathrm{F}} h_\Omega \|r^{i+\nu}_h\|_K + \|\nabla u^i_h + \mathbf{d}^i_h\|_K$ of the local indicators. Here the iteration step i and the number ν of additional iterations are set as the smallest values determined by the conditions (B.3a)–(B.3b) as described in Appendix B with $\gamma_{\text{alg}} = \gamma_{\text{rem}} = 0.5$.

7.4 Estimating the discretization error

We illustrate the discretization error bounds of Section 6. In Figures 12 and 13 we plot these bounds together with the estimator $\|\nabla u^i_h + \mathbf{d}^i_h\|$ that we have

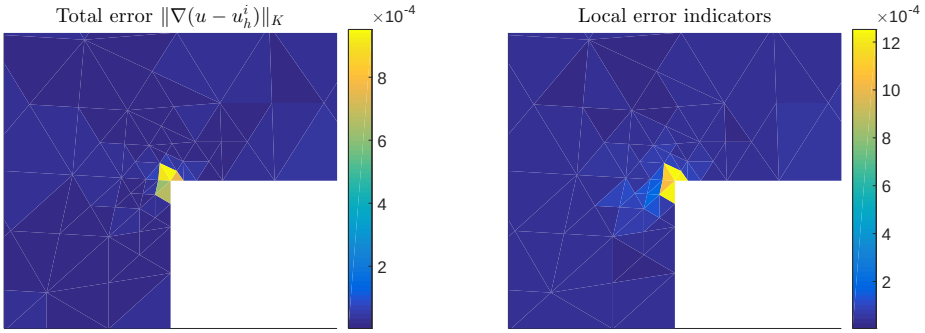

Figure 11: L-shape problem: elementwise distribution of the total error $\|\nabla(u-u_h^i)\|_K$ and the local error indicators $\eta_{\mathrm{osc},K} + \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_K + C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\|_K + \|\nabla u_h^i + \mathbf{d}_h^i\|_K$ in the iteration $i = 39$ with $\nu = 18$. We plot in both figures the part $[-0.02, 0.02] \times [-0.02, 0.02]$ of the discretization domain Ω.

identified with the discretization error in Theorem 1. As in the previous experiments, the number ν of additional iterations is determined by (7.3c) with $\gamma_{\mathrm{rem}} = 0.5$.

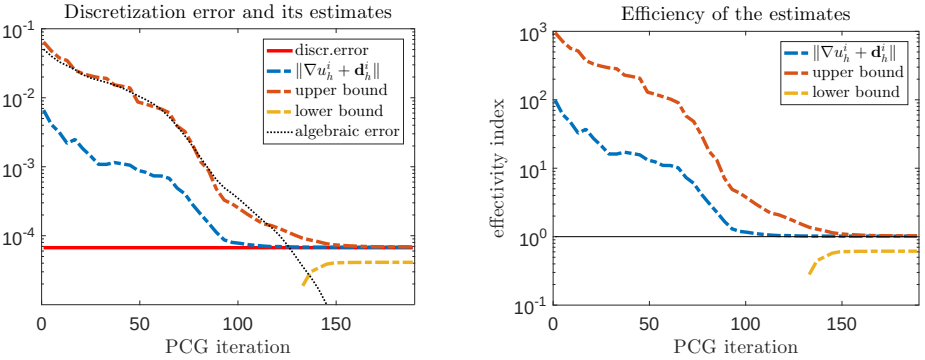

Figure 12: Peak problem: the discretization error $\|\nabla(u-u_h)\|$, the estimate $\|\nabla u_h^i + \mathbf{d}_h^i\|$, the upper bound $\eta_{\mathrm{discr}}^{i,\nu}$ of Theorem 6 with $\mu_{\mathrm{alg}}^{\mathrm{CG},i,\nu}$, and the lower bound $\mu_{\mathrm{discr}}^{i,\nu}$ of Theorem 5 (left); the efficiency of the estimates (right).

Estimating the discretization error via Theorems 5 and 6 is naturally inaccurate in the iterations where the energy norm of the total error is mostly dominated by the algebraic error; cf. upper left parts of Figures 2 and 3. When the algebraic error drops below the discretization error, our upper and lower bounds get close to each other and provide a tight estimate for the discretization error.

In all performed experiments with the Poisson model problem (here we present just a small sample), we have observed that $\|\nabla u_h^i + \mathbf{d}_h^i\| > \|\nabla(u-u_h)\|$, i.e., the estimate $\|\nabla u_h^i + \mathbf{d}_h^i\|$ gave an upper bound on the actual discretization error. However, an extrapolation from these *observations* can lead to false statements. As demonstrated below in Section 7.6 on the test problem with inhomogeneous diffusion tensor, the component associated with the discretiza-

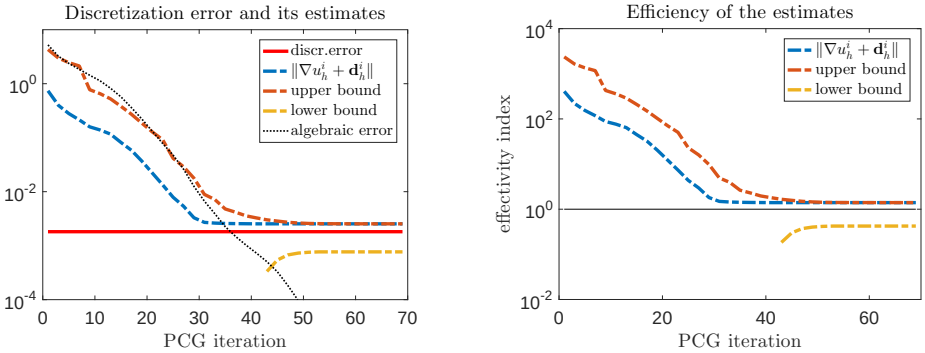

Figure 13: L-shape problem: the discretization error $\|\nabla(u-u_h)\|$, the estimate $\|\nabla u_h^i + \mathbf{d}_h^i\|$, the upper bound $\eta_{\mathrm{discr}}^{i,\nu}$ of Theorem 6 with $\mu_{\mathrm{alg}}^{\mathrm{CG},i,\nu}$, and the lower bound of Theorem 5 (left); the efficiency of the estimates (right).

tion error drops, in some iterations, below the energy norm of the discretization error. This emphasizes a need for guaranteed bounds on the errors and a need for mathematically justified stopping criteria that ensure balancing the error components as in (6.2).

7.5 Local stopping criteria and the spatial distribution of errors

We use the L-shape problem to illustrate that the local stopping criterion (6.4) prevents the algebraic error from dominating locally, as it can happen under the global criteria; cf. the numerical experiments of [45]. We consider the approximation u_h^{47} determined by the global stopping criterion (6.3) with $\gamma_{\mathrm{alg}} \equiv 0.5$ (the value of $\nu = 20$ is determined by (7.3c) with $\gamma_{\mathrm{rem}} \equiv 0.5$), and the approximation u_h^{79} satisfying the proposed local stopping criterion (6.4) with $\widetilde{\gamma}_{\mathrm{alg},\omega_{\mathsf{a}}} \equiv \gamma_{\mathrm{alg},\omega_{\mathsf{a}}}/(1+\gamma_{\mathrm{alg},\omega_{\mathsf{a}}}^2)^{1/2}$, $\gamma_{\mathrm{alg},\omega_{\mathsf{a}}} \equiv \gamma_{\mathrm{alg}}$, $\forall \mathsf{a} \in \mathcal{V}_h$ (the number $\nu = 20$ of the additional algebraic iterations is here determined by (B.8a) with $\gamma_{\mathrm{rem},K} \equiv \gamma_{\mathrm{rem}}$, $\forall K \in \mathcal{T}_h$).

Figure 14 depicts the differences $u-u_h^{47}$, $u-u_h^{79}$ and $u_h-u_h^{47}$, $u_h-u_h^{79}$ that visualize the total and algebraic errors respectively. We note that the algebraic part $u_h-u_h^{47}$ substantially affects the shape of $u-u_h^{47}$ in most of the domain Ω. This is not the case for $u-u_h^{79}$ as $|u(\mathbf{x})-u_h(\mathbf{x})| \geq 10^{-7}$ in most of the domain Ω.

7.6 Numerical results for a problem with inhomogeneous diffusion tensor

In order to further demonstrate a possible use of the presented methodology for obtaining the bounds on the total error and its components, we consider also the test problem with inhomogeneous diffusion tensor proposed in [43, Section 5.3] (based on the formulas published in [35]),

$$-\nabla\cdot(\mathbf{S}\nabla u) = 0 \quad \text{in } \Omega \equiv (-1,1)\times(-1,1), \qquad u = u_D \quad \text{on } \partial\Omega, \tag{7.6}$$

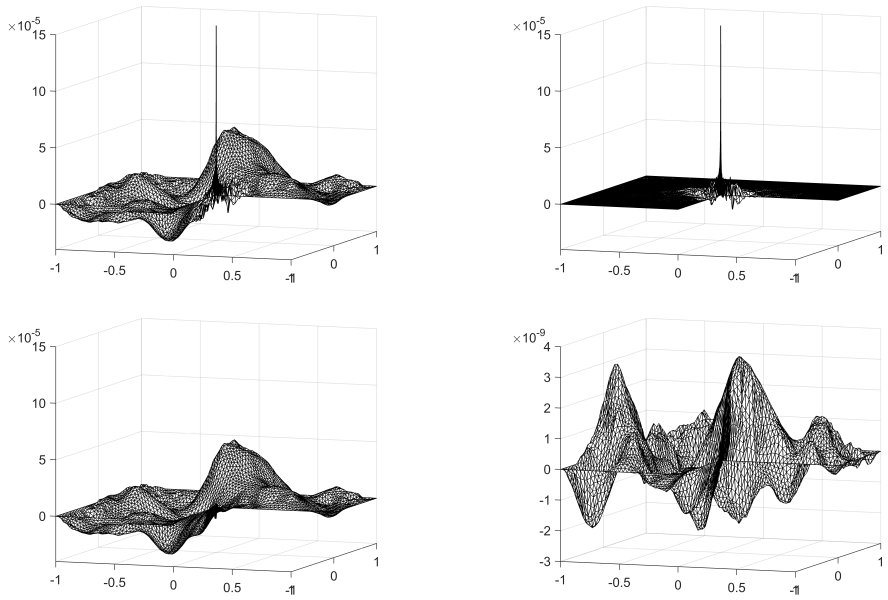

Figure 14: L-shape problem: the difference $u - u_h^{47}$ counting for the total error of the approximation u_h^{47} determined by the *global* stopping criterion (6.3) (upper left), its analogy $u - u_h^{79}$ for the approximation u_h^{79} determined by the *local* stopping criterion (6.4) (upper right), the algebraic part $u_h - u_h^{47}$ (bottom left), and its analogy $u_h - u_h^{79}$ (bottom right). Vertical axes are scaled by 10^{-5}, 10^{-5}, 10^{-5}, and 10^{-9}, respectively.

where the domain Ω is divided into four subdomains Ω_i corresponding to the axis quadrants numbered counterclockwise. The diffusion tensor $\mathbf{S}$ is a piecewise constant multiple of the identity matrix, $\mathbf{S}|_{\Omega_i} \equiv s_i\mathbf{I}$, with $s_1 = s_3 \approx 161.4$, $s_2 = s_4 = 1$. These values and the Dirichlet boundary condition u_D are used such that the solution u of (7.6) exhibits a singularity at the origin, $u \in H^{1.1-\epsilon}(\Omega)$, $\forall \epsilon > 0$. We discretize the problem using piecewise affine functions on adaptively refined mesh with 8040 nodes. The adaptive mesh refinement and the setting for iterative algebraic solver are analogous to those described above for peak and L-shape test problems. The stopping criteria are given by (B.3) with with $\gamma_{\mathrm{alg}} = \gamma_{\mathrm{rem}} = 0.5$.

The left part of Figure 15 gives, analogously to Figure 9, the energy norm of the total error $\|\mathbf{S}^{1/2}\nabla(u-u_h^i)\|$, its upper bound $\eta_{\mathrm{total}}^{i,\nu}$ of Theorem 1 modified for the test problem (7.6), and the corresponding error indicators $\|\mathbf{S}^{1/2}\nabla u_h^i + \mathbf{S}^{-1/2}\mathbf{d}_h^i\|$, $\|\mathbf{S}^{-1/2}(\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i)\|$, and $C_{\mathrm{F}} h_\Omega c_{\mathbf{S}}^{-1/2}\|r_h^{i+\nu}\|$. Here $c_{\mathbf{S}}$ denotes a uniform lower bound on the smallest eigenvalue of $\mathbf{S}$ in Ω; in the considered test problem, $c_{\mathbf{S}} = 1$. In this experiment, we can in some iterations observe

$$\|\mathbf{S}^{1/2}\nabla u_h^i + \mathbf{S}^{-1/2}\mathbf{d}_h^i\| < \|\mathbf{S}^{1/2}\nabla(u - u_h)\|,$$

i.e. the indicator $\|\mathbf{S}^{1/2}\nabla u_h^i + \mathbf{S}^{-1/2}\mathbf{d}_h^i\|$ associated with the discretization error $\|\mathbf{S}^{1/2}\nabla(u-u_h)\|$ does not provide, in general, its upper bound; cf. the discussion

in Sections 6.3 and 7.4. The right part of Figure 15 depicts the effectivity indices

$$\frac{\eta^{i,\nu}_{\text{total}}}{\|\mathbf{S}^{1/2}\nabla(u-u_h^i)\|}, \qquad \frac{\|\mathbf{S}^{-1/2}(\mathbf{d}_h^{i+\nu}-\mathbf{d}_h^i)\| + C_{\mathrm{F}} h_\Omega c_{\mathbf{S}}^{-1/2}\|r_h^{i+\nu}\|}{\|\mathbf{S}^{1/2}\nabla(u_h-u_h^i)\|} \tag{7.7}$$

of the upper bounds on the total and algebraic errors, respectively. We can see a similar behavior as for the Laplace operator; cf. Figure 9. Figure 16 gives, analogously to Figures 7–8, the local distribution of the algebraic error and the corresponding local error indicators. The local indicators provide again a very accurate description of the local distribution of the algebraic error; however, the evaluation of the error estimators is very costly because of $\nu = 50$ additional algebraic iterations.

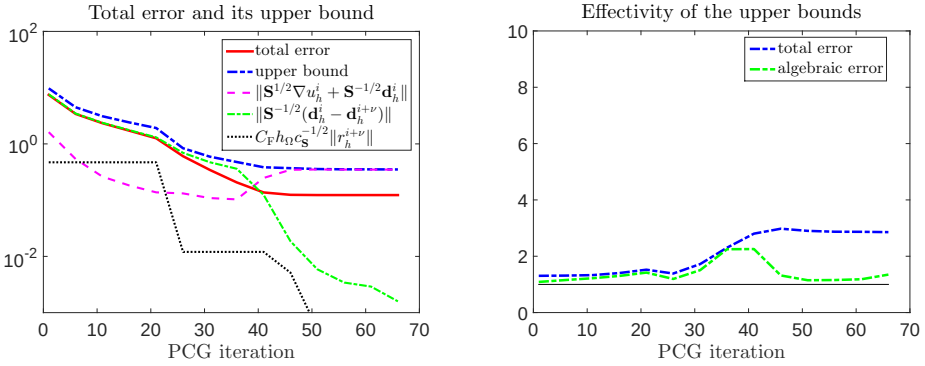

Figure 15: Problem with inhomogeneous diffusion tensor: total error $\|\mathbf{S}^{1/2}\nabla(u-u_h^i)\|$, the upper bound $\eta^{i,\nu}_{\text{total}}$ of Theorem 1 modified for the test problem (7.6), and the corresponding error indicators $\|\mathbf{S}^{1/2}\nabla u_h^i + \mathbf{S}^{-1/2}\mathbf{d}_h^i\|$, $\|\mathbf{S}^{-1/2}(\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i)\|$, and $C_{\mathrm{F}} h_\Omega c_{\mathbf{S}}^{-1/2}\|r_h^{i+\nu}\|$ (left); effectivity indices (7.7) of the upper bounds (right).

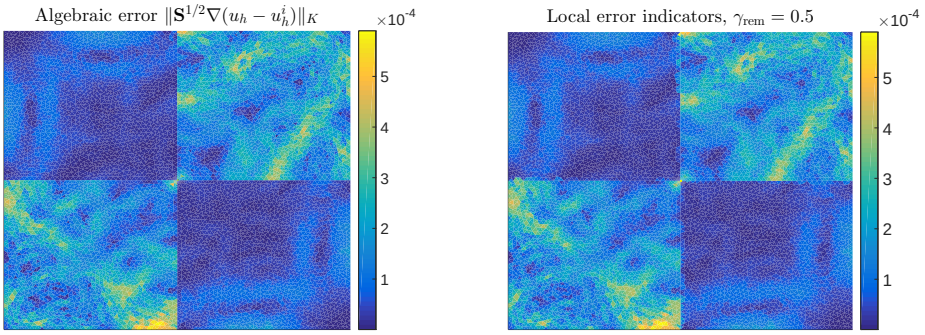

Figure 16: Problem with inhomogeneous diffusion tensor, iteration $i = 50$: elementwise distribution of the algebraic error $\|\mathbf{S}^{1/2}\nabla(u_h - u_h^i)\|_K$ and the local algebraic error indicators $\|\mathbf{S}^{-1/2}(\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i)\|_K + C_{\mathrm{F}} h_\Omega c_{\mathbf{S}}^{-1/2}\|r_h^{i+\nu}\|_K$. The value of ν, $\nu = 50$, is determined by (7.3c) with $\gamma_{\text{rem}} = 0.5$.

8 Conclusions and open questions

We have exposed in this paper in detail the methodology of $\mathbf{H}(\mathrm{div},\Omega)$-conforming flux and $H^1_0(\Omega)$-conforming residual reconstructions for estimating total, algebraic, and discretization errors for finite element discretizations and iterative algebraic solvers. The proposed upper and lower bounds are guaranteed and they contain no undetermined constants. We have used them for proposing stopping criteria for algebraic solvers that balance the algebraic and discretization errors and avoid stopping the algebraic iterations prematurely. As demonstrated on the model problems, they can practically localize very well the distribution of all errors and they can also avoid a possible local dominance of the algebraic error. The results provide a rigorous background for error estimators that can be extended to various problems and discretization techniques, including nonlinearity; see [24] for nonlinear problems and [59, 13] for unsteady nonlinear problems in an industrial application.

One part of the cost to be paid consists in a possibly nonnegligible amount of additional algebraic iterations that need to be performed. We have studied and reported this cost on two model examples in a rather unfavorable setting without a powerful preconditioner that would ensure very fast convergence and decrease this part of the cost to minimum. We believe that the presented methodology can be useful for many practical problems. Nevertheless, finding less costly alternatives within the presented framework is highly desirable and it represents one of our active research directions.

Acknowledgment. The authors wish to thank Ivana Pultarová, in particular for pointing out to us the inequality (5.9) including its proof. The authors are also grateful to anonymous referees for their numerous helpful comments.

A Details of the flux reconstruction

In this appendix we present the construction of the flux $\mathbf{d}_h^i$. It follows [24, Section 6.2.4] (see also [18, 10]) with the difference in the construction of the algebraic residual representation r_h^i satisfying (2.7), which allows to bound the algebraic error in Theorem 3.

For $K \in \mathcal{T}_h$, let $\mathbf{RTN}_{p'}(K) \equiv [\mathbb{P}_{p'}(K)]^d + \mathbf{x}\mathbb{P}_{p'}(K)$ be the Raviart–Thomas–Nédélec finite element space of order $p' \geq 0$. We set

$$\mathbf{RTN}_{p'}^{-1}(\mathcal{T}_h) \equiv \left\{ \mathbf{v}_h \in [L^2(\Omega)]^d, \mathbf{v}_h|_K \in \mathbf{RTN}_{p'}(K) \quad \forall K \in \mathcal{T}_h \right\}$$

and $\mathbf{RTN}_{p'}(\mathcal{T}_h) \equiv \mathbf{RTN}_{p'}^{-1}(\mathcal{T}_h) \cap \mathbf{H}(\mathrm{div},\Omega)$. We use a similar notation for these spaces on various patches. Let $\mathbf{RTN}_{p'}^{\mathrm{N},0}(\mathcal{T}_\mathtt{a})$ be the subspace of $\mathbf{RTN}_{p'}(\mathcal{T}_\mathtt{a})$ with zero normal flux through the boundary $\partial\omega_\mathtt{a}$ for $\mathtt{a} \in \mathcal{V}_h^{\mathrm{int}}$ and through $\partial\omega_\mathtt{a} \backslash \partial\Omega$ for $\mathtt{a} \in \mathcal{V}_h^{\mathrm{ext}}$ (corresponding to a homogeneous Neumann condition). Let $\mathbb{P}^*_{p'}(\mathcal{T}_\mathtt{a})$ be spanned by piecewise p'th order polynomials on $\mathcal{T}_\mathtt{a}$, with zero mean on $\mathcal{T}_\mathtt{a}$ when $\mathtt{a} \in \mathcal{V}_h^{\mathrm{int}}$.

For all vertices $\mathtt{a} \in \mathcal{V}_h$, we first solve the following mixed finite element problems on the patches $\mathcal{T}_\mathtt{a}$: find $\mathbf{d}_{h,\mathtt{a}}^i \in \mathbf{RTN}_{p'}^{\mathrm{N},0}(\mathcal{T}_\mathtt{a})$ and $q_{h,\mathtt{a}} \in \mathbb{P}^*_{p'}(\mathcal{T}_\mathtt{a})$, $p' = p$

or $p' = p+1$, such that

$$(\mathbf{d}^i_{h,\mathsf{a}}, \mathbf{v}_h)_{\omega_\mathsf{a}} - (q_{h,\mathsf{a}}, \nabla\cdot \mathbf{v}_h)_{\omega_\mathsf{a}} = -(\psi_\mathsf{a}\nabla u^i_h, \mathbf{v}_h)_{\omega_\mathsf{a}}, \tag{A.1a}$$

$$(\nabla\cdot \mathbf{d}^i_{h,\mathsf{a}}, \chi_h)_{\omega_\mathsf{a}} = (f_h\psi_\mathsf{a} - \nabla u^i_h\cdot\nabla\psi_\mathsf{a}, \chi_h)_{\omega_\mathsf{a}} - (r^i_h\psi_\mathsf{a}, \chi_h)_{\omega_\mathsf{a}} \tag{A.1b}$$

for all $(\mathbf{v}_h, \chi_h) \in \mathbf{RTN}^{\mathrm{N},0}_{p'}(\mathcal{T}_\mathsf{a}) \times \mathbb{P}^*_{p'}(\mathcal{T}_\mathsf{a})$. Then we set

$$\mathbf{d}^i_h \equiv \sum_{\mathsf{a}\in\mathcal{V}_h} \mathbf{d}^i_{h,\mathsf{a}}. \tag{A.1c}$$

We typically choose f_h to be the $L^2(\Omega)$-orthogonal projection of f onto the space of the piecewise polynomials of degree p', and $r^i_h \in \mathbb{P}_p(\mathcal{T}_h)$; see Section 5.1. Since $\psi_\mathsf{a} \in V_h$, (2.8) gives the Neumann compatibility condition of the problem (A.1a)–(A.1b),

$$(\nabla u^i_h, \nabla\psi_\mathsf{a})_{\omega_\mathsf{a}} = (f, \psi_\mathsf{a})_{\omega_\mathsf{a}} - (r^i_h, \psi_\mathsf{a})_{\omega_\mathsf{a}}.$$

Consequently, we can in (A.1b) take all test functions $\chi_h \in \mathbb{P}_{p'}(\mathcal{T}_\mathsf{a})$, which allows to show that $\mathbf{d}^i_h$ given by (A.1) satisfies (4.2), i.e., that $\nabla\cdot\mathbf{d}^i_h = f_h - r^i_h$ holds. Indeed, let $K \in \mathcal{T}_h$ and let $v_h \in \mathbb{P}_{p'}(K)$ be fixed. Since $\sum_{\mathsf{a}\in\mathcal{V}_h}\psi_\mathsf{a}|_K = 1$ and $\sum_{\mathsf{a}\in\mathcal{V}_h}\nabla\psi_\mathsf{a}|_K = 0$ (ψ_a form a partition of unity on K), we infer

$$\begin{aligned}(\nabla\cdot\mathbf{d}^i_h, v_h)_K &= \sum_{\mathsf{a}\in\mathcal{V}_h}(\nabla\cdot\mathbf{d}^i_{h,\mathsf{a}}, v_h)_K = \sum_{\mathsf{a}\in\mathcal{V}_h}\left[(f_h\psi_\mathsf{a} - \nabla u^i_h\cdot\nabla\psi_\mathsf{a}, v_h)_K - (r^i_h\psi_\mathsf{a}, v_h)_K\right]\\ &= (f_h, v_h)_K - (r^i_h, v_h)_K,\end{aligned}$$

and (4.2) is proved as $f_h - r^i_h \in \mathbb{P}_{p'}(\mathcal{T}_h)$.

We now briefly comment on the algorithmic construction of $\mathbf{d}^i_h$ in (A.1). Denote by by $\mathbf{\Phi}_\mathsf{a}$ the basis of $\mathbf{RTN}^{\mathrm{N},0}_{p'}(\mathcal{T}_\mathsf{a})$, and by $\widetilde{\mathcal{X}}_\mathsf{a}$ the basis of $\mathbb{P}^*_{p'}(\mathcal{T}_\mathsf{a})$, Then we construct $\mathbf{d}^i_h$ as

$$\mathbf{d}^i_h = \sum_{\mathsf{a}\in\mathcal{V}_h} \mathbf{\Phi}_\mathsf{a}\overline{\mathsf{D}}^i_\mathsf{a},$$

where $\overline{\mathsf{D}}^i_\mathsf{a}$ forms the part of the vector D^i_a solving the algebraic form of (A.1a)–(A.1b)

$$\mathsf{K}_\mathsf{a}\mathsf{D}^i_\mathsf{a} = \mathsf{E}^i_\mathsf{a}, \qquad \mathsf{K}_\mathsf{a} = \begin{bmatrix}\overline{\mathsf{K}}_\mathsf{a} & -\widetilde{\mathsf{K}}_\mathsf{a}\\ (\widetilde{\mathsf{K}}_\mathsf{a})^T & 0\end{bmatrix}, \qquad \mathsf{D}^i_\mathsf{a} = \begin{bmatrix}\overline{\mathsf{D}}^i_\mathsf{a}\\ \underline{\mathsf{D}}^i_\mathsf{a}\end{bmatrix}. \tag{A.2}$$

Here $\left(\overline{\mathsf{K}}_\mathsf{a}\right)_{kj} = (\boldsymbol{\phi}_j\,,\,\boldsymbol{\phi}_k)_{\omega_\mathsf{a}}$ and $\left(\widetilde{\mathsf{K}}_\mathsf{a}\right)_{k\ell} = (\widetilde{\chi}_\ell\,,\,\nabla\cdot\boldsymbol{\phi}_k)_{\omega_\mathsf{a}}$ with $\boldsymbol{\phi}_j, \boldsymbol{\phi}_k \in \mathbf{\Phi}_\mathsf{a}$, $\widetilde{\chi}_\ell \in \widetilde{\mathcal{X}}_\mathsf{a}$. The right-hand side vector is given as

$$\mathsf{E}^i_\mathsf{a} = \mathsf{E}_{\mathsf{a},f} - \mathsf{E}_{\mathsf{a},u^i_h} - \mathsf{E}_{\mathsf{a},r^i_h} = \begin{bmatrix}0\\ \underline{\mathsf{E}}_{\mathsf{a},f}\end{bmatrix} - \begin{bmatrix}\overline{\mathsf{E}}_{\mathsf{a},u^i_h}\\ \underline{\mathsf{E}}_{\mathsf{a},u^i_h}\end{bmatrix} - \begin{bmatrix}0\\ \underline{\mathsf{E}}_{\mathsf{a},r^i_h}\end{bmatrix},$$

where

$$(\overline{\mathsf{E}}_{\mathsf{a},u^i_h})_k = (\psi_\mathsf{a}\nabla u^i_h, \boldsymbol{\phi}_k)_{\omega_\mathsf{a}}, \qquad \boldsymbol{\phi}_k \in \mathbf{\Phi}_\mathsf{a},$$

$$(\underline{\mathsf{E}}_{\mathsf{a},f})_\ell = (f\psi_\mathsf{a}, \widetilde{\chi}_\ell)_{\omega_\mathsf{a}}, \quad (\underline{\mathsf{E}}_{\mathsf{a},u^i_h})_\ell = (\nabla u^i_h\cdot\nabla\psi_\mathsf{a}, \widetilde{\chi}_\ell)_{\omega_\mathsf{a}}, \quad (\underline{\mathsf{E}}_{\mathsf{a},r^i_h})_\ell = (r^i_h\psi_\mathsf{a}, \widetilde{\chi}_\ell)_{\omega_\mathsf{a}}, \quad \widetilde{\chi}_\ell \in \widetilde{\mathcal{X}}_\mathsf{a}.$$

Since $u_h^i = \Psi \mathsf{U}^i$, where, recall, Ψ is the basis of V_h, we have $u_h^i|_{\omega_\mathsf{a}} = \Psi_\mathsf{a} \mathsf{U}_\mathsf{a}^i$ for $\Psi_\mathsf{a} \subset \Psi$ a subset of basis functions that are nonvanishing on ω_a and U_a^i the associated entries of U^i. Then

$$\mathsf{E}_{\mathsf{a},u_h^i} = \mathbf{E}_{\mathsf{a},\Psi_\mathsf{a}} \mathsf{U}_\mathsf{a}^i, \quad \mathbf{E}_{\mathsf{a},\Psi_\mathsf{a}} = \begin{bmatrix} \overline{\mathbf{E}}_{\mathsf{a},\Psi_\mathsf{a}} \\ \underline{\mathbf{E}}_{\mathsf{a},\Psi_\mathsf{a}} \end{bmatrix}, \qquad \begin{aligned} \left(\overline{\mathbf{E}}_{\mathsf{a},\Psi_\mathsf{a}}\right)_{kj} &= (\psi_\mathsf{a} \nabla \psi_j, \boldsymbol{\phi}_k)_{\omega_\mathsf{a}}, \\ \left(\underline{\mathbf{E}}_{\mathsf{a},\Psi_\mathsf{a}}\right)_{\ell j} &= (\nabla \psi_j \cdot \nabla \psi_\mathsf{a}, \widetilde{\chi}_\ell)_{\omega_\mathsf{a}}, \end{aligned}$$

where $\psi_j \in \Psi_\mathsf{a}$, $\boldsymbol{\phi}_k \in \boldsymbol{\Phi}_\mathsf{a}$, $\widetilde{\chi}_\ell \in \widetilde{\mathcal{X}}_\mathsf{a}$. Similarly, denoting by $\mathcal{X}_\mathsf{a}$ the basis of $\mathbb{P}_p(\mathcal{T}_\mathsf{a})$, we have for the coefficient vector $\widehat{\mathsf{R}}_\mathsf{a}^i$ such that $r_h^i|_{\omega_\mathsf{a}} = \mathcal{X}_\mathsf{a} \widehat{\mathsf{R}}_\mathsf{a}^i$,

$$\mathsf{E}_{\mathsf{a},r_h^i} = \mathbf{E}_{\mathsf{a},\mathcal{X}_\mathsf{a}} \widehat{\mathsf{R}}_\mathsf{a}^i, \qquad \mathbf{E}_{\mathsf{a},\mathcal{X}_\mathsf{a}} = \begin{bmatrix} 0 \\ \underline{\mathbf{E}}_{\mathsf{a},\mathcal{X}_\mathsf{a}} \end{bmatrix}, \qquad \left(\underline{\mathbf{E}}_{\mathsf{a},\mathcal{X}_\mathsf{a}}\right)_{\ell j} = (\chi_j \psi_\mathsf{a}, \widetilde{\chi}_\ell)_{\omega_\mathsf{a}},$$

where $\chi_j \in \mathcal{X}_\mathsf{a}$, $\widetilde{\chi}_\ell \in \widetilde{\mathcal{X}}_\mathsf{a}$. Consequently, the vector D_a^i can be assembled as

$$\mathsf{D}_\mathsf{a}^i = \mathbf{K}_\mathsf{a}^{-1} \mathsf{E}_{\mathsf{a},f} - \left(\mathbf{K}_\mathsf{a}^{-1} \mathbf{E}_{\mathsf{a},\Psi_\mathsf{a}}\right) \mathsf{U}_\mathsf{a}^i - \left(\mathbf{K}_\mathsf{a}^{-1} \mathbf{E}_{\mathsf{a},\mathcal{X}_\mathsf{a}}\right) \widehat{\mathsf{R}}_\mathsf{a}^i. \tag{A.3}$$

This means that we can solve the system with $\mathbf{K}_\mathsf{a}$ only once with multiple right-hand sides $[\mathsf{E}_{\mathsf{a},f}, \mathbf{E}_{\mathsf{a},\Psi_\mathsf{a}}, \mathbf{E}_{\mathsf{a},\mathcal{X}_\mathsf{a}}]$ prior the start of the iterative solution of (2.5) and, at any iteration i, get the local coefficients $\overline{\mathsf{D}}_\mathsf{a}^i$ of the flux reconstruction $\mathbf{d}_h^i$ simply by matrix-vector multiplication and summing the vectors. This is particularly appealing when the error estimator is evaluated many times (e.g. when many iterations of the algebraic solver are performed). Note that assembling $\mathbf{K}_\mathsf{a}$, $\mathsf{E}_{\mathsf{a},f}$, $\mathbf{E}_{\mathsf{a},\Psi_\mathsf{a}}$, $\mathbf{E}_{\mathsf{a},\Psi_\mathsf{a}}$, $\mathsf{a} \in \mathcal{V}_h$, and solving the systems corresponding to (A.3) can be done in parallel (indeed, the individual patch problems (A.2) are mutually independent). Also, this can be done independently of assembling the system (2.5).

B Efficiency of the total error bound

We prove in this appendix the global and local efficiency of the upper bound of Theorem 1, which follows and extends the results in [24, 25, 46]. To simplify the presentation, we require that the source term f is piecewise polynomial, $f \in \mathbb{P}_{p'-1}(\mathcal{T}_h)$. Consequently, we choose $f_h = f$, and the oscillation term vanishes, $\eta_{\mathrm{osc}} = 0$.

The following lemma extends [14, Theorem 3.1] and [9, p. 1191] (see also [25, Lemma 3.12]) to the inexact algebraic solver case considered in this paper. Recall the space $H^1_*(\omega_\mathsf{a})$ introduced in (4.11).

Lemma 1. *Let $\mathsf{a} \in \mathcal{V}_h$ and let $m_\mathsf{a} \in H^1_*(\omega_\mathsf{a})$ be the solution of*

$$(\nabla m_\mathsf{a}, \nabla v)_{\omega_\mathsf{a}} = (f, \psi_\mathsf{a} v)_{\omega_\mathsf{a}} - \left(\nabla u_h^i, \nabla(\psi_\mathsf{a} v)\right)_{\omega_\mathsf{a}} - \left(r_h^i, \psi_\mathsf{a} v\right)_{\omega_\mathsf{a}} \qquad \forall v \in H^1_*(\omega_\mathsf{a}). \tag{B.1}$$

Then there holds

$$\|\nabla m_\mathsf{a}\|_{\omega_\mathsf{a}} \le C_{\mathrm{cont,PF},\omega_\mathsf{a}} \left(\|\nabla(u - u_h^i)\|_{\omega_\mathsf{a}} + \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_{\omega_\mathsf{a}}\right) + C_{\mathrm{PF},\omega_\mathsf{a}} h_{\omega_\mathsf{a}} \|r_h^{i+\nu}\|_{\omega_\mathsf{a}}.$$

Proof. From (B.1) and since, for $v \in H^1_*(\omega_\mathsf{a})$, $\psi_\mathsf{a} v \in H^1_0(\omega_\mathsf{a})$, we have, employing (2.2),

$$(\nabla m_\mathsf{a}, \nabla v)_{\omega_\mathsf{a}} = \left(\nabla(u - u_h^i), \nabla(\psi_\mathsf{a} v)\right)_{\omega_\mathsf{a}} - \left(r_h^i, \psi_\mathsf{a} v\right)_{\omega_\mathsf{a}}.$$

The Cauchy–Schwarz inequality and the bound (4.13) give

$$\big(\nabla(u-u_h^i),\nabla(\psi_{\mathtt{a}}v)\big)_{\omega_{\mathtt{a}}} \le \|\nabla(u-u_h^i)\|_{\omega_{\mathtt{a}}} C_{\mathrm{cont,PF},\omega_{\mathtt{a}}} \|\nabla v\|_{\omega_{\mathtt{a}}}.$$

Using (4.10), the Cauchy–Schwarz inequality, and (4.12),

$$\begin{aligned}
\big(r_h^i,\psi_{\mathtt{a}}v\big)_{\omega_{\mathtt{a}}} &= \big(\nabla\cdot\mathbf{d}_h^{i+\nu} - \nabla\cdot\mathbf{d}_h^i + r_h^{i+\nu},\psi_{\mathtt{a}}v\big)_{\omega_{\mathtt{a}}} \\
&= \big(-\mathbf{d}_h^{i+\nu}+\mathbf{d}_h^i,\nabla(\psi_{\mathtt{a}}v)\big)_{\omega_{\mathtt{a}}} + \big(r_h^{i+\nu},\psi_{\mathtt{a}}v\big)_{\omega_{\mathtt{a}}} \\
&\le \|\mathbf{d}_h^{i+\nu}-\mathbf{d}_h^i\|_{\omega_{\mathtt{a}}} C_{\mathrm{cont,PF},\omega_{\mathtt{a}}} \|\nabla v\|_{\omega_{\mathtt{a}}} + \|r_h^{i+\nu}\|_{\omega_{\mathtt{a}}}\|\psi_{\mathtt{a}}\|_{\infty,\omega_{\mathtt{a}}}\|v\|_{\omega_{\mathtt{a}}} \\
&\le \|\mathbf{d}_h^{i+\nu}-\mathbf{d}_h^i\|_{\omega_{\mathtt{a}}} C_{\mathrm{cont,PF},\omega_{\mathtt{a}}} \|\nabla v\|_{\omega_{\mathtt{a}}} + \|r_h^{i+\nu}\|_{\omega_{\mathtt{a}}} C_{\mathrm{PF},\omega_{\mathtt{a}}} h_{\omega_{\mathtt{a}}} \|\nabla v\|_{\omega_{\mathtt{a}}}.
\end{aligned}$$

Finally, using

$$\|\nabla m_{\mathtt{a}}\|_{\omega_{\mathtt{a}}} = \sup_{v\in H^1_*(\omega_{\mathtt{a}}),\|\nabla v\|=1} (\nabla m_{\mathtt{a}},\nabla v)_{\omega_{\mathtt{a}}}$$

and combining the above results yields the desired bound. □ □

The following crucial result has been shown in [9, Theorem 7] (see also [25, Corollary 3.16]) in the two-dimensional case. The three-dimensional proof is in [26, Corollary 3.3].

Lemma 2. *Let $\mathbf{d}_{h,\mathtt{a}}^i$ be given by (A.1) with $p'=p+1$ and let $m_{\mathtt{a}}$ be given by (B.1). Let $f\in\mathbb{P}_p(\mathcal{T}_h)$. Then there exists a constant $C_{\mathrm{st},\omega_{\mathtt{a}}}>0$ depending only on the shape of elements of the patch $\mathcal{T}_{\mathtt{a}}$ but not on their diameters such that*

$$\|\psi_{\mathtt{a}}\nabla u_h^i + \mathbf{d}_{h,\mathtt{a}}^i\|_{\omega_{\mathtt{a}}} \le C_{\mathrm{st},\omega_{\mathtt{a}}}\|\nabla m_{\mathtt{a}}\|_{\omega_{\mathtt{a}}}. \tag{B.2}$$

The constant $C_{\mathrm{st},\omega_{\mathtt{a}}}$ is not computable. It can, however, be bounded from above considering a finite-dimensional subspace of $H^1_*(\omega_{\mathtt{a}})$ and solving the discrete version of the problem (B.1); see [25, Lemma 3.23]. Hereafter we denote

$$C_{\mathrm{cont,PF}} \equiv \max_{\mathtt{a}\in\mathcal{V}_h} C_{\mathrm{cont,PF},\omega_{\mathtt{a}}}, \quad C_{\mathrm{PF}} \equiv \max_{\mathtt{a}\in\mathcal{V}_h} C_{\mathrm{PF},\omega_{\mathtt{a}}}, \quad C_{\mathrm{st}} \equiv \max_{\mathtt{a}\in\mathcal{V}_h} C_{\mathrm{st},\omega_{\mathtt{a}}}.$$

We now state the main result on the *global efficiency* of the estimators of Theorem 1, both for the *global stopping criteria* in the sense of [34, 24] and for the secure stopping criterion in the sense of (6.3), relying on the estimator μ_{total}^i of Theorem 2:

Theorem 7 (Global efficiency)**.** *Let the estimators of Theorem 1 satisfy the global stopping criteria*

$$C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\| \le \gamma_{\mathrm{rem}} \|\mathbf{d}_h^{i+\nu}-\mathbf{d}_h^i\|, \tag{B.3a}$$

$$\|\mathbf{d}_h^{i+\nu}-\mathbf{d}_h^i\| \le \gamma_{\mathrm{alg}} \|\nabla u_h^i + \mathbf{d}_h^i\| \tag{B.3b}$$

with positive parameters γ_{rem}, γ_{alg} such that

$$\gamma_{\mathrm{alg}} C_{\mathrm{st}} \left(C_{\mathrm{cont,PF}} + \gamma_{\mathrm{rem}} \frac{C_{\mathrm{PF}} \max_{\mathtt{a}\in\mathcal{V}_h} h_{\omega_{\mathtt{a}}}}{C_{\mathrm{F}} h_\Omega} \right) \le \frac{1}{2(d+1)}. \tag{B.4}$$

Alternatively, instead of (B.3)–(B.4), let

$$C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\| \le \gamma_{\mathrm{rem}} \|\mathbf{d}_h^{i+\nu}-\mathbf{d}_h^i\|, \tag{B.5a}$$

$$\|\mathbf{d}_h^{i+\nu}-\mathbf{d}_h^i\| \le \frac{\gamma_{\mathrm{alg}}}{(1+\gamma_{\mathrm{alg}}^2)^{1/2}}\, \mu_{\mathrm{total}}^i \tag{B.5b}$$

without any requirement on γ_{rem}, γ_{alg}, *supposing only*

$$\frac{C_{\mathrm{PF}} \max_{\mathtt{a}\in\mathcal{V}_h} h_{\omega_\mathtt{a}}}{C_{\mathrm{F}} h_\Omega} \le C_{\mathrm{cont,PF}}$$

that is typically satisfied, apart possibly the coarsest meshes. Let the assumptions of Lemma 2 hold. Then the upper bound of Theorem 1 is globally efficient,

$$\eta_{\mathrm{total}}^{i,\nu} \le C_{\mathrm{glob.\ eff.}} \|\nabla(u - u_h^i)\|$$

with the global efficiency constant

$$C_{\mathrm{glob.\ eff.}} \equiv (1 + \gamma_{\mathrm{alg}} + \gamma_{\mathrm{alg}}\gamma_{\mathrm{rem}}) 2(d+1) C_{\mathrm{st}} C_{\mathrm{cont,PF}}.$$

Recall that $\mathcal{V}_K$ stands for the vertices of the element K and that the functions $m_{h,\mathtt{a}}$ are specified in Theorem 2. Then the local version of Theorem 7 proving the *local efficiency* under the *local stopping criteria* is as follows:

Theorem 8 (Local efficiency)**.** *Let, for an element $K \in \mathcal{T}_h$, the estimators of Theorem 1 satisfy the local stopping criteria*

$$C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\|_{K'} \le \gamma_{\mathrm{rem},K} \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_{K'} \quad \forall K' \in \mathcal{T}_h \text{ such that } K' \cap K \neq \emptyset, \tag{B.6a}$$

$$\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_{\omega_\mathtt{a}} \le \gamma_{\mathrm{alg},K} \|\nabla u_h^i + \mathbf{d}_h^i\|_K \qquad \forall \mathtt{a} \in \mathcal{V}_K \tag{B.6b}$$

with positive parameters $\gamma_{\mathrm{rem},K}$, $\gamma_{\mathrm{alg},K}$ such that

$$\gamma_{\mathrm{alg},K} C_{\mathrm{st}} \left(C_{\mathrm{cont,PF}} + \gamma_{\mathrm{rem},K} \frac{C_{\mathrm{PF}} \max_{\mathtt{a}\in\mathcal{V}_K} h_{\omega_\mathtt{a}}}{C_{\mathrm{F}} h_\Omega} \right) \le \frac{1}{2(d+1)}. \tag{B.7}$$

Alternatively, instead of (B.6)–(B.7), let, for all $\mathtt{a} \in \mathcal{V}_K$,

$$C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\|_{\omega_\mathtt{a}} \le \gamma_{\mathrm{rem},K} \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_{\omega_\mathtt{a}}, \tag{B.8a}$$

$$\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_{\omega_\mathtt{a}} \le \frac{\gamma_{\mathrm{alg},K}}{(1 + \gamma_{\mathrm{alg},K}^2)^{1/2}} \frac{\|\nabla m_{h,\mathtt{a}}\|_{\omega_\mathtt{a}}}{C_{\mathrm{cont,PF},\omega_\mathtt{a}}}, \tag{B.8b}$$

without any requirement on $\gamma_{\mathrm{rem},K}$, $\gamma_{\mathrm{alg},K}$, supposing only

$$\frac{C_{\mathrm{PF}} \max_{\mathtt{a}\in\mathcal{V}_K} h_{\omega_\mathtt{a}}}{C_{\mathrm{F}} h_\Omega} \le C_{\mathrm{cont,PF}}$$

that is typically satisfied, apart possibly the coarsest meshes. Let the assumptions of Lemma 2 hold. Then we have the local efficiency of the upper bound,

$$\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_K + C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\|_K + \|\nabla u_h^i + \mathbf{d}_h^i\|_K \le C_{\mathrm{loc.\ eff.},K} \sum_{\mathtt{a}\in\mathcal{V}_K} \|\nabla(u - u_h^i)\|_{\omega_\mathtt{a}}$$

with the local efficiency constant

$$C_{\mathrm{loc.\ eff.},K} \equiv (1 + \gamma_{\mathrm{alg},K} + \gamma_{\mathrm{alg},K}\gamma_{\mathrm{rem},K}) 2 C_{\mathrm{st}} C_{\mathrm{cont,PF}}.$$

of Theorem 7. From the flux construction (A.1) of $\mathbf{d}_h^i$, using (B.2),

$$\begin{aligned}\|\nabla u_h^i + \mathbf{d}_h^i\|^2 &= \sum_{K\in\mathcal{T}_h} \Big\| \sum_{\mathtt{a}\in\mathcal{V}_K} (\psi_{\mathtt{a}} \nabla u_h^i + \mathbf{d}_{h,\mathtt{a}}^i)\Big\|_K^2 \\ &\le (d+1) \sum_{K\in\mathcal{T}_h} \sum_{\mathtt{a}\in\mathcal{V}_K} \|\psi_{\mathtt{a}} \nabla u_h^i + \mathbf{d}_{h,\mathtt{a}}^i\|_K^2 = (d+1) \sum_{\mathtt{a}\in\mathcal{V}_h} \|\psi_{\mathtt{a}} \nabla u_h^i + \mathbf{d}_{h,\mathtt{a}}^i\|_{\omega_{\mathtt{a}}}^2 \\ &\le (d+1) C_{\mathrm{st}}^2 \sum_{\mathtt{a}\in\mathcal{V}_h} \|\nabla m_{\mathtt{a}}\|_{\omega_{\mathtt{a}}}^2,\end{aligned}$$

as any element $K \in \mathcal{T}_h$ has $d+1$ vertices. From Lemma 1, we have

$$\begin{aligned}\Big[\sum_{\mathtt{a}\in\mathcal{V}_h} \|\nabla m_{\mathtt{a}}\|_{\omega_{\mathtt{a}}}^2\Big]^{1/2} &\le \Big[\sum_{\mathtt{a}\in\mathcal{V}_h} C_{\mathrm{cont,PF},\omega_{\mathtt{a}}}^2 \|\nabla(u-u_h^i)\|_{\omega_{\mathtt{a}}}^2\Big]^{1/2} \\ &+ \Big[\sum_{\mathtt{a}\in\mathcal{V}_h} C_{\mathrm{cont,PF},\omega_{\mathtt{a}}}^2 \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_{\omega_{\mathtt{a}}}^2\Big]^{1/2} + \Big[\sum_{\mathtt{a}\in\mathcal{V}_h} C_{\mathrm{PF},\omega_{\mathtt{a}}}^2 (h_{\omega_{\mathtt{a}}})^2 \|r_h^{i+\nu}\|_{\omega_{\mathtt{a}}}^2\Big]^{1/2}.\end{aligned}$$

Therefore, using $\Big[\sum_{\mathtt{a}\in\mathcal{V}_h} \|z\|_{\omega_{\mathtt{a}}}^2\Big]^{1/2} = (d+1)^{1/2}\|z\|$,

$$\begin{aligned}\|\nabla u_h^i + \mathbf{d}_h^i\| &\le (d+1) C_{\mathrm{st}} C_{\mathrm{cont,PF}} \|\nabla(u-u_h^i)\| \\ &+ (d+1) C_{\mathrm{st}} C_{\mathrm{cont,PF}} \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\| \;+\; (d+1) C_{\mathrm{st}} C_{\mathrm{PF}} \max_{\mathtt{a}\in\mathcal{V}_h} h_{\omega_{\mathtt{a}}} \|r_h^{i+\nu}\|. \quad \text{(B.9)}\end{aligned}$$

From the stopping criteria (B.3),

$$\begin{aligned}\|\nabla u_h^i + \mathbf{d}_h^i\| &\le (d+1) C_{\mathrm{st}} C_{\mathrm{cont,PF}} \|\nabla(u-u_h^i)\| \\ &+ (d+1)\gamma_{\mathrm{alg}} C_{\mathrm{st}} \left(C_{\mathrm{cont,PF}} + \gamma_{\mathrm{rem}} \frac{C_{\mathrm{PF}} \max_{\mathtt{a}\in\mathcal{V}_h} h_{\omega_{\mathtt{a}}}}{C_{\mathrm{F}} h_\Omega}\right) \|\nabla u_h^i + \mathbf{d}_h^i\|,\end{aligned}$$

and from (B.4),

$$\|\nabla u_h^i + \mathbf{d}_h^i\| \le 2(d+1) C_{\mathrm{st}} C_{\mathrm{cont,PF}} \|\nabla(u-u_h^i)\|.$$

Finally, we get the assertion for the stopping criteria (B.3),

$$\begin{aligned}\eta_{\mathrm{total}}^{i,\nu} &= \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\| + C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\| + \|\nabla u_h^i + \mathbf{d}_h^i\| \\ &\le (1 + \gamma_{\mathrm{alg}} + \gamma_{\mathrm{alg}}\gamma_{\mathrm{rem}}) \|\nabla u_h^i + \mathbf{d}_h^i\| \le C_{\mathrm{glob.\ eff.}} \|\nabla(u-u_h^i)\|.\end{aligned}$$

The efficiency under the stopping criteria (B.5) actually does not request any restrictive assumptions of the form (B.4). Using (B.5b) and the bound of Theorem 2,

$$\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\| \le \frac{\gamma_{\mathrm{alg}}}{(1+\gamma_{\mathrm{alg}}^2)^{1/2}} \|\nabla(u-u_h^i)\|.$$

Now a combination with (B.9) and (B.5a) gives

$$\begin{aligned}&\|\nabla u_h^i + \mathbf{d}_h^i\| \le (d+1) C_{\mathrm{st}} C_{\mathrm{cont,PF}} \|\nabla(u-u_h^i)\| \\ &+ (d+1) \frac{\gamma_{\mathrm{alg}}}{(1+\gamma_{\mathrm{alg}}^2)^{1/2}} C_{\mathrm{st}} \left(C_{\mathrm{cont,PF}} + \gamma_{\mathrm{rem}} \frac{C_{\mathrm{PF}} \max_{\mathtt{a}\in\mathcal{V}_h} h_{\omega_{\mathtt{a}}}}{C_{\mathrm{F}} h_\Omega}\right) \|\nabla(u-u_h^i)\|,\end{aligned}$$

so that the assertion for the stopping criteria (B.5) follows with the constant

$$(d+1)C_{\mathrm{st}}\left(C_{\mathrm{cont,PF}}+\frac{\gamma_{\mathrm{alg}}}{(1+\gamma_{\mathrm{alg}}^2)^{1/2}}C_{\mathrm{cont,PF}}+\gamma_{\mathrm{rem}}\frac{\gamma_{\mathrm{alg}}}{(1+\gamma_{\mathrm{alg}}^2)^{1/2}}\frac{C_{\mathrm{PF}}\max_{\mathtt{a}\in\mathcal{V}_h}h_{\omega_{\mathtt{a}}}}{C_{\mathrm{F}}h_\Omega}\right)$$
$$\leq (1+\gamma_{\mathrm{alg}}+\gamma_{\mathrm{alg}}\gamma_{\mathrm{rem}})(d+1)C_{\mathrm{st}}C_{\mathrm{cont,PF}}\leq\frac{C_{\text{glob. eff.}}}{2}.$$

□ □

of Theorem 8. For the proof of the local efficiency, we first note that

$$\|\nabla u_h^i+\mathbf{d}_h^i\|_K\leq\sum_{\mathtt{a}\in\mathcal{V}_K}\|\psi_{\mathtt{a}}\nabla u_h^i+\mathbf{d}_{h,\mathtt{a}}^i\|_{\omega_{\mathtt{a}}}\leq\sum_{\mathtt{a}\in\mathcal{V}_K}C_{\mathrm{st},\omega_{\mathtt{a}}}\|\nabla m_{\mathtt{a}}\|_{\omega_{\mathtt{a}}}.$$

From Lemma 1,

$$\|\nabla u_h^i+\mathbf{d}_h^i\|_K\leq C_{\mathrm{st}}C_{\mathrm{cont,PF}}\sum_{\mathtt{a}\in\mathcal{V}_K}\|\nabla(u-u_h^i)\|_{\omega_{\mathtt{a}}}$$
$$+C_{\mathrm{st}}C_{\mathrm{cont,PF}}\sum_{\mathtt{a}\in\mathcal{V}_K}\|\mathbf{d}_h^{i+\nu}-\mathbf{d}_h^i\|_{\omega_{\mathtt{a}}}+C_{\mathrm{st}}C_{\mathrm{PF}}\max_{\mathtt{a}\in\mathcal{V}_K}h_{\omega_{\mathtt{a}}}\sum_{\mathtt{a}\in\mathcal{V}_K}\|r_h^{i+\nu}\|_{\omega_{\mathtt{a}}}.\quad\text{(B.10)}$$

Thus, under the stopping criteria (B.6),

$$\|\nabla u_h^i+\mathbf{d}_h^i\|_K\leq C_{\mathrm{st}}C_{\mathrm{cont,PF}}\sum_{\mathtt{a}\in\mathcal{V}_K}\|\nabla(u-u_h^i)\|_{\omega_{\mathtt{a}}}$$
$$+(d+1)C_{\mathrm{st}}\gamma_{\mathrm{alg},K}\left(C_{\mathrm{cont,PF}}+\gamma_{\mathrm{rem},K}\frac{C_{\mathrm{PF}}\max_{\mathtt{a}\in\mathcal{V}_K}h_{\omega_{\mathtt{a}}}}{C_{\mathrm{F}}h_\Omega}\right)\|\nabla u_h^i+\mathbf{d}_h^i\|_K.$$

From (B.7), we further obtain

$$\|\nabla u_h^i+\mathbf{d}_h^i\|_K\leq 2C_{\mathrm{st}}C_{\mathrm{cont,PF}}\sum_{\mathtt{a}\in\mathcal{V}_K}\|\nabla(u-u_h^i)\|_{\omega_{\mathtt{a}}},$$

so that finally

$$\|\mathbf{d}_h^{i+\nu}-\mathbf{d}_h^i\|_K+C_{\mathrm{F}}h_\Omega\|r_h^{i+\nu}\|_K+\|\nabla u_h^i+\mathbf{d}_h^i\|_K$$
$$\leq(1+\gamma_{\mathrm{alg},K}+\gamma_{\mathrm{alg},K}\gamma_{\mathrm{rem},K})\|\nabla u_h^i+\mathbf{d}_h^i\|$$
$$\leq C_{\text{loc. eff.},K}\sum_{\mathtt{a}\in\mathcal{V}_K}\|\nabla(u-u_h^i)\|_{\omega_{\mathtt{a}}}.$$

Let $\widetilde{m}_{\mathtt{a}}\in H^1_*(\omega_{\mathtt{a}})$ be the solution of

$$(\nabla\widetilde{m}_{\mathtt{a}},\nabla v)_{\omega_{\mathtt{a}}}=(f,\psi_{\mathtt{a}}v)_{\omega_{\mathtt{a}}}-\big(\nabla u_h^i,\nabla(\psi_{\mathtt{a}}v)\big)_{\omega_{\mathtt{a}}}\qquad\forall v\in H^1_*(\omega_{\mathtt{a}}),$$

in the continuous counterpart to $m_{h,\mathtt{a}}$ of Theorem 2 and similarly to (B.1). The fact that $m_{h,\mathtt{a}}$ is a projection of $\widetilde{m}_{\mathtt{a}}$ from $H^1_*(\omega_{\mathtt{a}})$ onto $W_h^{\mathtt{a}}$ gives $\|\nabla m_{h,\mathtt{a}}\|_{\omega_{\mathtt{a}}}\leq\|\nabla\widetilde{m}_{\mathtt{a}}\|_{\omega_{\mathtt{a}}}$. Proceeding as in the proof of Lemma 1 with $r_h^i=0$, we get the inequality $\|\nabla\widetilde{m}_{\mathtt{a}}\|_{\omega_{\mathtt{a}}}\leq C_{\mathrm{cont,PF},\omega_{\mathtt{a}}}\|\nabla(u-u_h^i)\|_{\omega_{\mathtt{a}}}$, so that

$$\|\nabla m_{h,\mathtt{a}}\|_{\omega_{\mathtt{a}}}\leq C_{\mathrm{cont,PF},\omega_{\mathtt{a}}}\|\nabla(u-u_h^i)\|_{\omega_{\mathtt{a}}}.$$

Thus, under the secure local stopping criterion (B.8b), we obtain

$$\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_{\omega_\mathtt{a}} \leq \frac{\gamma_{\mathrm{alg},K}}{(1+\gamma_{\mathrm{alg},K}^2)^{1/2}} \|\nabla(u - u_h^i)\|_{\omega_\mathtt{a}},$$

and, employing (B.10) and (B.8a),

$$\|\nabla u_h^i + \mathbf{d}_h^i\|_K \leq C_{\mathrm{st}} C_{\mathrm{cont,PF}} \sum_{\mathtt{a}\in\mathcal{V}_K} \|\nabla(u - u_h^i)\|_{\omega_\mathtt{a}}$$
$$+C_{\mathrm{st}} \frac{\gamma_{\mathrm{alg},K}}{(1+\gamma_{\mathrm{alg},K}^2)^{1/2}} \left(C_{\mathrm{cont,PF}} + \gamma_{\mathrm{rem},K} \frac{C_{\mathrm{PF}} \max_{\mathtt{a}\in\mathcal{V}_K} h_{\omega_\mathtt{a}}}{C_{\mathrm{F}} h_\Omega} \right) \sum_{\mathtt{a}\in\mathcal{V}_K} \|\nabla(u - u_h^i)\|_{\omega_\mathtt{a}}.$$

The claim in this case thus follows from

$$C_{\mathrm{st}} \left(C_{\mathrm{cont,PF}} + \frac{\gamma_{\mathrm{alg},K}}{(1+\gamma_{\mathrm{alg},K}^2)^{1/2}} C_{\mathrm{cont,PF}} + \gamma_{\mathrm{rem},K} \frac{\gamma_{\mathrm{alg},K}}{(1+\gamma_{\mathrm{alg},K}^2)^{1/2}} \frac{C_{\mathrm{PF}} \max_{\mathtt{a}\in\mathcal{V}_K} h_{\omega_\mathtt{a}}}{C_{\mathrm{F}} h_\Omega} \right)$$
$$\leq (1 + \gamma_{\mathrm{alg},K} + \gamma_{\mathrm{alg},K}\gamma_{\mathrm{rem},K}) C_{\mathrm{st}} C_{\mathrm{cont,PF}} \leq \frac{C_{\mathrm{loc.\ eff.},K}}{2}.$$

□ □

References

[1] M. Ainsworth, *Robust a posteriori error estimation for nonconforming finite element approximation*, SIAM J. Numer. Anal., 42 (2005), pp. 2320–2341.

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

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

[4] I. Babuška and T. Strouboulis, *The finite element method and its reliability*, Numerical Mathematics and Scientific Computation, The Clarendon Press Oxford University Press, New York, 2001.

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

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

[7] R. Becker, S. Mao, and Z. Shi, *A convergent nonconforming adaptive finite element method with quasi-optimal complexity*, SIAM J. Numer. Anal., 47 (2010), pp. 4639–4659.

[8] M. Berndt, T. A. Manteuffel, and S. F. McCormick, *Local error estimates and adaptive refinement for first-order system least squares (FOSLS)*, Electron. Trans. Numer. Anal., 6 (1997), pp. 35–43. Special issue on multilevel methods (Copper Mountain, CO, 1997).

[9] D. Braess, V. Pillwein, and J. Schöberl, *Equilibrated residual error estimates are p-robust*, Comput. Methods Appl. Mech. Engrg., 198 (2009), pp. 1189–1197.

[10] D. Braess and J. Schöberl, *Equilibrated residual error estimator for edge elements*, Math. Comp., 77 (2008), pp. 651–672.

[11] C. Burstedde and A. Kunoth, *A wavelet-based nested iteration-inexact conjugate gradient algorithm for adaptively solving elliptic PDEs*, Numer. Algorithms, 48 (2008), pp. 161–188.

[12] D. Calvetti, S. Morigi, L. Reichel, and F. Sgallari, *Computable error bounds and estimates for the conjugate gradient method*, Numer. Algorithms, 25 (2000), pp. 75–88.

[13] C. Cancès, I. S. Pop, and M. Vohralík, *An a posteriori error estimate for vertex-centered finite volume discretizations of immiscible incompressible two-phase flow*, Math. Comp., 83 (2014), pp. 153–188.

[14] C. Carstensen and S. A. Funken, *Fully reliable localized error control in the FEM*, SIAM J. Sci. Comput., 21 (1999/00), pp. 1465–1484.

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

[16] ——, *Linear and nonlinear functional analysis with applications*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013.

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

[18] P. Destuynder and B. Métivet, *Explicit error bounds in a conforming finite element method*, Math. Comp., 68 (1999), pp. 1379–1396.

[19] P. Deuflhard, *Cascadic conjugate gradient methods for elliptic partial differential equations: algorithm and numerical results*, in Domain decomposition methods in scientific and engineering computing (University Park, PA, 1993), vol. 180 of Contemp. Math., American Mathematical Society, Providence, RI, 1994, pp. 29–42.

[20] V. Dolean, P. Jolivet, and F. Nataf, *An Introduction to Domain Decomposition Methods : Algorithms, Theory, and Parallel Implementation*, Other Titles in Applied Mathematics, SIAM, Philadelphia, 2015.

[21] V. Dolejší, I. Šebestová, and M. Vohralík, *Algebraic and discretization error estimation by equilibrated fluxes for discontinuous Galerkin methods on nonmatching grids*, J. Sci. Comput., 64 (2015), pp. 1–34.

[22] V. Dolejší, A. Ern, and M. Vohralík, *hp-adaptation driven by polynomial-degree-robust a posteriori error estimates for elliptic problems*, SIAM J. Sci. Comput., 38 (2016), pp. A3220–A3246.

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

[24] A. Ern and M. Vohralík, *Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs*, SIAM J. Sci. Comput., 35 (2013), pp. A1761–A1791.

[25] ——, *Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations*, SIAM J. Numer. Anal., 53 (2015), pp. 1058–1081.

[26] ——, *Stable broken H^1 and $\boldsymbol{H}(\mathrm{div})$ polynomial extensions for polynomial-degree-robust potential and flux reconstruction in three space dimensions*. HAL Preprint 01422204, submitted for publication, 2016.

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

[28] G. H. Golub and G. Meurant, *Matrices, moments and quadrature*, in Numerical analysis 1993 (Dundee, 1993), vol. 303 of Pitman Res. Notes Math. Ser., Longman Sci. Tech., Harlow, 1994, pp. 105–156.

[29] ——, *Matrices, moments and quadrature. II. How to compute the norm of the error in iterative methods*, BIT, 37 (1997), pp. 687–705.

[30] G. H. Golub and Z. Strakoš, *Estimates in quadratic formulas*, Numer. Algorithms, 8 (1994), pp. 241–268.

[31] H. Harbrecht and R. Schneider, *On error estimation in finite element methods without having Galerkin orthogonality*, Berichtsreihe des SFB 611 457, Universität Bonn, 2009.

[32] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436.

[33] R. Hiptmair, *Operator preconditioning*, Comput. Math. Appl., 52 (2006), pp. 699–706.

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

[35] R. B. Kellogg, *On the Poisson equation with intersecting interfaces*, Applicable Anal., 4 (1974/75), pp. 101–129. Collection of articles dedicated to Nikolai Ivanovich Muskhelishvili.

[36] J. Liesen and Z. Strakoš, *Krylov subspace methods: principles and analysis*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2013.

[37] R. Luce and B. I. Wohlmuth, *A local a posteriori error estimator based on equilibrated fluxes*, SIAM J. Numer. Anal., 42 (2004), pp. 1394–1414.

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

[39] G. Meurant, *The computation of bounds for the norm of the error in the conjugate gradient algorithm*, Numer. Algorithms, 16 (1997), pp. 77–87 (1998).

[40] ——, *Numerical experiments in computing bounds for the norm of the error in the preconditioned conjugate gradient algorithm*, Numer. Algorithms, 22 (1999), pp. 353–365 (1999).

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

[42] G. Meurant and P. Tichý, *On computing quadrature-based bounds for the A-norm of the error in conjugate gradients*, Numer. Algorithms, 62 (2013), pp. 163–191.

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

[44] J. Papež, *Algebraic Error in Matrix Computations in the Context of Numerical Solution of Partial Differential Equations*, PhD thesis, Charles University, Prague, November 2016.

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

[46] J. Papež, U. Rüde, M. Vohralík, and B. Wohlmuth, *Guaranteed algebraic, discretization, and total error bounds in multigrid finite element solvers*. In preparation, 2016.

[47] A. T. Patera and E. M. Rønquist, *A general output bound result: application to discretization and iteration error estimation and control*, Math. Models Methods Appl. Sci., 11 (2001), pp. 685–712.

[48] L. E. Payne and H. F. Weinberger, *An optimal Poincaré inequality for convex domains*, Arch. Rational Mech. Anal., 5 (1960), pp. 286–292 (1960).

[49] R. Rannacher, *Error control in finite element computations. An introduction to error estimation and mesh-size adaptation*, in Error control and adaptivity in scientific computing (Antalya, 1998), vol. 536 of NATO Sci. Ser. C Math. Phys. Sci., Kluwer Acad. Publ., Dordrecht, 1999, pp. 247–278.

[50] K. Rektorys, *Variational methods in mathematics, science and engineering*, D. Reidel Publishing Co., Dordrecht-Boston, Mass., second ed., 1980. Translated from the Czech by Michael Basch.

[51] S. Repin, *A posteriori estimates for partial differential equations*, vol. 4 of Radon Series on Computational and Applied Mathematics, Walter de Gruyter GmbH & Co. KG, Berlin, 2008.

[52] V. V. Shaidurov, *Some estimates of the rate of convergence for the cascadic conjugate-gradient method*, Comput. Math. Appl., 31 (1996), pp. 161–171.

[53] D. J. Silvester and V. Simoncini, *An optimal iterative solver for symmetric indefinite systems stemming from mixed approximation*, ACM Trans. Math. Software, 37 (2011), pp. Art. 42, 22.

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

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

[56] ——, *Error estimation in preconditioned conjugate gradients*, BIT, 45 (2005), pp. 789–817.

[57] A. Veeser and R. Verfürth, *Poincaré constants for finite element stars*, IMA J. Numer. Anal., 32 (2012), pp. 30–47.

[58] R. Verfürth, *A posteriori error estimation techniques for finite element methods*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2013.

[59] M. Vohralík and M. F. Wheeler, *A posteriori error estimates, stopping criteria, and adaptivity for two-phase flows*, Comput. Geosci., 17 (2013), pp. 789–812.

[60] B. I. Wohlmuth and R. H. W. Hoppe, *A comparison of a posteriori error estimators for mixed finite element discretizations by Raviart-Thomas elements*, Math. Comp., 68 (1999), pp. 1347–1378.