A DISCRETIZATION-ACCURATE STOPPING CRITERION FOR ITERATIVE SOLVERS FOR FINITE ELEMENT APPROXIMATION*

ZHIQIANG CAI†, SHUHAO CAO‡, AND ROBERT D. FALGOUT§

Abstract. This paper introduces a discretization-accurate stopping criterion of symmetric iterative methods for solving systems of algebraic equations resulting from the finite element approximation. The stopping criterion consists of the evaluations of the discretization and the algebraic error estimators, that are based on the respective duality error estimator and the difference of two consecutive iterates. Iterations are terminated when the algebraic estimator is of the same magnitude as the discretization estimator. Numerical results for multigrid $V(1,1)$-cycle and symmetric Gauss-Seidel iterative methods are presented for the linear finite element approximation to the Poisson equations. A large reduction in computational cost is observed compared to the standard residual-based stopping criterion.

1. Introduction.

Consider the Dirichlet boundary value problem in a bounded polygonal/polyhedral domain $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ for the diffusion equation as follows:

$$\begin{cases} -\nabla\cdot(A\nabla u) = f, & \text{in } \Omega, \\ u = g, & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where A is a scalar diffusion coefficient, and the data $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$.

In practice, the system of algebraic equations resulting from the finite element approximation to (1.1) is often solved by iterative methods, e.g., Gauss-Seidel, conjugate gradient, multigrid methods, etc. Instead of having the exact solution $u_\mathcal{T}$ of the algebraic system at hand, $\bar{u}_\mathcal{T} := u_\mathcal{T}^{(k)}$ is the current output from an iterative solver, where k is the number of iterations. The total energy error of $\bar{u}_\mathcal{T}$ to the solution u of the continuous problem in (1.1) consists of both discretization and algebraic errors as follows:

$$\underbrace{\|u - \bar{u}_\mathcal{T}\|_A^2}_{\text{total error}} = \underbrace{\|u_\mathcal{T} - \bar{u}_\mathcal{T}\|_A^2}_{\text{algebraic error}} + \underbrace{\|u - u_\mathcal{T}\|_A^2}_{\text{discretization error}}, \tag{1.2}$$

where $\|\cdot\|_A$ is the energy norm associated with the problem in (1.1) (for the norm notations, see section 2).

The goal of this paper is to propose a stopping criterion for iterative solvers. To do so, we need to develop two error estimators for the respective discretization and algebraic errors. Since the discretization error is fixed for a given finite element space, (1.2) clearly indicates that the stopping criterion of the iterative solver is when the algebraic estimator is of the same magnitude as the discretization estimator, provided that both represent their error counterparts reliably.

Discretization error estimators for the exact finite element approximations have been intensively studied during the past four decades (see books [1, 26] and references

* This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 (LLNL-JRNL-789117). This work was supported in part by the National Science Foundation under grants DMS-1320608, DMS-1418934, and DMS-1522707.

† Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2067, zcai@math.purdue.edu.

‡ Department of Mathematics, University of California Irvine, Irvine, CA 92697, scao@math.uci.edu.

§ Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA 94551-0808, falgout2@llnl.gov.

therein). In the context of stopping criterion for iterative solvers, the residual-based a posteriori error estimator was employed for the conforming finite element approximation by several researchers (see, e.g., [5, 24, 3, 2, 4]); recovery-based estimators were used by Vohralik and his collaborators in [20] for the finite volume discretization and in [17] for the discontinuous finite element approximation on non-matching grids.

In this paper, we will adopt the equilibrated flux error estimator (see, e.g., [8, 16, 21, 25, 13]) for the discretization error. This is because the reliability bound of estimators of this type is constant free. Using this technique, a locally post-processed flux based on the iterate $\bar{u}_{\mathcal{T}}$ will be constructed. Unlike the exact finite element approximation $u_{\mathcal{T}}$, the local problems based on the current iterate $\bar{u}_{\mathcal{T}}$ on vertex patches are not consistent. To overcome this difficulty, we modify the local problems by adding back the algebraic errors. The resulting discretization error estimator plus the algebraic error is proved to be reliable and the reliability bound for the discretization estimator component is constant free (see Theorem 4).

To construct the algebraic error estimator, we first bound $\big\|u_{\mathcal{T}} - u_{\mathcal{T}}^{(k)}\big\|_A$ above by the energy norm of the difference of consecutive iterates, and the constant in the upper bound depends on the spectral radius of the error propagation operator (see Theorem 5). The unknown spectral radius is further approximated by the ratio of the l^2 norms of the residuals of consecutive iterates. The resulting algebraic error estimator is then shown to be reliable when sufficiently many iterations have been performed.

Lastly, in Section 6, based on the discretizaton and algebraic estimators, a new stopping criterion for a given linear solver is verified numerically by some test problems. The numerics shows promising results in that the bounds are independent of the coefficient jump ratio even without the quasi-monotonicy assumption [6, 18, 23] for the distribution of the diffusion coefficient A.

2. Finite element method and iterative solver. In this section, all preliminaries are presented. Denote $H^1(\Omega)$ with a specified boundary value as $H^1_g(\Omega) := \{v \in H^1(\Omega) : \, v = g \text{ on } \partial\Omega\}$, and then the variational problem of (1.1) is

$$\text{(2.1)} \qquad \text{Find } \; u \in H^1_g(\Omega) \; \text{ such that } \; \big(A\nabla u,\, \nabla v\big) = \big(f,\, v\big), \quad \forall v \in H^1_0(\Omega),$$

where $\big(\cdot,\, \cdot\big)$ denotes the L^2-inner product on the whole domain.

Let $\mathcal{T} = \{K\}$ be a triangulation of Ω using simplicial element, where $\mathcal{T}$ is assumed to be quasi-uniform and regular. For each $K \in \mathcal{T}$, $h_K := \operatorname{diam}(K) = O(|K|^{1/d})$. The set of all the vertices of this triangulation is denoted by $\mathcal{N}$. Throughout this paper, the term "face" is used to refer to the $(d-1)$-facet of a d-simplex in this triangulation $(d = 2, 3)$. For the $d = 2$ case, a face actually represents an edge. The set of all the interior faces is denoted by $\mathcal{F}$. For any $F \in \mathcal{F}$, $h_F := \operatorname{diam}(F) = O(|F|^{1/(d-1)})$. Each face $F \in \mathcal{F}$ is associated with a fixed unit normal $\boldsymbol{n}_F$ globally. For any function or distribution v well-defined on the two elements sharing a face F respectively, define $[\![v]\!]_F = v^- - v^+$ on an interior face. The v^- and v^+ are defined in the limiting sense of $v^\pm = \lim\limits_{\epsilon \to 0^\pm} v(\boldsymbol{x} + \epsilon \boldsymbol{n}_F)$. If F is a boundary face, the function v is extended by zero outside the domain to compute $[\![v]\!]_F$. For every geometrical object D and for every integer $k \geq 0$, $P_k(D)$ denotes the set of polynomials of degree $\leq k$ on D.

For the purpose of constructing the local error estimation procedure for the finite element approximation, notations of the following local geometric objects are used in this paper. First, denote by $\mathcal{N}_K$ the set of all the vertices of $K \in \mathcal{T}$. For any vertex

$\boldsymbol{z} \in \mathcal{N}$, denote by

$$\omega_{\boldsymbol{z}} := \bigcup\nolimits_{\{K\in\mathcal{T}:\, \boldsymbol{z}\in\mathcal{N}_K\}} K$$

as the vertex patch, which is the union of all elements sharing $\boldsymbol{z}$ as a common vertex. Now $\mathcal{T}_{\boldsymbol{z}}$ stands for the triangulation of this patch such that $\mathcal{T}_{\boldsymbol{z}} := \{K : K \subset \omega_{\boldsymbol{z}}\}$. Denote

$$\omega_K := \bigcup\nolimits_{\boldsymbol{z}\in\mathcal{N}_K} \omega_{\boldsymbol{z}}$$

as the element patch for K that contains all the elements sharing a vertex with K. For a face $F \in \mathcal{F}$, denote the face patch as

$$\omega_F := \bigcup\nolimits_{F\cap\partial K\neq\emptyset} K,$$

which contains the elements sharing F as a common face. The L^2-inner product and norm on $\omega = \cup K \subset \Omega$ are denoted by

$$\big(u,\, v\big)_\omega := \sum_{K\subset\omega} (u, v)_K \quad \text{and} \quad \|v\|_{0,\omega}^2 := (v, v)_\omega,$$

respectively. These notations carry through for vector-valued functions. The "energy" seminorm associated with the problem (2.1) is (with slight abuse of notation, because the local seminorm is denoted as a norm):

$$\|v\|_A^2 := \big(A\nabla u,\, \nabla v\big) \;\text{ and }\; \|v\|_{A,\omega}^2 := \big(A\nabla u,\, \nabla v\big)_\omega. \tag{2.2}$$

Let $\mathcal{F}_K$ be the set of faces of an element $K \in \mathcal{T}$. Denote the set of the interior faces within $\omega_{\boldsymbol{z}}$ as:

$$\mathcal{F}_{\boldsymbol{z}} := \{F \in \mathcal{F} : F \in \mathcal{F}_K \text{ for } K \subset \omega_{\boldsymbol{z}},\; F \cap \partial\omega_{\boldsymbol{z}} = \emptyset\}.$$

Denote the H^1-conforming linear finite element space by

$$\mathcal{S}^1 := \{v \in H^1(\Omega) :\, v\big|_K \in P_1(K),\;\; \forall\, K \in \mathcal{T}\}, \tag{2.3}$$

and the piecewise constant space with respect to the triangulation $\mathcal{T}$ by

$$\mathcal{S}^0 := \{v \in L^2(\Omega) :\, v\big|_K \in P_0(K),\;\; \forall\, K \in \mathcal{T}\}, \tag{2.4}$$

Then the finite element approximation to (2.1) is

$$\begin{cases} \text{Find } \; u_{\mathcal{T}} \in \mathcal{S}^1 \cap H_g^1(\Omega) \;\text{ such that} \\ \big(A\nabla u_{\mathcal{T}},\, \nabla v\big) = \big(f,\, v\big), \quad \forall\, v \in \mathcal{S}^1 \cap H_0^1(\Omega). \end{cases} \tag{2.5}$$

For the presentation purpose, here it is assumed that both the diffusion coefficient A and the data f are in $\mathcal{S}^0$, and denote $A\big|_K = A_K$, $f\big|_K = f_K$. Additionally, the Dirichlet boundary data g can be represented by the trace of a function in $\mathcal{S}^1$. In this setting, no data oscillation term will be present in the final error estimate bounds.

Let $\phi_{\boldsymbol{z}_i}$ be the Lagrange nodal basis function of $\mathcal{S}^1$ associated with an interior vertex $\boldsymbol{z}_i \in \mathcal{N}$. Using these nodal basis functions, the discrete problem in (2.5) may be written as the following system of linear equations:

$$\mathbf{A}\mathbf{u} = \mathbf{f}, \tag{2.6}$$

where the stiffness matrix $\mathbf{A}$ is $\mathbf{A}[i,j] = a_{ij}$ with $a_{ij} = \big(A\nabla\phi_{\boldsymbol{z}_j},\, \nabla\phi_{\boldsymbol{z}_i}\big)$; the $\mathbf{u}$ is the vector representation of the exact solution $u_{\mathcal{T}}$; and the $\mathbf{f}$ is the vector representation of the right hand side with i-th row $\mathbf{f}[i]$ of $\mathbf{f}$ being $(f, \phi_{\boldsymbol{z}_i})$. For a given initial guess $\mathbf{u}^{(0)}$, an iterative solver for problem (2.6) has the following form

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \mathbf{B}(\mathbf{f} - \mathbf{A}\mathbf{u}^{(k)}), \tag{2.7}$$

where $\mathbf{u}^{(l)}$ is the vector representation of the l-th iterate $u_{\mathcal{T}}^{(l)}$ for $l = 0, 1, \cdots$. Our attention in this paper is restricted to symmetric iterative methods, i.e., the matrix $\mathbf{B}$ in (2.7) is symmetric.

Next we define the norms for vectors and matrices: with the help of the context, the usual 2-norm $\|\cdot\|_2$ for a vector $\mathbf{v} \in \mathbb{R}^n$ and a non-singular symmetric matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ is defined by:

$$\|\mathbf{v}\|_2 := \sqrt{\mathbf{v}\cdot\mathbf{v}} \text{ and } \|\mathbf{M}\|_2 := \sup_{\|\mathbf{v}\|_2 = 1} \|\mathbf{M}\mathbf{v}\|_2 = \rho(\mathbf{M}), \tag{2.8}$$

respectively, where $\rho(\mathbf{M})$ is the spectral radius of $\mathbf{M}$ equaling its largest eigenvalue.

The stiffness matrix $\mathbf{A}$ is symmetric positive definite for the Dirichlet boundary value problem. As a result, $\mathbf{A}^{1/2}$ is non-singular and can be used to induce a norm:

$$\|\mathbf{v}\|_{\mathbf{A}} := \sqrt{\mathbf{A}\mathbf{v}\cdot\mathbf{v}} = \left\|\mathbf{A}^{1/2}\mathbf{v}\right\|_2 \text{ and } \|\mathbf{M}\|_{\mathbf{A}} := \sup_{\|\mathbf{v}\|_{\mathbf{A}} = 1} \|\mathbf{M}\mathbf{v}\|_{\mathbf{A}}. \tag{2.9}$$

By definition it is straightforward to verify that:

$$\|\mathbf{M}\|_{\mathbf{A}} = \sup_{\|\mathbf{A}^{1/2}\mathbf{v}\|_2 = 1} \left\|\mathbf{A}^{1/2}\mathbf{M}\mathbf{v}\right\|_2 = \left\|\mathbf{A}^{1/2}\mathbf{M}\mathbf{A}^{-1/2}\right\|_2. \tag{2.10}$$

For a finite element function v and its vector representation $\mathbf{v}$, the following equivalence between vector norm and Sobolev norm holds as well:

$$\|v\|_A = \|\mathbf{v}\|_{\mathbf{A}}. \tag{2.11}$$

3. Discretization error estimator using an equilibrated flux.

In this section, firstly the duality theory for the error estimation is introduced. Then a locally post-processed flux based on the iterate $\bar{u}_{\mathcal{T}} := u_{\mathcal{T}}^{(k)}$ for a fixed $k \geq 1$ is constructed. Lastly the reliability of the estimator based on this recovered flux is proved in order that a stopping criterion can be designed for the iterative solver.

3.1. Duality theory.

It is known that the variational problem in (2.1) can be rewritten as a functional minimization problem, where the primal functional is:

$$\mathcal{J}(v) := \frac{1}{2}\big(A\nabla v,\, \nabla v\big) - \big(f,\, v\big) \tag{3.1}$$

Then problem (2.1) is equivalent to the following minimization problem:

$$\text{Find } u \in H_g^1(\Omega) \text{ such that } \mathcal{J}(u) = \min_{v\in H_g^1(\Omega)} \mathcal{J}(v). \tag{3.2}$$

The dual functional with respect to (3.1) is:

$$\mathcal{J}^*(\boldsymbol{\tau}) := -\frac{1}{2}\big(A^{-1}\boldsymbol{\tau},\, \boldsymbol{\tau}\big). \tag{3.3}$$

The dual problem is then to maximize $\mathcal{J}^*(\boldsymbol{\tau})$ in the following space:

$$\boldsymbol{\Sigma} := \{\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div};\Omega) : \nabla\cdot\boldsymbol{\tau} = f\}, \tag{3.4}$$

and can be phrased as:

$$\text{Find } \boldsymbol{\sigma} \in \boldsymbol{\Sigma} \text{ such that } \mathcal{J}^*(\boldsymbol{\sigma}) = \max_{\boldsymbol{\tau}\in\boldsymbol{\Sigma}} \mathcal{J}^*(\boldsymbol{\tau}). \tag{3.5}$$

The foundation to use the dual problem in constructing a posteriori error estimator is that the minimum of the primal functional $\mathcal{J}(\cdot)$ coincides with the maximum of the dual functional $\mathcal{J}^*(\boldsymbol{\sigma})$ (see [19] Chapter 3):

$$\mathcal{J}(u) = \mathcal{J}^*(\boldsymbol{\sigma}) \text{ and } \boldsymbol{\sigma} = -A\nabla u. \tag{3.6}$$

Now that (3.6) is satisfied, then a guaranteed upper bound can be obtained as follows: for any $\boldsymbol{\sigma}_{\mathcal{T}} \in \boldsymbol{\Sigma}_{\mathcal{T}} := \boldsymbol{\Sigma} \cap \boldsymbol{\mathcal{RT}}^0$ being a subspace of $\boldsymbol{\Sigma}$, where $\boldsymbol{\mathcal{RT}}^0$ is the lowest order Raviart-Thomas element (e.g., see [9]),

$$\|u - \bar{u}_{\mathcal{T}}\|_A^2 = 2\Big(\mathcal{J}(\bar{u}_{\mathcal{T}}) - \mathcal{J}(u)\Big) = 2\Big(\mathcal{J}(\bar{u}_{\mathcal{T}}) - \mathcal{J}^*(\boldsymbol{\sigma})\Big) \le 2\Big(\mathcal{J}(\bar{u}_{\mathcal{T}}) - \mathcal{J}^*(\boldsymbol{\sigma}_{\mathcal{T}})\Big). \tag{3.7}$$

One of the main goals of this paper is to locally construct such $\boldsymbol{\sigma}_{\mathcal{T}}$ based on the current iterate $\bar{u}_{\mathcal{T}}$, so that the global reliability bound in (3.7) is automatically met.

3.2. Localized flux recovery. Let $\boldsymbol{\sigma}^\Delta$ be the correction from the numerical flux $\overline{\boldsymbol{\sigma}}_{\mathcal{T}} := -A\nabla\bar{u}_{\mathcal{T}}$ to the true flux $\boldsymbol{\sigma} := -A\nabla u$:

$$\boldsymbol{\sigma}^\Delta := \boldsymbol{\sigma} - \overline{\boldsymbol{\sigma}}_{\mathcal{T}} \tag{3.8}$$

Decompose $\boldsymbol{\sigma}^\Delta$ by a partition of unity $\{\phi_{\boldsymbol{z}}\}_{\boldsymbol{z}\in\mathcal{N}}$, which is the set of the nodal basis functions for the linear finite element space $\mathcal{S}^1$, as follows:

$$\boldsymbol{\sigma}^\Delta = \sum_{\boldsymbol{z}\in\mathcal{N}} \boldsymbol{\sigma}_{\boldsymbol{z}}^\Delta \text{ with } \boldsymbol{\sigma}_{\boldsymbol{z}}^\Delta := \phi_{\boldsymbol{z}}\boldsymbol{\sigma}^\Delta. \tag{3.9}$$

Denote the element residual on an element K and the jump of the normal component of the numerical flux on a face F by

$$r_K := \big\{f + \nabla\cdot(A\nabla\bar{u}_{\mathcal{T}})\big\}\big|_K = f_K \tag{3.10}$$

$$\text{and} \quad j_F := -[\![A\nabla(u-\bar{u}_{\mathcal{T}})\cdot\boldsymbol{n}_F]\!]_F = \begin{cases} [\![A\nabla\bar{u}_{\mathcal{T}}\cdot\boldsymbol{n}_F]\!]_F, & \text{if } F\in\mathcal{F}_{\boldsymbol{z}},\\ A\nabla(u-\bar{u}_{\mathcal{T}})\cdot\boldsymbol{n}_F, & \text{if } F\subset\partial\Omega, \end{cases} \tag{3.11}$$

respectively. Note that r_K and j_F are constants in K and on F if F is an interior face, respectively. When $\boldsymbol{z}\notin\partial\Omega$ is an interior vertex, $\boldsymbol{\sigma}_{\boldsymbol{z}}^\Delta$ satisfies the following local problem:

$$\begin{cases} \nabla\cdot\boldsymbol{\sigma}_{\boldsymbol{z}}^\Delta = \phi_{\boldsymbol{z}} r_K - \nabla\phi_{\boldsymbol{z}}\cdot\nabla(u-\bar{u}_{\mathcal{T}}), & \text{on } K\subset\omega_{\boldsymbol{z}},\\ [\![\boldsymbol{\sigma}_{\boldsymbol{z}}^\Delta\cdot\boldsymbol{n}_F]\!]_F = \phi_{\boldsymbol{z}} j_F, & \text{on } F\in\mathcal{F}_{\boldsymbol{z}},\\ \boldsymbol{\sigma}_{\boldsymbol{z}}^\Delta\cdot\boldsymbol{n}_F = 0, & \text{on } F\subset\partial\omega_{\boldsymbol{z}}. \end{cases} \tag{3.12}$$

If $\boldsymbol{z}\in\partial\Omega$, then the first equation in (3.12) is unchanged, and the flux jump equations change to

$$\begin{cases} [\![\boldsymbol{\sigma}_{\boldsymbol{z}}^\Delta\cdot\boldsymbol{n}_F]\!]_F = \phi_{\boldsymbol{z}} j_F, & \text{on } F\in\mathcal{F}_{\boldsymbol{z}} \text{ and } F\not\subset\partial\omega_{\boldsymbol{z}}\cap\partial\Omega,\\ \boldsymbol{\sigma}_{\boldsymbol{z}}^\Delta\cdot\boldsymbol{n}_F = 0, & \text{on } F\subset\partial\omega_{\boldsymbol{z}}\backslash\partial\Omega. \end{cases} \tag{3.13}$$

To approximate problem (3.12), an approximated correction flux $\boldsymbol{\sigma}^{\Delta}_{\boldsymbol{z},\mathcal{T}}$ is sought in the following broken lowest-order Raviart-Thomas space:

$$\mathcal{RT}^0_{-1,\omega_{\boldsymbol{z}}} := \Big\{\boldsymbol{\tau} \in \boldsymbol{L}^2(\omega_{\boldsymbol{z}}) : \boldsymbol{\tau}\big|_K \in \mathcal{RT}^0(K),\ \forall K \subset \omega_{\boldsymbol{z}}\Big\}, \tag{3.14}$$

where $\mathcal{RT}^0(K)$ denotes the local lowest-order Raviart-Thomas space on K (see [9]).

An explicit procedure called the hypercircle method or equilibration (see [7, 8]) is used to construct $\boldsymbol{\sigma}^{\Delta}_{\boldsymbol{z},\mathcal{T}}$. The correction flux $\boldsymbol{\sigma}^{\Delta}_{\boldsymbol{z},\mathcal{T}}$ satisfies the following problem on an interior vertex patch $\omega_{\boldsymbol{z}}$ ($\boldsymbol{z} \notin \partial\Omega$):

$$\begin{cases} \nabla\cdot\boldsymbol{\sigma}^{\Delta}_{\boldsymbol{z},\mathcal{T}} = \bar{r}_{K,\boldsymbol{z}} + c_{\boldsymbol{z}}, & \text{on } K \subset \omega_{\boldsymbol{z}}, \\ [\![\boldsymbol{\sigma}^{\Delta}_{\boldsymbol{z},\mathcal{T}} \cdot \boldsymbol{n}_F]\!]_F = \bar{j}_{F,\boldsymbol{z}}, & \text{on } F \in \mathcal{F}_{\boldsymbol{z}}, \\ \boldsymbol{\sigma}^{\Delta}_{\boldsymbol{z},\mathcal{T}} \cdot \boldsymbol{n}_F = 0, & \text{on } F \subset \partial\omega_{\boldsymbol{z}}, \end{cases} \tag{3.15}$$

where $\bar{r}_{K,\boldsymbol{z}}$ and $\bar{j}_{F,\boldsymbol{z}}$ are defined as the L^2-projection of $\phi_{\boldsymbol{z}} r_K$ and $\phi_{\boldsymbol{z}} j_F$ onto the constant space of K and interior F, respectively, for $d = 2, 3$:

$$\begin{aligned} \bar{r}_{K,\boldsymbol{z}} &:= \Pi_K(\phi_{\boldsymbol{z}} r_K) = \frac{1}{d+1} f_K = \frac{1}{d+1} r_K, \\ \bar{j}_{F,\boldsymbol{z}} &:= \Pi_F(\phi_{\boldsymbol{z}} j_F) = \frac{1}{d}[\![(A\nabla\bar{u}_{\mathcal{T}})\cdot\boldsymbol{n}_F]\!]_F = \frac{1}{d} j_F. \end{aligned} \tag{3.16}$$

When $\boldsymbol{z} \in \partial\Omega$, $c_{\boldsymbol{z}} = 0$, and the normal fluxes in (3.15) are modified accordingly by (3.13).

Note that, without $c_{\boldsymbol{z}}$, the compatibility condition for (3.15) is not automatically satisfied, that is,

$$\sum_{K\subset\omega_{\boldsymbol{z}}} \big(\bar{r}_{K,\boldsymbol{z}},\, 1\big)_K - \sum_{F\in\mathcal{F}_{\boldsymbol{z}}} \big(\bar{j}_{F,\boldsymbol{z}},\, 1\big)_F \neq 0,$$

which implies that (3.15) does not have a solution. To guarantee the existence of a solution to (3.15), an element-wise compensation term $c_{\boldsymbol{z}}$ is added on the right hand side of the divergence equation in (3.15). Notice that the normal fluxes are kept unchanged so that the final recovered flux can still fulfill the $\boldsymbol{H}(\text{div})$-continuity condition of the space in (3.4). The $c_{\boldsymbol{z}}$ is defined as a constant on this vertex patch $\omega_{\boldsymbol{z}}$ enforcing the compatibility condition for (3.15):

$$\sum_{K\subset\omega_{\boldsymbol{z}}} \big(\bar{r}_{K,\boldsymbol{z}} + c_{\boldsymbol{z}},\, 1\big)_K - \sum_{F\in\mathcal{F}_{\boldsymbol{z}}} \big(\bar{j}_{F,\boldsymbol{z}},\, 1\big)_F = 0, \tag{3.17}$$

which, together with (3.16), yields for an interior vertex $\boldsymbol{z}$

$$\begin{aligned} c_{\boldsymbol{z}} :=& \frac{1}{|\omega_{\boldsymbol{z}}|}\left(\sum_{F\in\mathcal{F}_{\boldsymbol{z}}} \big(\bar{j}_{F,\boldsymbol{z}},\, 1\big)_F - \sum_{K\subset\omega_{\boldsymbol{z}}} \big(\bar{r}_{K,\boldsymbol{z}},\, 1\big)_K\right) \\ =& \frac{1}{|\omega_{\boldsymbol{z}}|}\left(\sum_{F\in\mathcal{F}_{\boldsymbol{z}}} \big(j_F,\, \phi_{\boldsymbol{z}}\big)_F - \sum_{K\subset\omega_{\boldsymbol{z}}} \big(r_K,\, \phi_{\boldsymbol{z}}\big)_K\right) \\ =& \frac{1}{|\omega_{\boldsymbol{z}}|}\big(A\nabla(u - \bar{u}_{\mathcal{T}}),\, \nabla\phi_{\boldsymbol{z}}\big)_{\omega_{\boldsymbol{z}}}. \end{aligned} \tag{3.18}$$

With $c_{\boldsymbol{z}}$, the solution to (3.15) exists since the compatibility condition (3.17) is met (see [8, 13]). We note that if $\bar{u}_{\mathcal{T}}$ solves (2.5) exactly, i.e., $\bar{u}_{\mathcal{T}} = u_{\mathcal{T}}$, then $c_{\boldsymbol{z}} = 0$ for an interior vertex by (3.18), and this is a consequence of the Galerkin orthogonality.

In the case that $\bar{u}_{\mathcal{T}}$ is not an exact solution to problem (2.5), we emphasize again that problem (3.15) is not solvable without the presence of $c_{\boldsymbol{z}}$. The Galerkin orthogonality, which occurs as the compatibility condition for (3.15) if $\boldsymbol{z} \notin \partial\Omega$, is violated if $\bar{u}_{\mathcal{T}}$ is not the exact finite element approximation.

We also note that if $\boldsymbol{z} \in \partial\Omega$, the Galerkin orthogonality does not hold either, $\big(A\nabla u_{\mathcal{T}},\, \nabla\phi_{\boldsymbol{z}}\big) \neq \big(f,\, \phi_{\boldsymbol{z}}\big) = \big(A\nabla u,\, \nabla\phi_{\boldsymbol{z}}\big)$, since the nodal basis $\phi_{\boldsymbol{z}}$ is not in the test function space for the discretized problem in (2.5). A direct usage of (3.18) implies $c_{\boldsymbol{z}} \neq 0$, yet, the degrees of freedom for $\boldsymbol{\sigma}^{\Delta}_{\boldsymbol{z},\mathcal{T}}$ on the faces on $\partial\omega_{\boldsymbol{z}} \cap \partial\Omega$ are treated as unknowns in (3.19), and $c_{\boldsymbol{z}}$ is not needed in (3.15) on a boundary vertex $\boldsymbol{z} \in \partial\Omega$.

The flux correction is postprocessed by a minimization procedure locally on $\omega_{\boldsymbol{z}}$:

$$\left\| A^{-1/2}\boldsymbol{\sigma}^{\Delta}_{\boldsymbol{z},\mathcal{T}} \right\|_{0,\omega_{\boldsymbol{z}}} = \min_{\boldsymbol{\tau}\in\boldsymbol{\Sigma}_{\boldsymbol{z},\mathcal{T}}} \left\| A^{-1/2}\boldsymbol{\tau} \right\|_{0,\omega_{\boldsymbol{z}}}, \tag{3.19}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{z},\mathcal{T}} := \big\{ \boldsymbol{\tau} \in \boldsymbol{\mathcal{RT}}^{0}_{-1,\omega_{\boldsymbol{z}}} : \boldsymbol{\tau} \text{ satisfies (3.15)} \big\}$. The element-wise and the global flux corrections are then:

$$\boldsymbol{\sigma}^{\Delta}_{K,\mathcal{T}} := \sum_{\boldsymbol{z}\in\mathcal{N}_K} \boldsymbol{\sigma}^{\Delta}_{\boldsymbol{z},\mathcal{T}} \quad \text{and} \quad \boldsymbol{\sigma}^{\Delta}_{\mathcal{T}} := \sum_{\boldsymbol{z}\in\mathcal{N}} \boldsymbol{\sigma}^{\Delta}_{\boldsymbol{z},\mathcal{T}}. \tag{3.20}$$

Lastly, a compensatory flux $\boldsymbol{\sigma}^{c}_{\mathcal{T}}$, which is in the globally $\boldsymbol{H}(\mathrm{div})$-conforming $\boldsymbol{\mathcal{RT}}^{0}$ space, is then sought using $c_{\boldsymbol{z}}$ defined in (3.18) as data:

$$\nabla\cdot\boldsymbol{\sigma}^{c}_{\mathcal{T}} = -\sum_{\boldsymbol{z}\in\mathcal{N}_K} c_{\boldsymbol{z}}, \quad \text{in any } K \in \mathcal{T}, \tag{3.21}$$

By the surjectivity of the divergence operator from $\boldsymbol{\mathcal{RT}}^{0}$ to $\mathcal{S}^{0}$, the above problem has a solution (e.g., [9, ?]). If $\boldsymbol{\sigma}^{c}_{\mathcal{T}}$ is sought by minimizing a weighted L^2-norm, with (3.21) being a constraint, then it is equivalent to seeking the solution to a mixed finite element approximation problem in the $\boldsymbol{\mathcal{RT}}^{0}$–$\mathcal{S}^{0}$ pair. The energy estimate in a weighted L^2-norm for $\boldsymbol{\sigma}^{c}_{\mathcal{T}}$, which bridges it with the algebraic error, will be shown later in Lemma 3.

The recovered flux based on the $\bar{u}_{\mathcal{T}}$ is defined as:

$$\boldsymbol{\sigma}_{\mathcal{T}} := -A\nabla\bar{u}_{\mathcal{T}} + \boldsymbol{\sigma}^{\Delta}_{\mathcal{T}} + \boldsymbol{\sigma}^{c}_{\mathcal{T}}. \tag{3.22}$$

In practice, only $\boldsymbol{\sigma}^{\Delta}_{\mathcal{T}}$ is explicitly computed. For explicit local constructions of $\boldsymbol{\sigma}^{\Delta}_{\mathcal{T}}$, we refer the readers to [13, 8]. The $\boldsymbol{\sigma}^{c}_{\mathcal{T}}$ is here to compensate the change in divergence caused by the correction term $c_{\boldsymbol{z}}$, and is not needed, nor explicitly computed for the estimator defined in (3.23).

LEMMA 1. *The recovered flux $\boldsymbol{\sigma}_{\mathcal{T}}$ is in the conforming finite element subspace of the duality space: $\boldsymbol{\sigma}_{\mathcal{T}} \in \boldsymbol{\Sigma}_{\mathcal{T}} := \boldsymbol{\Sigma} \cap \boldsymbol{\mathcal{RT}}^{0}$.*

Proof. Using (3.15) and (3.21), together with the fact that $A\nabla\bar{u}_{\mathcal{T}}$ is a constant vector on each element K, we have:

$$\nabla\cdot\boldsymbol{\sigma}_{\mathcal{T}}\big|_K = \nabla\cdot\boldsymbol{\sigma}^{\Delta}_{\mathcal{T}} + \nabla\cdot\boldsymbol{\sigma}^{c}_{\mathcal{T}} = \sum_{\boldsymbol{z}\in\mathcal{N}_K} \bar{r}_{K,\boldsymbol{z}} = f_K.$$

On $F \in \mathcal{F}$, the continuity of the normal component implies $\boldsymbol{\sigma}^{c}_{\mathcal{T}} \in \boldsymbol{H}(\mathrm{div};\Omega)$

$$[\![\boldsymbol{\sigma}_{\mathcal{T}}\cdot\boldsymbol{n}]\!]_F = [\![\boldsymbol{\sigma}^{\Delta}_{\mathcal{T}}\cdot\boldsymbol{n}]\!]_F - [\![A\nabla\bar{u}_{\mathcal{T}}\cdot\boldsymbol{n}]\!]_F = \sum_{\boldsymbol{z}\in\mathcal{N}(F)} \bar{j}_{F,\boldsymbol{z}} - j_F = 0.$$

∎

3.3. Discretization error estimator and reliability. With the recovered flux correction defined in (3.20), we define the discretization error estimator η_d as:

$$\eta_{d,\,K} = \left\| A^{-1/2}\boldsymbol{\sigma}^{\Delta}_{K,\mathcal{T}} \right\|_{0,K}, \quad \text{and} \quad \eta_d = \left\| A^{-1/2}\boldsymbol{\sigma}^{\Delta}_{\mathcal{T}} \right\|_{0}. \tag{3.23}$$

The reliability we show in this section is: the total error $\|u-\bar{u}_{\mathcal{T}}\|_A$ is bounded by the error estimator η_d plus the algebraic error.

In (3.18), the representation of $c_{\boldsymbol{z}}$ uses $u-\bar{u}_{\mathcal{T}}$. Nevertheless, inserting the Galerkin orthogonality into (3.18), which reads $\left(A\nabla(u-u_{\mathcal{T}}),\, \nabla\phi_{\boldsymbol{z}}\right)_{\omega_{\boldsymbol{z}}} = 0$ for any interior vertex $\boldsymbol{z}$, we have

$$c_{\boldsymbol{z}} = \frac{1}{|\omega_{\boldsymbol{z}}|}\left(A\nabla(u_{\mathcal{T}}-\bar{u}_{\mathcal{T}}),\, \nabla\phi_{\boldsymbol{z}}\right)_{\omega_{\boldsymbol{z}}} = \frac{1}{|\omega_{\boldsymbol{z}}|}\sum_{K\subset\omega_{\boldsymbol{z}}}\left(A\nabla(u_{\mathcal{T}}-\bar{u}_{\mathcal{T}}),\, \nabla\phi_{\boldsymbol{z}}\right)_{K}. \tag{3.24}$$

Now the compatibility compensation term $c_{\boldsymbol{z}}$ can be decomposed as follows:

$$c_{\boldsymbol{z}} = \sum_{K\subset\omega_{\boldsymbol{z}}} c_{\boldsymbol{z},K}, \quad \text{with} \quad c_{\boldsymbol{z},K} := \frac{1}{|\omega_{\boldsymbol{z}}|}\left(A\nabla(u_{\mathcal{T}}-\bar{u}_{\mathcal{T}}),\, \nabla\phi_{\boldsymbol{z}}\right)_{K}. \tag{3.25}$$

LEMMA 2 (Nodal estimate for the compensation term). *For any interior vertex $\boldsymbol{z}\in\mathcal{N}_K$, on $K\subset\omega_{\boldsymbol{z}}$, $c_{\boldsymbol{z},K}$ satisfies the following L^2-estimate with C depending on the shape regularity of the patch $\omega_{\boldsymbol{z}}$:*

$$h_K A_K^{-1/2}\left\|c_{\boldsymbol{z},K}\right\|_{0,K} \le C\,\left\|u_{\mathcal{T}}-\bar{u}_{\mathcal{T}}\right\|_{A,K}, \tag{3.26}$$

Proof. By the representation in (3.25), it follows from the Cauchy-Schwarz inequality, the fact that $\|\nabla\phi_{\boldsymbol{z}}\|_{0,K} \le C\,h_K^{\frac{d}{2}-1}$, and the shape regularity of the patch that

$$\begin{aligned} |c_{\boldsymbol{z},K}| &= \frac{1}{|\omega_{\boldsymbol{z}}|}\left|\left(A\nabla(u-\bar{u}_{\mathcal{T}}),\, \nabla\phi_{\boldsymbol{z}}\right)_K\right| \le \frac{1}{|\omega_{\boldsymbol{z}}|}\left\|u-\bar{u}_{\mathcal{T}}\right\|_{A,K}\left\|\phi_{\boldsymbol{z}}\right\|_{A,K} \\ &\le C\,h_K^{-\frac{d}{2}-1}A_K^{1/2}\,\left\|u-\bar{u}_{\mathcal{T}}\right\|_{A,K}. \end{aligned}$$

Since $c_{\boldsymbol{z},K}$ is a constant on K, $\|c_{\boldsymbol{z}}\|_{0,K} \le h_K^{\frac{d}{2}}\,|c_{\boldsymbol{z},K}|$, the validity of (3.26) is then verified. ∎

To bridge the energy estimate for $\boldsymbol{\sigma}^c_{\mathcal{T}}$ with the algebraic error, the following norms are need: let $A_F := \max_{K\subset\omega_F} A_K$, for $p\in\mathcal{S}^0$, and $f\in L^2(\Omega)$

$$\|f\|_{-1,h} := \sup_{q\in\mathcal{S}^0}\frac{(f,q)}{\|q\|_{1,h}}, \quad \text{and} \quad \|p\|_{1,h} := \left(\sum_{F\in\mathcal{F}} h_F^{-1}A_F\left\|[\![p]\!]\right\|^2_{0,F}\right)^{1/2}. \tag{3.27}$$

LEMMA 3 (A discrete energy estimate for $\boldsymbol{\sigma}^c_{\mathcal{T}}$). *If $\boldsymbol{\sigma}^c_{\mathcal{T}}$ is obtained by*

$$\left\|A^{-1/2}\boldsymbol{\sigma}^c_{\mathcal{T}}\right\|_0 = \min_{\substack{\boldsymbol{\tau}\in\boldsymbol{\mathcal{RT}}^0,\\ \nabla\cdot\boldsymbol{\tau}=f^c}}\left\|A^{-1/2}\boldsymbol{\tau}\right\|_0, \tag{3.28}$$

where f^c is defined as follows on an element K using (3.21),

$$f^c|_K := -\sum_{\boldsymbol{z}\in\mathcal{N}_K} c_{\boldsymbol{z}} \tag{3.29}$$

then the following estimate holds:

$$\left\| A^{-1/2}\boldsymbol{\sigma}_{\mathcal{T}}^c \right\|_0 \le C_A \left\| u_{\mathcal{T}} - \bar{u}_{\mathcal{T}} \right\|_A , \tag{3.30}$$

in which C depends on the shape regularity of the triangulation, the maximum number of elements in each ω_K, and the diffusion coefficient A.

Proof. The minimizer of problem (3.28) satisfies the following global mixed problem: find $(\boldsymbol{\sigma}_{\mathcal{T}}^c, p) \in \boldsymbol{\mathcal{RT}}^0 \times \mathcal{S}^0$

$$\begin{cases} \big(A^{-1}\boldsymbol{\sigma}_{\mathcal{T}}^c,\, \boldsymbol{\tau}\big) - \big(p,\, \nabla\cdot\boldsymbol{\tau}\big) = 0, & \forall\, \boldsymbol{\tau} \in \boldsymbol{\mathcal{RT}}^0, \\ \big(\nabla\cdot\boldsymbol{\sigma}_{\mathcal{T}}^c,\, q\big) = \big(f^c,\, q\big), & \forall\, q \in \mathcal{S}^0. \end{cases} \tag{3.31}$$

By the inf-sup stability of discrete H^1-$\boldsymbol{L}^2$ analysis of the mixed problem when the shape regularity of the mesh is assumed ($h_F \eqsim h_K$ for F's neighboring elements) ([7, Chapter 3 §5.7]), problem (3.31) has a unique solution satisfying the following energy estimate: letting $\boldsymbol{\tau} = \boldsymbol{\sigma}_{\mathcal{T}}^c$, $q = p$, we have

$$\begin{aligned} \left\| A^{-1/2}\boldsymbol{\sigma}_{\mathcal{T}}^c \right\|_0^2 &\le \|f^c\|_{-1,h} \|p\|_{1,h} \le \|f^c\|_{-1,h} \sup_{\boldsymbol{\tau}\in\boldsymbol{\mathcal{RT}}^0} \frac{(p, \nabla\cdot\boldsymbol{\tau})}{\left\|A^{-1/2}\boldsymbol{\tau}\right\|_0} \\ &= \|f^c\|_{-1,h} \sup_{\boldsymbol{\tau}\in\boldsymbol{\mathcal{RT}}^0} \frac{\big(A^{-1}\boldsymbol{\sigma}_{\mathcal{T}}^c,\, \boldsymbol{\tau}\big)}{\left\|A^{-1/2}\boldsymbol{\tau}\right\|_0} \le \|f^c\|_{-1,h} \left\| A^{-1/2}\boldsymbol{\sigma}_{\mathcal{T}}^c \right\|_0 . \end{aligned} \tag{3.32}$$

Now, to prove the validity of the lemma, by (3.27), it suffices to show that for $q \in \mathcal{S}^0$

$$(f^c, q) \le C \left\| u_{\mathcal{T}} - \bar{u}_{\mathcal{T}} \right\|_A \|q\|_{1,h} . \tag{3.33}$$

To this end, first denote $q_K := q|_K$, and f_c is written out explicitly using (3.29),

$$(f^c, q) = -\sum_{K\in\mathcal{T}} \left(\sum_{\boldsymbol{z}\in\mathcal{N}_K} c_{\boldsymbol{z}}, q \right)_K = -\sum_{K\in\mathcal{T}} \sum_{\boldsymbol{z}\in\mathcal{N}_K} c_{\boldsymbol{z}} q_K |K|. \tag{3.34}$$

Using $c_{\boldsymbol{z}} = \sum_{K\subset\omega_{\boldsymbol{z}}} c_{\boldsymbol{z},K}$ in (3.25) for interior vertices and $c_{\boldsymbol{z}} = 0$ for $\boldsymbol{z} \in \partial\Omega$ yields,

$$(f^c, q) = -\sum_{K\in\mathcal{T}} \sum_{\boldsymbol{z}\in\mathcal{N}_K, \boldsymbol{z}\notin\partial\Omega} \left(\sum_{T\subset\omega_{\boldsymbol{z}}} c_{\boldsymbol{z},T} \right) q_K |K|. \tag{3.35}$$

We switch the order of the summation, by summing up the inner terms $c_{\boldsymbol{z},T}$ last, then the above equation becomes

$$\begin{aligned} &- \sum_{K\in\mathcal{T}} \sum_{\boldsymbol{z}\in\mathcal{N}_K, \boldsymbol{z}\notin\partial\Omega} \left(\sum_{T\subset\omega_{\boldsymbol{z}}} c_{\boldsymbol{z},T} \right) q_K |K| \\ &= -\sum_{K\in\mathcal{T}} \sum_{\boldsymbol{z}\in\mathcal{N}_K, \boldsymbol{z}\notin\partial\Omega} \left\{ c_{\boldsymbol{z},K} \Big(\sum_{T\subset\omega_{\boldsymbol{z}}} q_T |T| \Big) \right\} =: -(*), \end{aligned} \tag{3.36}$$

in which for each vertex $\boldsymbol{z} \in \mathcal{N}_K$, the term $c_{\boldsymbol{z},K}$ is only summed against $q_T|T|$ for $T \subset \omega_{\boldsymbol{z}}$. The reason is that among the terms in the original summation in (3.35), a term involving $c_{\boldsymbol{z},T}$ is summed up multiplying $q_K|K|$ only when $\omega_{\boldsymbol{z}} \subset \omega_K$.

Now on each K not touching $\partial\Omega$, we have the following weighted average of $c_{\boldsymbol{z},K}$, using $|\omega_{\boldsymbol{z}}|m_K$ as weights, being zero for any m_K that is a constant on the patch ω_K:

$$\sum_{\boldsymbol{z}\in\mathcal{N}_K,\boldsymbol{z}\notin\partial\Omega} c_{\boldsymbol{z},K}(|\omega_{\boldsymbol{z}}|m_K) = m_K \sum_{\boldsymbol{z}\in\mathcal{N}_K,\boldsymbol{z}\notin\partial\Omega} \big(A\nabla(u_{\mathcal{T}} - \bar{u}_{\mathcal{T}}),\, \nabla\phi_{\boldsymbol{z}}\big)_K = 0. \tag{3.37}$$

As a result, $|\omega_{\boldsymbol{z}}|m_K$ can be inserted into (3.36), and m_K is chosen as the average of q on ω_K, i.e., $m_K := (\sum_{P\subset\omega_K} q_P|P|)/|\omega_K|$, thus $(*)$ in (3.36) becomes

$$\begin{aligned}
&\sum_{K\in\mathcal{T}} \sum_{\boldsymbol{z}\in\mathcal{N}_K,\boldsymbol{z}\notin\partial\Omega} \left\{ c_{\boldsymbol{z},K}\Big(\sum_{T\subset\omega_{\boldsymbol{z}}} q_T|T| - \frac{|\omega_{\boldsymbol{z}}|}{|\omega_K|} \sum_{P\subset\omega_K} q_P|P| \Big) \right\} \\
&= \sum_{K\in\mathcal{T}} \sum_{\boldsymbol{z}\in\mathcal{N}_K,\boldsymbol{z}\notin\partial\Omega} \left\{ c_{\boldsymbol{z},K} \sum_{T\subset\omega_{\boldsymbol{z}}} \Big(\frac{|T|}{|\omega_K|} \sum_{P\subset\omega_K} (q_T - q_P)|P| \Big) \right\} =: \sum_{K\in\mathcal{T}} \sum_{\boldsymbol{z}\in\mathcal{N}_K,\boldsymbol{z}\notin\partial\Omega} \beta_{\boldsymbol{z},K}.
\end{aligned} \tag{3.38}$$

For any $T \subset \omega_{\boldsymbol{z}}$, if T and $P \subset \omega_K$ have a common face $F = \partial T \cap \partial P$, $|q_T - q_P| = |[\![q]\!]_F|$ on F; otherwise, there always exists a path consisting of finite many elements $K_i \subset \omega_K$ $(i = 1, \ldots, n_{TP})$ starting from $K_1 := T$ to $K_{n_{TP}} := P$, such that K_i and K_{i-1} share a face F_i, then

$$|q_T - q_P| = |q_{K_1} - q_{K_2} + q_{K_2} - q_{K_3} \ldots| \le \sum_{i=1}^{n_{TP}} \left|[\![q]\!]_{F_i}\right| \le \sum_{F\in\mathcal{F}_{\omega_K}} \left|[\![q]\!]_F\right|. \tag{3.39}$$

Applying above on the innermost summation for P of (3.38), exploiting the local shape regularity on every element in ω_K, and using the fact that $c_{\boldsymbol{z},K}$ and $[\![q]\!]_F$ are constants on K and F, respectively, yields:

$$\begin{aligned}
\beta_{\boldsymbol{z},K} &\le |c_{\boldsymbol{z},K}| \left| \sum_{T\subset\omega_{\boldsymbol{z}}} \Big(\frac{|T|}{|\omega_K|} \sum_{P\subset\omega_K} (q_T - q_P)|P| \Big) \right| \le |c_{\boldsymbol{z},K}| \left(\sum_{T\subset\omega_{\boldsymbol{z}}} |T| \sum_{F\in\mathcal{F}_{\omega_K}} \left|[\![q]\!]_F\right| \right) \\
&\le C A_K^{-1/2} h_K \left\|c_{\boldsymbol{z},K}\right\|_{0,K} \cdot A_K^{1/2} h_K^{-1} |K|^{1/2} \left(\sum_{F\in\mathcal{F}_{\omega_K}} \left\|[\![q]\!]_F\right\|_{0,F} |F|^{-1/2} \right).
\end{aligned} \tag{3.40}$$

Using the Cauchy-Schwarz inequality and the shape regularity of the triangulation, $(*)$ can be estimated as follows:

$$\begin{aligned}
(*) \le C &\left(\sum_{K\in\mathcal{T}} \sum_{\boldsymbol{z}\in\mathcal{N}_K,\boldsymbol{z}\notin\partial\Omega} A_K^{-1} h_K^2 \left\|c_{\boldsymbol{z},K}\right\|_{0,K}^2 \right)^{1/2} \\
&\left(\sum_{K\in\mathcal{T}} \sum_{\boldsymbol{z}\in\mathcal{N}_K,\boldsymbol{z}\notin\partial\Omega} A_K \sum_{F\in\mathcal{F}_{\omega_K}} h_F^{-1} \left\|[\![q]\!]_F\right\|_{0,F}^2 \right)^{1/2}.
\end{aligned} \tag{3.41}$$

Finally, the lemma follows from Lemma 2 and definition (3.27). ∎

THEOREM 4. *There exists a positive constant C_A, depending on the shape regularity of the mesh and the coefficient A, such that*

$$\left\|u - \bar{u}_{\mathcal{T}}\right\|_A \le \eta_d + C_A \left\|u_{\mathcal{T}} - \bar{u}_{\mathcal{T}}\right\|_A. \tag{3.42}$$

Proof. The proof of (3.42) starts from (3.7)

$$\|u - \bar{u}_{\mathcal{T}}\|_A^2 \le 2\Big(\mathcal{J}(\bar{u}_{\mathcal{T}}) - \mathcal{J}^*(\boldsymbol{\sigma}_{\mathcal{T}})\Big) = \left\|A^{1/2}\nabla \bar{u}_{\mathcal{T}}\right\|_0^2 - 2\big(f,\, \bar{u}_{\mathcal{T}}\big) + \big(A^{-1}\boldsymbol{\sigma}_{\mathcal{T}},\, \boldsymbol{\sigma}_{\mathcal{T}}\big).$$

With $\boldsymbol{\sigma}_{\mathcal{T}} = -A\nabla\bar{u}_{\mathcal{T}} + \boldsymbol{\sigma}_{\mathcal{T}}^{\Delta} + \boldsymbol{\sigma}_{\mathcal{T}}^{c}$ defined in (3.22), we have

$$\big(A^{-1}\boldsymbol{\sigma}_{\mathcal{T}},\, \boldsymbol{\sigma}_{\mathcal{T}}\big) = \left\|A^{-1/2}(\boldsymbol{\sigma}_{\mathcal{T}}^{\Delta} + \boldsymbol{\sigma}_{\mathcal{T}}^{c})\right\|_0^2 - 2\big(\boldsymbol{\sigma}_{\mathcal{T}},\, \nabla\bar{u}_{\mathcal{T}}\big) - \left\|A^{1/2}\nabla\bar{u}_{\mathcal{T}}\right\|_0^2,$$

which, together with the above inequality, implies

$$\begin{aligned}\|u - \bar{u}_{\mathcal{T}}\|_A^2 &\le \left\|A^{-1/2}(\boldsymbol{\sigma}_{\mathcal{T}}^{\Delta} + \boldsymbol{\sigma}_{\mathcal{T}}^{c})\right\|_0^2 - 2\big(\boldsymbol{\sigma}_{\mathcal{T}},\, \nabla\bar{u}_{\mathcal{T}}\big) - 2\big(f,\, \bar{u}_{\mathcal{T}}\big)\\ &= \left\|A^{-1/2}(\boldsymbol{\sigma}_{\mathcal{T}}^{\Delta} + \boldsymbol{\sigma}_{\mathcal{T}}^{c})\right\|_0^2.\end{aligned}$$

The last equality uses the fact that $\big(\boldsymbol{\sigma}_{\mathcal{T}},\, \nabla\bar{u}_{\mathcal{T}}\big) + \big(f,\, \bar{u}_{\mathcal{T}}\big) = 0$, which follows from integration by parts element-wise and Lemma 1. By the triangle inequality, we have

$$\|u - \bar{u}_{\mathcal{T}}\|_A \le \left\|A^{-1/2}\boldsymbol{\sigma}_{\mathcal{T}}^{\Delta}\right\|_0 + \left\|A^{-1/2}\boldsymbol{\sigma}_{\mathcal{T}}^{c}\right\|_0 = \eta_d + \left\|A^{-1/2}\boldsymbol{\sigma}_{\mathcal{T}}^{c}\right\|_0.$$

Now, the theorem simply follows from estimate (3.30) in Lemma 3. ∎

4. Algebraic error estimator.

The upper bound in (3.42) contains the algebraic error $\|u_{\mathcal{T}} - \bar{u}_{\mathcal{T}}\|_A$. This section introduces an algebraic error estimator in terms of the energy norm of two consecutive iterates with a constant depending on an approximation of the spectral radius of the error propagation matrix.

Recall the stiffness matrix $\mathbf{A}$ introduced in Section 2 and the iteration in (2.7). Denote the algebraic iteration error at the k-th iteration by

$$\mathbf{e}^{(k)} := \mathbf{u}_{\mathcal{T}} - \mathbf{u}_{\mathcal{T}}^{(k)}, \tag{4.1}$$

then the error propagation can be verified to be:

$$\mathbf{e}^{(k+1)} = (\mathbf{I} - \mathbf{BA})\mathbf{e}^{(k)}. \tag{4.2}$$

Let $e^{(k)}$ be the function in the finite element space having $\mathbf{e}^{(k)}$ as its vector representation in the nodal basis. Define the spectral radius of the error propagation matrix $\mathbf{I} - \mathbf{BA}$ as ρ_{err}:

$$\rho_{\mathrm{err}} := \rho(\mathbf{I} - \mathbf{BA}) = \|\mathbf{I} - \mathbf{BA}\|_2. \tag{4.3}$$

THEOREM 5 (Upper bound of the algebraic error). *Let $\{\mathbf{u}^{(k)}\}$ be the sequence generated by (2.7), then the algebraic error $\mathbf{e}^{(k)}$ defined in (4.1) satisfies the following estimate:*

$$\left\|\mathbf{e}^{(k+1)}\right\|_{\mathbf{A}} \le \frac{\rho_{\mathrm{err}}}{1 - \rho_{\mathrm{err}}}\left\|\mathbf{u}^{(k+1)} - \mathbf{u}^{(k)}\right\|_{\mathbf{A}}, \tag{4.4}$$

or in the finite element function form:

$$\left\|u_{\mathcal{T}} - u_{\mathcal{T}}^{(k+1)}\right\|_A \le \frac{\rho_{\mathrm{err}}}{1 - \rho_{\mathrm{err}}}\left\|u_{\mathcal{T}}^{(k+1)} - u_{\mathcal{T}}^{(k)}\right\|_A. \tag{4.5}$$

Proof. By the norm equivalence in (2.10) and the fact that $(\mathbf{I}-\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{-1/2})$ is similar to $(\mathbf{I}-\mathbf{BA})$ (they have the same eigenvalues), we have

$$\|\mathbf{I}-\mathbf{BA}\|_{\mathbf{A}} = \left\|\mathbf{A}^{1/2}(\mathbf{I}-\mathbf{BA})\mathbf{A}^{-1/2}\right\|_2 = \rho\big(\mathbf{I}-\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{-1/2}\big) = \rho_{\mathrm{err}}. \tag{4.6}$$

Hence, $\left\|\mathbf{e}^{(k+1)}\right\|_{\mathbf{A}} \le \rho_{\mathrm{err}} \left\|\mathbf{e}^{(k)}\right\|_{\mathbf{A}}$, and the result follows from a standard contraction mapping convergence theorem (see, e.g., [22, Theorem 12.1.2]). ∎

In Theorem 5, ρ_{err} is the true rate of convergence of the solver. However, in practice, ρ_{err} is not available during any iteration of the solver, unless an eigenvalue problem is solved for the error propagation matrix $\mathbf{I}-\mathbf{BA}$. What we have access to is the following quantity:

$$\rho_{\mathrm{err}}^{(k)} := \frac{\|\mathbf{r}_k\|_2}{\|\mathbf{r}_{k-1}\|_2}, \tag{4.7}$$

where $\mathbf{r}_k := \mathbf{A}\mathbf{e}^{(k)}$ with j-th entry given by $(f,\phi_{\boldsymbol{z}_j}) - \big(A\nabla u_{\mathcal{T}}^{(k)},\, \nabla\phi_{\boldsymbol{z}_j}\big)$. The following lemma describes the convergence of $\rho_{\mathrm{err}}^{(k)}$ provided that the iterative solver is convergent.

Lemma 6 (Convergence of $\rho_{\mathrm{err}}^{(k)}$). *Assuming the error propagation matrix $\mathbf{I}-\mathbf{BA}$ has eigenvalues $1 > \rho_{\mathrm{err}} = \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_N > 0$, then $\rho_{\mathrm{err}}^{(k)} \to \rho_{\mathrm{err}}$ as $k \to \infty$.*

Proof. First notice that, by applying (4.2) from 0 to k in a cascading fashion,

$$\mathbf{r}_k = \mathbf{A}\mathbf{e}^{(k)} = \mathbf{A}(\mathbf{I}-\mathbf{BA})^k\mathbf{e}^{(0)} = (\mathbf{I}-\mathbf{AB})^k\mathbf{A}\mathbf{e}^{(0)} = (\mathbf{I}-\mathbf{AB})^k\mathbf{r}_0.$$

Since $\mathbf{A}^{-1}(\mathbf{I}-\mathbf{AB})\mathbf{A} = \mathbf{I}-\mathbf{BA}$, $\mathbf{I}-\mathbf{AB}$ and $\mathbf{I}-\mathbf{BA}$ share the same eigenvalues and eigenvectors. Suppose that $\{\mathbf{v}_i\}_{i=1}^N$ are the set of orthonormal eigenvectors in the ℓ^2-sense corresponding to the eigenvalue set $\{\lambda_i\}_{i=1}^N$. Let $c_i = \mathbf{r}_0\cdot\mathbf{v}_i$ be the coefficient of the eigen-expansion of $\mathbf{r}_0$. Without loss of generality, assume the multiplicity of the largest eigenvalue λ_1 is 1. Then we have:

$$\begin{aligned} \rho_{\mathrm{err}}^{(k)} &= \frac{\left\|(\mathbf{I}-\mathbf{AB})^k\mathbf{r}_0\right\|_2}{\left\|(\mathbf{I}-\mathbf{AB})^{k-1}\mathbf{r}_0\right\|_2} = \frac{\left\|\sum_{i=1}^N \lambda_i^k c_i\mathbf{v}_i\right\|_2}{\left\|\sum_{i=1}^N \lambda_i^{k-1} c_i\mathbf{v}_i\right\|_2} \\ &= \lambda_1 \frac{\left\|\sum_{i=1}^N \left(\frac{\lambda_i}{\lambda_1}\right)^k c_i\mathbf{v}_i\right\|_2}{\left\|\sum_{i=1}^N \left(\frac{\lambda_i}{\lambda_1}\right)^{k-1} c_i\mathbf{v}_i\right\|_2} = \lambda_1 \frac{1+\sum_{i=2}^N b_i\gamma_i^k}{1+\sum_{i=2}^N b_i\gamma_i^{k-1}}, \end{aligned} \tag{4.8}$$

where $b_i := (c_i/c_1)^2$, and $\gamma_i := (\lambda_i/\lambda_1)^2$. The lemma follows from letting $k\to\infty$. When the multiplicity of λ_1 is $m \ge 2$, factoring out the first m terms and i starts from $(m+1)$ in the eigen-expansion in (4.8) yields the same result. ∎

Lemma 7 (Monotonicity of $\rho_{\mathrm{err}}^{(k)}$). *Under the same assumption as in Lemma 6, $\rho_{\mathrm{err}}^{(k)} \le \rho_{\mathrm{err}}^{(k+1)}$, for any fixed $k \in \mathbb{R}^+$.*

Proof. By (4.8), to prove the validity of the lemma, it suffices to show that:

$$\left(1+\sum_{i=2}^N b_i\gamma_i^k\right)^2 \le \left(1+\sum_{i=2}^N b_i\gamma_i^{k-1}\right)\left(1+\sum_{i=2}^N b_i\gamma_i^{k+1}\right), \tag{4.9}$$

which is equivalent to

$$2\sum_{i=2}^{N} b_i\gamma_i^k + \left(\sum_{i=2}^{N} b_i\gamma_i^k\right)^2 \le \sum_{i=2}^{N} b_i\left(\gamma_i^{k-1}+\gamma_i^{k-1}\right) + \left(\sum_{i=2}^{N} b_i\gamma_i^{k-1}\right)\left(\sum_{i=2}^{N} b_i\gamma_i^{k+1}\right). \tag{4.10}$$

Since $b_i \ge 0$, $\lambda_i \ge 0$, and $2\gamma_i \le 1+\gamma_i^2$, we have

$$2\sum_{i=2}^{N} b_i\gamma_i^k \le \sum_{i=2}^{N} b_i\left(\gamma_i^{k-1}+\gamma_i^{k-1}\right),$$

Then it suffices to show the following inequality:

$$a := \left(\sum_{i=2}^{N} b_i\gamma_i^k\right)^2 - \left(\sum_{i=2}^{N} b_i\gamma_i^{k-1}\right)\left(\sum_{i=2}^{N} b_i\gamma_i^{k+1}\right) \le 0, \tag{4.11}$$

which will be proved by a standard inductive argument. To this end, let $N=2$, it is easy to see that (4.11) holds with equality. Next, assume that (4.11) holds for $N=n$. For $N=n+1$: we have

$$\begin{aligned} a &= \left(\sum_{i=2}^{n} b_i\gamma_i^k + b_{n+1}\gamma_{n+1}^k\right)^2 - \left(\sum_{i=2}^{n} b_i\gamma_i^{k-1} + b_{n+1}\gamma_{n+1}^{k-1}\right)\left(\sum_{i=2}^{n} b_i\gamma_i^{k+1} + b_{n+1}\gamma_{n+1}^{k+1}\right) \\ &\le \left(\sum_{i=2}^{n} b_i\gamma_i^k\right)^2 - \left(\sum_{i=2}^{n} b_i\gamma_i^{k-1}\right)\left(\sum_{i=2}^{n} b_i\gamma_i^{k+1}\right) - b_{n+1}\gamma_{n+1}^{k-1}\sum_{i=2}^{n} b_i(\gamma_{n+1}-\gamma_i)^2\gamma_i^{k-1}. \end{aligned} \tag{4.12}$$

Now (4.11) is a direct consequence of the induction hypothesis. This completes the proof of the lemma. □

After the preparation, now we define the algebraic error estimator as follows at the $(k+1)$-th iteration of the solver: for $k \ge 1$

$$\eta_a^{(k+1)} := e^{1/k}\frac{\rho_{\text{err}}^{(k)}}{1-\rho_{\text{err}}^{(k)}}\left\|\mathbf{u}^{(k+1)}-\mathbf{u}^{(k)}\right\|_{\mathbf{A}} = e^{1/k}\frac{\rho_{\text{err}}^{(k)}}{1-\rho_{\text{err}}^{(k)}}\left\|u_{\mathcal{T}}^{(k+1)}-u_{\mathcal{T}}^{(k)}\right\|_{A}. \tag{4.13}$$

The $e^{1/k}$ factor is added to remedy the fact that $\rho_{\text{err}}^{(k)}$ converges to ρ_{err} from below. Without it, the solver might stop too early, before a good estimate of ρ_{err} is obtained.

Theorem 8 (Reliability of the algebraic error estimator). *Under the same setting with Theorem 5 and Lemma 6, there exists an $N \in \mathbb{R}^+$ such that for all $k \ge N$,*

$$\left\|\mathbf{e}^{(k+1)}\right\|_{\mathbf{A}} = \left\|u_{\mathcal{T}} - u_{\mathcal{T}}^{(k+1)}\right\|_{A} \le \eta_a^{(k+1)}. \tag{4.14}$$

Proof. Denote $p(k) := \rho_{\text{err}}^{(k)}$, $\xi(k) := p(k)/\left(1-p(k)\right)$, and $\xi := \rho_{\text{err}}/(1-\rho_{\text{err}})$. By Theorem 5, it suffices to show that: there exists an N such that for $k \ge N$

$$\xi \le e^{1/k}\xi(k). \tag{4.15}$$

It is straightforward to verify that $\xi(k) \to \xi$ from below as $p(k) \to \rho_{\text{err}}$. Moreover, $e^{1/k}\xi(k) \to \xi$ as $k \to \infty$. Now it suffices to show that when k is sufficiently large,

$e^{1/k}\xi(k)$ is a decreasing function of k. Recalling from (4.8) in Lemma 6 that if k is sufficiently large,

$$p(k) = \rho_{\text{err}}^{(k)} \simeq \rho_{\text{err}} \frac{1 + b\gamma^k}{1 + b\gamma^{k-1}}, \tag{4.16}$$

where $\gamma := (\lambda_{m+1}/\lambda_1)^2 < 1$, and $b := (c_{m+1}/c_1)^2 \geq 0$ where m is the multiplicity of the largest eigenvalue. Taking the derivative of $e^{1/k}\xi(k)$ with respect to k leads to:

$$\frac{d}{dk}\big(e^{1/k}\xi(k)\big) = e^{1/k} \frac{-k^{-2}p(k)\big(1 - p(k)\big) + p'(k)}{\big(1 - p(k)\big)^2}. \tag{4.17}$$

By (4.16), we have

$$p'(k) \simeq \rho_{\text{err}} \frac{b(\gamma - 1)\gamma^{k-1} \ln \gamma}{\big(1 + b\gamma^{k-1}\big)^2} = O(\gamma^{k-1}). \tag{4.18}$$

Using (4.18) in (4.17) and noting that k^{-2} decreases at a slower rate than γ^{k-1}, then for sufficiently large k, $\frac{d}{dk}(e^{1/k}\xi(k)) < 0$, and the theorem follows. ∎

REMARK 9 (Speed up of the rate of convergence estimate). *We notice that without the correction factor in (4.13), the closer $\rho_{\text{err}}^{(k)}$ is to ρ_{err}, the more accurate the algebraic estimator is. The convergence of $\rho_{\text{err}}^{(k)}$ can be accelerated in the following way:*

$$\rho_{\text{err}}^{(k)} \approx \rho_{\text{err}} \frac{1 + b\gamma^k}{1 + b\gamma^{k-1}} \quad , \textit{ and } \quad \rho_{\text{err}}^{(k-1)} \approx \rho_{\text{err}} \frac{1 + b\gamma^{k-1}}{1 + b\gamma^{k-2}}.$$

Now we define for $k \geq 2$:

$$\widehat{\rho}_{\text{err}}^{(k)} := \rho_{\text{err}}^{(k)} \frac{\rho_{\text{err}}^{(k)}}{\rho_{\text{err}}^{(k-1)}}. \tag{4.19}$$

And $\widehat{\rho}_{\text{err}}^{(k)}$ converges to ρ_{err} faster than the original $\rho_{\text{err}}^{(k)}$. To see this, taking derivative of $\widehat{\rho}_{\text{err}}^{(k)}$ with respect to k gives,

$$\frac{d}{dk}\big(\widehat{\rho}_{\text{err}}^{(k)}\big) = O(\gamma^{k-2}),$$

which is an order faster than the convergence of $\rho_{\text{err}}^{(k)}$ in (4.18).

5. Discretization-accurate stopping criterion. Identity (1.2) clearly indicates that the iterative solver should be stopped when the algebraic error is of the same magnitude as the discretization error. This observation suggests the following stopping criterion: let $\eta_d^{(k)}$ be η_d from (3.23) computed using the iterate $u_{\mathcal{T}}^{(k)}$, the iterative solver shall stop when

$$\eta_a^{(k)} < \varepsilon^{-1} \cdot \eta_d^{(k)} \quad \text{and} \quad \left|\rho_{\text{err}}^{(k)}/\rho_{\text{err}}^{(k-1)} - 1\right| < \varepsilon_\rho, \tag{5.1}$$

where $\varepsilon = \eta_d / \left\|u - u_{\mathcal{T}}\right\|_A$ is the effectivity index. In light of the proof of Theorem 8, the second condition implies that $\rho_{\text{err}}^{(k)}$ is a good approximation to ρ_{err} and, hence, $\eta_a^{(k)}$ is an accurate representation of the algebraic error $\left\|u_{\mathcal{T}} - u_{\mathcal{T}}^{(k)}\right\|_A$ at the k-th iteration. Together with the first condition, the estimated algebraic error is of the same magnitude in the discretization error.

6. Numerical examples. In this section, several examples are presented to verify the reliability of the estimators proposed, as well as the stopping criterion. The error estimator η_d, using a localized equilibrated flux to solve (3.15), is implemented in iFEM [14]. The initial guess for all examples presented in this section is a random guess with each entry of $\mathbf{u}^{(0)}$ satisfying a uniform distribution in $[-1, 1]$ using a fixed seed. An effectivity index of $\varepsilon = 1.5$ or $\varepsilon^{-1} = 2/3 \approx 0.67$ is used in (5.1). This is similar to typical values used in practice when $u_{\mathcal{T}}$ is computed with a direct solver.

The first test problem is the Poisson equation

$$-\Delta u = f, \text{ in } \Omega = (-1, 1)^2$$

with Dirichlet boundary conditions and the exact solution is given by

$$u = \alpha\Big(\sin(\pi x)\sin(\pi y) + 0.5\sin(4\pi x)\sin(4\pi y)\Big),$$

where the constant α is chosen such that $\|u\|_A = 1$. This problem is discretized by the continuous piecewise linear finite element method on a uniform triangular mesh with mesh size $h = 1/32$.

The resulting system of algebraic equations is first solved by a multigrid method with $V(1,1)$-cycle. Convergence of the multigrid solver in the energy norm along with the algebraic estimator are depicted in Figure 1a (see the red and blue dot-circle lines), which numerically verify Theorem 8 for the algebraic estimator η_a being an upper bound of the algebraic error. The total and the discretization errors along with the discretization estimator are also depicted in Figure 1a (see the red solid-diamond, the red dot, and the blue solid-diamond lines, respectivley). Estimated convergence rates based on both $\rho_{\text{err}}^{(k)}$ and $\hat{\rho}_{\text{err}}^{(k)}$ are presented to numerically verify Remark 9.

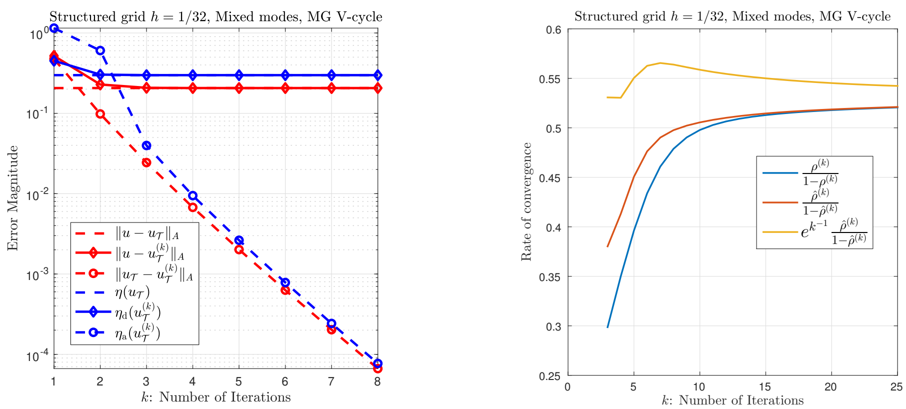
Using the first stopping criterion in (5.1) with $\varepsilon^{-1} = 0.67$, the multigrid iteration stops after merely two iterations, and Figure 1a shows that the algebraic error already drops below the discretization error. For a conventional stopping criterion using the relative residual measured in the ℓ^2-norm: $\left\|\mathbf{A}\mathbf{e}^{(k)}\right\|_0 / \left\|\mathbf{A}\mathbf{e}^{(0)}\right\|_0 \le 10^{-7}$, the multigrid iteration stops after fifteen iterations. For a slower iterative solver, we also implement symmetric Gauss-Seidel iterative method. The first stopping criterion in (5.1) with $\varepsilon^{-1} = 0.67$ requires only thirty-one iterations, while the conventional stopping criterion with the tolerance 10^{-5} needs more than two hundred eighty iterations. These results show a dramatic reduction in computational cost when using the discretization-accurate stopping criterion introduced in this paper. The numbers of iterations for the multigrid and the symmetric Gauss-Seidel iterative methods with both the stopping criterions as well as the total and the algebraic errors are summarized in Table 1. As observed from Table 1, additional iterations needed by the conventional stopping criterion significantly decrease the algebraic errors but not the total errors. Figure 2 compares the solution u_h obtained by a direct solver with that of a multigrid solver after 2 iterations.

The second test problem tests the stopping criterion on a non-uniform mesh for the Kellogg intersecting interface problem. The Kellogg problem with a checkerboard coefficient distribution [10] is a commonly used benchmark for testing the efficiency and robustness of a posteriori error estimators ([13, 11, 12, 15, 23]):

$$-\nabla\cdot(A\nabla u) = 0, \text{ in } \Omega = (-1, 1)^2 \tag{6.1}$$

Table 1: The number of iterations and the total and algebraic errors for the Poisson problem.

	Stopping	# Iter	$\left\|u - u_{\mathcal{T}}^{(k)}\right\|_A$	$\left\|u_{\mathcal{T}} - u_{\mathcal{T}}^{(k)}\right\|_A$
MG V(1,1)	$\eta_a \leq 0.67\eta_d$	2	0.0821	3.5×10^{-1}
MG V(1,1)	$\|\mathbf{r}_k\|_2/\|\mathbf{r}_0\|_2 \leq 10^{-7}$	15	0.0741	3.4×10^{-8}
Sym GS	$\eta_a \leq 0.67\eta_d$	31	0.1051	7.5×10^{-1}
Sym GS	$\|\mathbf{r}_k\|_2/\|\mathbf{r}_0\|_2 \leq 10^{-5}$	289	0.0741	2.7×10^{-4}



(a) The convergence of the $V(1,1)$-cycles.

(b) The convergence of the estimated rate of convergence

Fig. 1: The convergence results for the Poisson problem: the solution has mixed modes, the problem is discretized on a uniform triangular mesh, and the linear system is approximated using $V(1,1)$-cycle iterations.

with Dirichlet boundary condition, where the diffusion coefficient A is given by

$$A = \begin{cases} R & \text{in } (0,1)^2 \cup (-1,0)^2, \\ 1 & \text{in } \Omega \backslash \Big((0,1)^2 \cup (-1,0)^2 \Big). \end{cases}$$

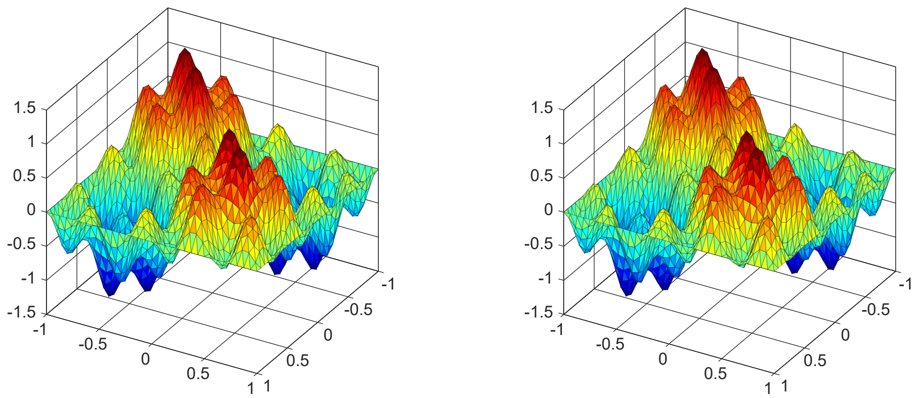
The exact solution u of (6.1) is given in polar coordinates (r,θ):

$$u = r^{\gamma}\psi(\theta) \in H^{1+\gamma-\epsilon}(\Omega) \ \text{ for any } \ \epsilon > 0,$$

where the definition of $\psi(\theta)$ is given in, e.g., [15]. Here the parameters are:

$$\gamma = 0.5, \quad R \approx 5.8284271247461907, \quad \rho = \pi/4, \ \text{ and } \ \sigma \approx -2.3561944901923448.$$

For this example, $\mathcal{T}$ is a graded mesh on which the relative error for the direct solve $\|u - u_{\mathcal{T}}\|_A \,/\, \|u\|_A \approx 10\%$, in addition, we choose $\varepsilon^{-1} = 0.67$ and $\varepsilon_\rho = 0.1$ for the stopping criterion. The stopping criterion (5.1) is checked every three $V(1,1)$-cycles. The local error distribution is shown in Figure 3.



(a) u_h obtained by a direct solver.

(b) $u_h^{(2)}$ obtained by two $V(1,1)$-cycles.

Fig. 2: The comparison of the direct-solved approximation u_h and the multigrid iterate $u_h^{(2)}$ in the first test problem.

Table 2: The number of iterations and the total and algebraic errors for the Kellogg problem.

	Stopping	# Iter	$\left\|u - u_\mathcal{T}^{(k)}\right\|_A$	$\left\|u_\mathcal{T} - u_\mathcal{T}^{(k)}\right\|_A$
MG V(1,1)	$\eta_a \le 0.67\eta_d$	2	0.05141	1.577×10^{-3}
MG V(1,1)	$\|\mathbf{r}_k\|_2 / \|\mathbf{r}_0\|_2 \le 10^{-7}$	6	0.05139	8.026×10^{-8}

REFERENCES

[1] M. Ainsworth and J. T. Oden, *A posteriori error estimation in finite element analysis*, vol. 37, John Wiley & Sons, 2011. 1

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

[3] M. Arioli, J. Liesen, A. Miedlar, and Z. Strakoš, *Interplay between discretization and algebraic computation in adaptive numerical solution of elliptic PDE problems*, STFC, 2012. 2

[4] M. Arioli, D. Loghin, and A. J. Wathen, *Stopping criteria for iterations in finite element methods*, Numerische Mathematik, 99 (2005), pp. 381–410. 2

[5] R. Becker, C. Johnson, and R. Rannacher, *Adaptive error control for multigrid finite element*, Computing, 55 (1995), pp. 271–288. 2

[6] C. Bernardi and R. Verfürth, *Adaptive finite element methods for elliptic equations with non-smooth coefficient*, Numerische Mathematik, 85 (2000), pp. 579–608. 2

[7] D. Braess, *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*, Cambridge University Press, 2007. 6, 9

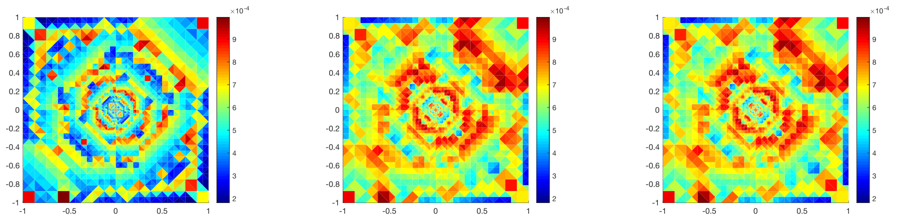
[8] D. Braess and J. Schöberl, *Equilibrated residual error estimator for edge elements*, Math. Comp., 77 (2008), pp. 651–672. 2, 6, 7

[9] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer, 1991. 5, 6, 7

[10] R. Bruce Kellogg, *On the poisson equation with intersecting interfaces*, Applicable Analysis, 4 (1974), pp. 101–129. 15

[11] Z. Cai and S. Cao, *A recovery-based a posteriori error estimator for H(curl) interface problems*, Comput. Methods in Appl. Mech. Eng., 296 (2015), pp. 169–195. 15

[12] Z. Cai, S. Cao, and R. Falgout, *Robust a posteriori error estimation for finite element approximation to H(curl) problem*, Comput. Methods in Appl. Mech. Eng., 309 (2016),



(a) $\|u - u_{\mathcal{T}}\|_{A,K}$, the local energy error.

(b) The local error indicator $\eta_{d,K}$ in (3.23) using direct solve $u_{\mathcal{T}}$.

(c) The local error indicator $\eta_{d,K}$ using iterate $\bar{u}_{\mathcal{T}}$.

Fig. 3: The comparison the local error and the error indicator distributions.

pp. 182–201. 15

[13] Z. Cai and S. Zhang, *Robust equilibrated residual error estimator for diffusion problems: Conforming elements*, SIAM J. Numer. Anal., 50 (2012), pp. 151–170. 2, 7, 15

[14] L. Chen, *iFEM: an innovative finite element methods package in MATLAB*, (2008). 15

[15] Z. Chen and S. Dai, *On the efficiency of adaptive finite element methods for elliptic problems with discontinuous coefficients*, SIAM Journal on Scientific Computing, 24 (2002), pp. 443–462. 15, 16

[16] P. Destuynder and B. Métivet, *Explicit error bounds in a conforming finite element method*, Mathematics of Computation of the American Mathematical Society, 68 (1999), pp. 1379–1396. 2

[17] V. Dolejší, I. Šebestová, and M. Vohralík, *Algebraic and discretization error estimation by equilibrated fluxes for discontinuous galerkin methods on nonmatching grids*, Journal of Scientific Computing, (2013), pp. 1–34. 2

[18] M. Dryja, M. V. Sarkis, and O. B. Widlund, *Multilevel schwarz method for elliptic problems with discontinuous coefficients in three dimensions*, Numerische Mathematik, 72 (1996), pp. 313–348. 2

[19] I. Ekeland and R. Temam, *Convex analysis and variational problems*, (1976). 5

[20] P. Jiránek, Z. Strakoš, and M. Vohralík, *A posteriori error estimates including algebraic error and stopping criteria for iterative solvers*, SIAM Journal on Scientific Computing, 32 (2010), pp. 1567–1590. 2

[21] R. Luce and B. I. Wohlmuth, *A local a posteriori error estimator based on equilibrated fluxes*, SIAM Journal on Numerical Analysis, 42 (2004), pp. 1394–1414. 2

[22] J. Ortega and W. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Classics in Applied Mathematics, SIAM, 1970. 12

[23] M. Petzoldt, *A posteriori error estimators for elliptic equations with discontinuous coefficients*, Advances in Computational Mathematics, 16 (2002), pp. 47–75. 2, 15

[24] R. Rannacher, *Error control in finite element computations*, Springer, 1999. 2

[25] R. Verfürth, *A note on constant-free a posteriori error estimates*, SIAM Journal on Numerical Analysis, 47 (2009), pp. 3180–3194. 2

[26] R. Verfürth, *A posteriori error estimation techniques for finite element methods*, OUP Oxford, 2013. 1