A posteriori error estimates including algebraic error: computable upper bounds and stopping criteria for iterative solvers

Pavel Jiranek, Zdenek Strakos, Martin Vohralík

A POSTERIORI ERROR ESTIMATES INCLUDING ALGEBRAIC ERROR: COMPUTABLE UPPER BOUNDS AND STOPPING CRITERIA FOR ITERATIVE SOLVERS

PAVEL JIRÁNEK*, ZDENĚK STRAKOŠ†, AND MARTIN VOHRALÍK‡

Abstract. We consider the finite volume and the lowest-order mixed finite element discretizations of a second-order elliptic pure diffusion model problem. The first goal of this paper is to derive guaranteed and fully computable a posteriori error estimates which take into account an inexact solution of the associated linear algebraic system. We show that the algebraic error can be simply bounded using the algebraic residual vector. Much better results are, however, obtained using the complementary energy of an equilibrated Raviart–Thomas–Nédélec discrete vector field whose divergence is given by a proper weighting of the residual vector. The second goal of this paper is to construct efficient stopping criteria for iterative solvers such as the conjugate gradients, GMRES, or Bi-CGStab. We claim that the discretization error, implied by the given numerical method, and the algebraic one should be in balance, or, more precisely, that it is enough to solve the linear algebraic system to the accuracy which guarantees that the algebraic part of the error does not contribute significantly to the whole error. Our estimates allow a reliable and cheap comparison of the discretization and algebraic errors. One can thus use them to stop the iterative algebraic solver at the desired accuracy level, without performing an excessive number of unnecessary additional iterations. Under the assumption of the relative balance between the two errors, we also prove the efficiency of our a posteriori estimates, i.e., we show that they also represent a lower bound, up to a generic constant, for the overall energy error. A local version of this result is also stated. Several numerical experiments illustrate the theoretical results.

Key words. Second-order elliptic partial differential equation, finite volume method, mixed finite element method, a posteriori error estimates, iterative methods for linear algebraic systems, stopping criteria.

AMS subject classifications. 65N15, 65N30, 76M12, 65N22, 65F10

1. Introduction.

In numerical solution of partial differential equations, the computed result is an approximate solution found in some finite-dimensional space. A natural question is whether this solution is a sufficiently accurate approximation of the exact (weak) solution of the problem at hand. A posteriori error estimates aim at giving an answer to this question while providing upper bounds on the error between the approximate and exact solutions that can be easily computed. Their mathematical theory was started by the pioneering paper by Babuška and Rheinboldt [6] and a vast amount of literature on this subject exists nowadays; we refer, e.g., to the books by Verfürth [45] or Ainsworth and Oden [2]. Apart from few exceptions, they rely on the assumption that the linear system resulting from discretization *is solved exactly*. This is not assumed, e.g., in the work by Wohlmuth and Hoppe [52], but the bounds,

*Faculty of Mechatronics and Interdisciplinary Engineering Studies, Technical University of Liberec, Hálkova 6, 46117 Liberec, Czech Republic (pavel.jiranek@tul.cz). The work of this author was supported by the MSMT CR under the project 1M0554 "Advanced Remedial Technologies" and by the project IAA100300802 of the GAAS.

†Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 18207 Prague, Czech Republic (strakos@cs.cas.cz). The work of this author was supported by the project IAA100300802 of the GAAS and by the Institutional Research Plan AV0Z10300504 "Computer Science for the Information Society: Models, Algorithms, Applications."

‡UPMC Univ. Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions, 75005 Paris, France & CNRS, UMR 7598, Laboratoire Jacques-Louis Lions, 75005 Paris, France (vohralik@ann.jussieu.fr). The work of this author was supported by the GNR MoMaS project "Numerical Simulations and Mathematical Modeling of Underground Nuclear Waste Disposal", PACEN/CNRS, ANDRA, BRGM, CEA, EdF, IRSN, France.

taking into account possible errors of the linear algebraic solver, are valid only for a sufficiently refined mesh, and/or contain various unspecified constants. Therefore a practical overall error control and balancing the discretization and algebraic error is not possible. Rüde gives in [35] estimates of the energy norm of the error based on the norms of the residual functionals obtained from some particular stable splitting of the underlying Hilbert space. Repin [32, 33] and Korotov [22] do not use any information about the discretization method and the method for solving the resulting linear algebraic system. This makes their estimates very general but the price is that they may be quite costly and not sufficiently accurate, see the theoretical comparison in [49] and [15]. It is again not possible to compare the discretization and algebraic errors and to construct stopping criteria for algebraic iterative solvers.

Ignoring (for the moment) rounding errors, a moderate size system of linear algebraic equations can be solved exactly. If a direct method can be used and the linear algebraic problem is reasonably well conditioned, then one can get a highly accurate solution even in finite precision arithmetic. The same is true for some iterative methods providing that one performs a sufficient number of iterations. When the linear algebraic system is large, (preconditioned) iterative methods become competitive with direct ones, and in many cases they represent the only viable alternative. Here it should be emphasized that applications of direct and iterative methods are *principally different*. While in direct methods the whole solution process must be completed to the very end in order to get a meaningful numerical solution, iterative methods can produce an approximation of the solution *at each iteration step*. The amount of computational work depends on the number of iterations performed, and an efficient PDE solver should use this principal advantage by stopping the algebraic solver whenever the algebraic error drops to the level at which it does not significantly affect the whole error (cf. [5]). In other words, the linear algebraic system is affected by the errors on the preceding stages (modeling and discretization errors) of the solution process. Therefore it does not represent the investigated problem accurately. Solving an inaccurate linear system to a non-needed high accuracy is meaningless. It represents nothing but wasting computational time and resources. Similarly, comparison of direct and iterative algebraic solvers at the same high accuracy level can be misleading, since the high accuracy may not be needed (for a detailed discussion we refer to [40]).

Efficient use of iterative solvers requires reliable stopping criteria. The simplest, most often used, and mathematically most questionable stopping criterion is based on evaluation of the relative Euclidean norm of the residual vector, see, e.g., the discussion in [21, Section 17.5]. There is only a rough connection of the relative residual norm to the whole error in approximation of the continuous problem (we discuss this point in detail in Section 7.1 below) and, usually, not even this connection is considered. Consequently, one either continues the iterations until the residual norm is not further reduced (i.e., one uses the iterative solver essentially as a direct solver, possibly wasting resources and computational time without getting any further improvement of the whole error), or stops earlier at a risk that the computed solution is not sufficiently accurate. For some enlightening comments we refer, e.g., to [28].

The question of stopping criteria has been already addressed by, e.g., Becker et al. [8]. In connection with numerical discretization, this paper uses the residual error estimate and develops a stopping criterion for the multigrid solver. However, the constants resulting from the general interpolation bounds can deteriorate the effectivity index, i.e., the ratio of the error estimate and the actual error. A remarkable

approach relating the algebraic and discretization errors is represented by the so-called cascadic conjugate gradient method of Deuflhard [12], which was further studied by several other authors, see, e.g., [38]. In [3], Arioli compares the bound on the discretization error with the error of the iterative method when solving self-adjoint second-order elliptic problems. He uses the relationship between the energy norm defined in the underlying Hilbert space for the weak formulation and its restriction onto the discrete space, in combination with the numerically stable algebraic error bounds [41], see also [42]. Arioli et al. [4] extend these results for non self-adjoint problems. Their approach is interesting and useful in some applications but relies on an *a priori* knowledge, not an a posteriori bound for the discretization error. Stopping the algebraic iterative solver based on a priori information on the discretization error is also applied in the context of wavelet discretizations of elliptic partial differential equations by Burstedde and Kunoth [10]. Finally, the interesting technique of Patera and Rønquist [28], see also Maday and Patera [23], gives computable lower and upper asymptotic bounds of a linear functional of an approximate linear system solution and hence, if the asymptotic bound property is obtained for some reasonable number of iterations, a stopping criterion. It is, however, tailored to a fast converging preconditioned primal-dual conjugate gradient Lanczos method, and, at least in the presented form, it does not relate the discretization and algebraic parts of the error. Moreover, it does not fully eliminate numerical uncertainty.

In this paper we consider a second-order elliptic pure diffusion model problem: find a real-valued function p defined on Ω such that

$$-\nabla\cdot(\mathbf{S}\nabla p) = f \quad \text{in} \quad \Omega, \qquad p = g \quad \text{on} \quad \Gamma := \partial\Omega, \tag{1.1}$$

where Ω is a polygonal/polyhedral domain (open, bounded, and connected set) in $\mathbb{R}^d$, $d = 2, 3$, $\mathbf{S}$ is a diffusion tensor, f is a source term, and g prescribes the Dirichlet boundary condition. Details are given in Section 2. For the discretization of problem (1.1) on simplicial meshes we consider two classes of numerical methods recalled in Section 3. First, cell-centered finite volume schemes are included under the condition that they are written, by prescribing the discrete diffusive fluxes, as a conservation equation over each computational cell. For a general survey of such methods we refer to Eymard et al. [16]. The second class consists of lowest-order mixed finite element methods, cf. Brezzi and Fortin [9] or Quarteroni and Valli [30]. In certain parts we build upon the close relationships of these methods derived in [47].

The first goal of this paper is to derive a posteriori error estimates which take into account an *inexact solution of the associated linear algebraic system*. After describing in Section 4 the inexact solution of linear algebraic equations, we extend in Section 5 for this purpose the a posteriori error estimates proposed and analyzed in [48, 50]. The derived upper bound for the overall error is *guaranteed and fully computable*. It consists of three independent estimators: an estimator measuring the nonconformity of the approximate solution, which essentially reflects the discretization error; a residual estimator which in general turns out to be a higher-order term corresponding to the interpolation error in the approximation of the source term f; and an abstract algebraic error estimator corresponding to the inexact solution of the discrete linear algebraic problem. The abstract algebraic error estimator is quite general. It is based on equilibrated vector fields $\mathbf{r}_h$ from the lowest-order Raviart–Thomas–Nédélec space whose divergences are given by a proper weighting of the algebraic residual vector.

The second goal of this paper is to construct, in the context of solving problem (1.1), efficient *stopping criteria for iterative solvers* such as the conjugate gradient

(CG) method [20], GMRES [37], or Bi-CGStab [43], see, e.g., the standard monograph Saad [36]. We undertake it in Section 6. We claim that the discretization and the algebraic errors should be in balance, or, more precisely, that it is enough to solve the linear algebraic system to the accuracy which guarantees that the algebraic part of the error does not contribute significantly to the whole error. Our approach allows a reliable and cheap comparison of the discretization and algebraic errors and one can thus use it to stop the iterative algebraic solver at the desired accuracy level. Under the assumption of the relative balance between the estimates on the two errors we also prove the efficiency of our a posteriori estimates. Recall that our estimates represent an upper bound for the overall energy error. Efficiency then means that they also represent a lower bound for the overall energy error, up to a generic constant only dependent on the space dimension, shape regularity of the mesh, and the tensor $\mathbf{S}$. In other words, they guarantee an upper bound for the error which is such that the overestimation is moderate and *independent* of the weak solution regularity, domain size, mesh refinement level, specific discretization (of locally conservative type), algebraic solver, and other factors. Moreover, using a stopping criterion where the estimate on the algebraic error is bounded by the estimate on the discretization one *locally* in each mesh element, we also prove the *local efficiency* of our estimates. This means that the estimated error in each mesh element represents, as above, a lower bound for the energy error in the given element and in its close neighborhood, up to a generic constant. Consequently, our estimates are suitable for adaptive mesh refinement as they can correctly predict the overall error size and distribution.

The algebraic error estimator of Section 5 is abstract and it cannot be computed in practice. The purpose of Section 7 is to give its fully computable upper bounds. The first upper bound is given directly by the components of the algebraic residual vector, and the vector field $\mathbf{r}_h$ actually does not appear here. Though this way is simple and it can be used in some cases, it in general highly overestimates the algebraic error, in particular in connection with adaptive mesh refinement and for highly discontinuous tensor $\mathbf{S}$. In the second approach, we relate the abstract algebraic error to the complementary energy $\|\mathbf{S}^{-\frac{1}{2}}\mathbf{q}_h\|$ of such $\mathbf{q}_h$ which minimizes $\|\mathbf{S}^{-\frac{1}{2}}\mathbf{r}_h\|$ among all Raviart–Thomas–Nédélec discrete vector fields $\mathbf{r}_h$ whose divergence is given by the weighted residual vector. Consider now for the moment the mixed finite element discretization of the residual problem (cf. (5.9) below) with its corresponding Schur complement matrix. It then follows that the algebraic energy error induced by this matrix is equal to the above minimal discrete complementary energy. Consequently, known estimates on the algebraic energy error, see, e.g., [41, 42], can in this case be used to estimate the algebraic error. They can be numerically very efficient, although they do not give mathematically guaranteed upper bounds. Finally, the last approach is based on a factual construction of a vector field $\mathbf{r}_h$ and on the use of its complementary energy as the algebraic error estimator. It is simple and gives a guaranteed and fully computable upper bound on the algebraic error for all discretizations considered in this paper. It also bounds from above the preceding algebraic error estimator. In comparison with this preceding estimator, only a few more iterations of the iterative solver are necessary to guarantee the given accuracy. All three approaches are numerically illustrated in Section 8 on several examples.

2. Preliminaries. In this section we introduce the notation, partitions of the domain, state the assumptions on the data, and give details on the continuous problem (1.1).

2.1. Notation and assumptions. The notation that we use is standard, see [11, 9, 16], and it is included here for completeness. It can be skipped and used as a reference, if needed, while reading the rest of the paper.

Recall that Ω is a polygonal domain in $\mathbb{R}^2$ or a polyhedral domain in $\mathbb{R}^3$ with the boundary Γ. Let $\mathcal{T}_h$ be a partition of Ω into closed simplices, i.e., triangles if $d = 2$ and tetrahedra if $d = 3$, such that $\overline{\Omega} = \cup_{K\in\mathcal{T}_h} K$. Moreover, we assume that the partition is conforming in the sense that if $K, L \in \mathcal{T}_h$, $K \neq L$, then $K \cap L$ is either an empty set, a common face, edge, or vertex of K and L. For $K \in \mathcal{T}_h$, we denote by $\mathcal{E}_K$ the set of sides (edges if $d = 2$, faces if $d = 3$) of K, by $\mathcal{E}_h = \cup_{K\in\mathcal{T}_h}\mathcal{E}_K$ the set of all sides of $\mathcal{T}_h$, and by $\mathcal{E}_h^{\rm int}$ and $\mathcal{E}_h^{\rm ext}$, respectively, the interior and exterior sides. We also use the notation $\mathfrak{E}_K$ for the set of all $\sigma \in \mathcal{E}_h^{\rm int}$ which share at least a vertex with a $K \in \mathcal{T}_h$. For interior sides such that $\sigma = \sigma_{K,L} := \partial K \cap \partial L$, i.e., $\sigma_{K,L}$ is a part of the boundary ∂K and, at the same time, a part of the boundary ∂L, we shall call K and L neighbors and we denote the set of neighbors of a given element $K \in \mathcal{T}_h$ by $\mathcal{T}_K$; $\mathfrak{T}_K$ stands for all triangles sharing at least a vertex with $K \in \mathcal{T}_h$. For $K \in \mathcal{T}_h$, $\mathbf{n}$ will always denote its exterior normal vector; we shall also employ the notation $\mathbf{n}_\sigma$ for a normal vector of a side $\sigma \in \mathcal{E}_h$, whose orientation is chosen arbitrarily but fixed for interior sides and coinciding with the exterior normal of Ω for exterior sides. For $\sigma_{K,L} \in \mathcal{E}_h^{\rm int}$ such that $\mathbf{n}_\sigma$ points from K to L and a function φ we also define the jump operator $[\![\cdot]\!]$ by $[\![\varphi]\!] := (\varphi|_K)|_\sigma - (\varphi|_L)|_\sigma$. Finally, a family of meshes $\mathcal{T} := \{\mathcal{T}_h;\, h > 0\}$ is parameterized by $h := \max_{K\in\mathcal{T}_h} h_K$, where h_K is the diameter of K (we also denote by h_σ the diameter of $\sigma \in \mathcal{E}_h$).

For a given domain $S \subset \mathbb{R}^d$, let $L^2(S)$ be the space of square-integrable (in the Lebesgue sense) functions over S, $(\cdot,\cdot)_S$ the $L^2(S)$ inner product, and $\|\cdot\|_S$ the associated norm (we omit the index S when $S = \Omega$). By $|S|$ we denote the Lebesgue measure of S and by $|\sigma|$ the $(d-1)$-dimensional Lebesgue measure of a $(d-1)$-dimensional surface σ in $\mathbb{R}^d$. Let $\mathcal{H}(S)$ be a set of real-valued functions defined on S. By $[\mathcal{H}(S)]^d$ we denote the set of vector functions with d components each belonging to $\mathcal{H}(S)$. Let next $H^1(S)$ be the Sobolev space with square-integrable weak derivatives up to order one, $H_0^1(S) \subset H^1(S)$ its subspace of functions with traces vanishing on Γ, $H^{1/2}(S)$ the trace space, $\mathbf{H}(\mathrm{div}, S) := \{\mathbf{v} \in [L^2(S)]^d;\, \nabla\cdot\mathbf{v} \in L^2(S)\}$ the space of functions with square-integrable weak divergences, and let finally $\langle\cdot,\cdot\rangle_{\partial S}$ stand for $(d-1)$-dimensional $L^2(\partial S)$-inner product on ∂S. We also let $H^1_\Gamma(\Omega) := \{\varphi \in H^1(\Omega);\, \varphi|_\Gamma = g\}$ be the set of functions satisfying the Dirichlet boundary condition on Γ in the sense of traces. For a given partition $\mathcal{T}_h$ of Ω, let $H^1(\mathcal{T}_h) := \{\varphi \in L^2(\Omega);\, \varphi|_K \in H^1(K)\ \forall K \in \mathcal{T}_h\}$ be the broken Sobolev space. Finally, we let $W_0(\mathcal{T}_h)$ be the space of functions with mean values of the traces continuous across interior sides, i.e., $W_0(\mathcal{T}_h) := \{\varphi \in H^1(\mathcal{T}_h);\, \langle[\![\varphi]\!], 1\rangle_\sigma = 0\ \forall\sigma \in \mathcal{E}_h^{\rm int}\}$.

We next denote by $\mathbb{P}_k(S)$ the space of polynomials on S of total degree less than or equal to k and by $\mathbb{P}_k(\mathcal{T}_h) := \{\varphi_h \in L^2(\Omega);\, \varphi_h|_K \in \mathbb{P}_k(K)\ \forall K \in \mathcal{T}_h\}$ the space of piecewise k-degree polynomials on $\mathcal{T}_h$. We define $\mathbf{RTN}(K) := [\mathbb{P}_0(K)]^d + \mathbf{x}\mathbb{P}_0(K)$ for an element $K \in \mathcal{T}_h$ the local and $\mathbf{RTN}(\mathcal{T}_h) := \{\mathbf{v}_h \in [L^2(\Omega)]^d;\, \mathbf{v}_h|_K \in \mathbf{RTN}(K)\ \forall K \in \mathcal{T}_h\} \cap \mathbf{H}(\mathrm{div}, \Omega)$ the global lowest-order Raviart–Thomas–Nédélec space of specific piecewise linear vector functions. Recall that the normal components of $\mathbf{v}_h \in \mathbf{RTN}(K)$, $\mathbf{v}_h \cdot \mathbf{n}$, are constant on each $\sigma \in \mathcal{E}_K$ and that they represent the degrees of freedom of $\mathbf{RTN}(K)$. By consequence, the constraint $\mathbf{v}_h \in \mathbf{H}(\mathrm{div}, \Omega)$ imposing the normal continuity of the traces is expressed as $\mathbf{v}_h|_K \cdot \mathbf{n} + \mathbf{v}_h|_L \cdot \mathbf{n} = 0$ for all $\sigma_{K,L} \in \mathcal{E}_h^{\rm int}$ and there is still one degree of freedom per side in $\mathbf{RTN}(\mathcal{T}_h)$. Recall also that $\nabla\cdot\mathbf{v}_h|_K$ is constant for $\mathbf{v}_h \in \mathbf{RTN}(K)$. For more details, we refer to Brezzi

and Fortin [9] or Quarteroni and Valli [30].

In the paper, we make the following assumption on the data in problem (1.1):

ASSUMPTION 2.1 (Data). *Let* $\mathbf{S}$ *be a symmetric, bounded, and uniformly positive definite tensor, piecewise constant on* $\mathcal{T}_h$. *Let in particular* $c_{\mathbf{S},K} > 0$ *and* $C_{\mathbf{S},K} > 0$ *denote its smallest and biggest eigenvalues on each* $K \in \mathcal{T}_h$. *In addition, let* $f \in \mathbb{P}_l(\mathcal{T}_h)$ *be an elementwise* l*-degree polynomial function and* $g \in H^{1/2}(\Gamma)$.

The assumptions on $\mathbf{S}$ and f are made for the sake of simplicity and are usually satisfied in practice. Otherwise, interpolation can be used in order to get the desired properties. In the sequel, we will employ the notation $\mathbf{S}_K := \mathbf{S}|_K$, and, in general, $\varphi_K := \varphi_h|_K$ for $\varphi_h \in \mathbb{P}_0(\mathcal{T}_h)$.

2.2. Continuous problem. We define a bilinear form $\mathcal{B}$ by

$$\mathcal{B}(p,\varphi) := \sum_{K\in\mathcal{T}_h} (\mathbf{S}\nabla p, \nabla\varphi)_K, \qquad p,\varphi \in H^1(\mathcal{T}_h)$$

and the corresponding energy norm by

$$|||\varphi|||^2 := \mathcal{B}(\varphi,\varphi). \tag{2.1}$$

Note that $\mathcal{B}$ is well-defined for functions from the space $H^1(\Omega)$ as well as from the broken space $H^1(\mathcal{T}_h)$. The weak formulation of problem (1.1) is then to find $p \in H^1_\Gamma(\Omega)$ such that

$$\mathcal{B}(p,\varphi) = (f,\varphi) \qquad \forall\varphi \in H^1_0(\varphi). \tag{2.2}$$

Assumption 2.1 implies that problem (2.2) admits a unique solution [11].

3. Finite volume methods, mixed finite element methods, and postprocessing. We first introduce here the finite volume and mixed finite element methods for problem (1.1), see [16, 9]. The original approximations p_h in these methods are only piecewise constant and they are not appropriate for an energy a posteriori error estimate, as $\nabla p_h = 0$. We therefore construct a locally postprocessed approximation using information about the known fluxes. Finally, we will in the a posteriori error estimates need a $H^1(\Omega)$-conforming approximation using the Oswald interpolation operator.

3.1. Finite volume methods. A general cell-centered finite volume method for problem (1.1) can be written in the following form: find $p_h \in \mathbb{P}_0(\mathcal{T}_h)$ such that

$$\sum_{\sigma\in\mathcal{E}_K} U_{K,\sigma} = f_K|K| \qquad \forall K \in \mathcal{T}_h, \tag{3.1}$$

where $f_K := (f,1)_K/|K|$ and $U_{K,\sigma}$ is the diffusive flux through the side σ of an element K, see, e.g., [16]. We assume that the fluxes $U_{K,\sigma}$ depend linearly on the values of p_h, so that equations (3.1) represent a system of linear algebraic equations of the form

$$\mathbb{S}P = H, \tag{3.2}$$

where $\mathbb{S} \in \mathbb{R}^{N\times N}$ and $P, H \in \mathbb{R}^N$ with N being the number of elements in the partition $\mathcal{T}_h$. Here we only assume the continuity of the fluxes, i.e., $U_{K,\sigma} = -U_{L,\sigma}$ for all $\sigma = \sigma_{K,L} \in \mathcal{E}_h^{\mathrm{int}}$, so that practically all finite volume schemes can be included in our analysis. We give an example which clarifies the ideas.

Let there be a point $\mathbf{x}_K \in K$ for each $K \in \mathcal{T}_h$ such that if $\sigma_{K,L} \in \mathcal{E}_h^{\mathrm{int}}$, then $\mathbf{x}_K \neq \mathbf{x}_L$ and the straight line connecting $\mathbf{x}_K$ and $\mathbf{x}_L$ is orthogonal to $\sigma_{K,L}$. Let an analogous orthogonality condition hold also on the boundary. Then $\mathcal{T}_h$ is admissible in the sense of [16, Definition 9.1]. Under the additional assumption $\mathbf{S}_K = s_K \mathbb{I}$ ($\mathbb{I}$ denotes the identity matrix) on each $K \in \mathcal{T}_h$, the following choice is possible:

$$\begin{aligned} U_{K,\sigma} &= -s_{K,L} \frac{|\sigma_{K,L}|}{d_{K,L}} (p_L - p_K) \quad \text{for } \sigma = \sigma_{K,L} \in \mathcal{E}_h^{\mathrm{int}}, \\ U_{K,\sigma} &= -s_K \frac{|\sigma|}{d_{K,\sigma}} (g_\sigma - p_K) \qquad \text{for } \sigma \in \mathcal{E}_K \cap \mathcal{E}_h^{\mathrm{ext}}. \end{aligned} \tag{3.3}$$

Here p_K are the cell values of p_h ($p_K := p_h|_K$ for all $K \in \mathcal{T}_h$) and the value of $s_{K,L}$ on a side $\sigma = \sigma_{K,L} \in \mathcal{E}_h^{\mathrm{int}}$ is given by

$$s_{K,L} = \omega_{\sigma,K} s_K + \omega_{\sigma,L} s_L,$$

where $\omega_{\sigma,K} = \omega_{\sigma,L} = \frac{1}{2}$ in the case of the arithmetic averaging and $\omega_{\sigma,K} = s_L/(s_K + s_L)$ and $\omega_{\sigma,L} = s_K/(s_K + s_L)$ in the case of the harmonic averaging. The symbol $d_{K,L}$ stands for the Euclidean distance between the points $\mathbf{x}_K$ and $\mathbf{x}_L$ and $d_{K,\sigma}$ for the distance between $\mathbf{x}_K$ and $\sigma \in \mathcal{E}_K \cap \mathcal{E}_h^{\mathrm{ext}}$. Finally, $g_\sigma := \langle g, 1 \rangle_\sigma / |\sigma|$ is the mean value of g on a side $\sigma \in \mathcal{E}_h^{\mathrm{ext}}$. To express (3.1), (3.3) in the matrix form (3.2), let the elements of $\mathcal{T}_h$ be enumerated using a bijection $\ell : \mathcal{T}_h \to \{1, \dots, N\}$. With the corresponding ordering of unknown values p_K of p_h defined by $(P)_{\ell(K)} = p_K$ for each $K \in \mathcal{T}_h$, and denoting respectively by $(\cdot)_{kl}$ and $(\cdot)_k$ the matrix and vector components, the system matrix $\mathbb{S}$ and the right-hand side vector H are all zero except the elements defined by

$$\begin{aligned} (\mathbb{S})_{\ell(K),\ell(K)} &= \sum_{L \in \mathcal{T}_K} s_{K,L} \frac{|\sigma_{K,L}|}{d_{K,L}} + \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_h^{\mathrm{ext}}} s_K \frac{|\sigma|}{d_{K,\sigma}}, \\ (\mathbb{S})_{\ell(K),\ell(L)} &= -s_{K,L} \frac{|\sigma_{K,L}|}{d_{K,L}}, \qquad L \in \mathcal{T}_K, \\ (H)_{\ell(K)} &= f_K |K| + \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_h^{\mathrm{ext}}} s_K \frac{|\sigma|}{d_{K,\sigma}} g_\sigma. \end{aligned}$$

The system matrix $\mathbb{S}$ is therefore symmetric and positive definite and, moreover, irreducibly diagonally dominant (for the definition of this term, see, e.g., [44]).

3.2. The lowest-order mixed finite element method and its classical solutions. In the lowest-order Raviart–Thomas–Nédélec mixed finite element scheme (cf. [9] or [30]), one seeks simultaneously the approximations of p and $-\mathbf{S}\nabla p$. It reads: find $\mathbf{u}_h \in \mathbf{RTN}(\mathcal{T}_h)$ and $p_h \in \mathbb{P}_0(\mathcal{T}_h)$ such that

$$(\mathbf{S}^{-1}\mathbf{u}_h, \mathbf{v}_h) - (p_h, \nabla \cdot \mathbf{v}_h) = -\langle \mathbf{v}_h \cdot \mathbf{n}, g \rangle_\Gamma \qquad \forall\, \mathbf{v}_h \in \mathbf{RTN}(\mathcal{T}_h), \tag{3.5a}$$

$$-(\nabla \cdot \mathbf{u}_h, 1)_K = -f_K |K| \qquad \forall K \in \mathcal{T}_h. \tag{3.5b}$$

In a matrix notation, using the mapping ℓ from the previous section and a similar bijection $\wp : \mathcal{E}_h \to \{1, \dots, M\}$, where M is the number of sides in $\mathcal{E}_h$, so that U is composed of the fluxes of $\mathbf{u}_h$ through the sides, i.e., $(U)_{\wp(\sigma)} = \langle \mathbf{u}_h \cdot \mathbf{n}_\sigma, 1 \rangle_\sigma$ for each $\sigma \in \mathcal{E}_h$, the scheme (3.5a)–(3.5b) writes

$$\begin{pmatrix} \mathbb{A} & \mathbb{B}^t \\ \mathbb{B} & 0 \end{pmatrix} \begin{pmatrix} U \\ P \end{pmatrix} = \begin{pmatrix} F \\ G \end{pmatrix}, \tag{3.6}$$

where the matrix $\mathbb{A}$ is symmetric and positive definite. A possible approach to the solution of (3.6) consists in eliminating the unknowns U,

$$U = \mathbb{A}^{-1}(F - \mathbb{B}^t P), \tag{3.7}$$

which leaves the Schur complement equation for P,

$$\mathbb{B}\mathbb{A}^{-1}\mathbb{B}^t\, P = \mathbb{B}\mathbb{A}^{-1}F - G. \tag{3.8}$$

The Schur complement matrix $\mathbb{B}\mathbb{A}^{-1}\mathbb{B}^t$ is symmetric and positive definite. In practical computations, it is never explicitly formed. Though $\mathbb{A}$ is sparse, its inverse $\mathbb{A}^{-1}$ is typically *dense*, and, moreover, its construction would be in any case *too expensive*.

3.3. The lowest-order mixed finite element method viewed as a finite volume scheme and an alternative solution. An alternative approach to the lowest-order mixed finite element method is derived in [47]. It is shown that in the mixed finite element method, contrary to the common belief, there also exist local flux expressions. More precisely, it is shown that the mixed finite element scheme (3.5a)–(3.5b) can equivalently be written in the form (3.1), where the diffusive flux $U_{K,\sigma} = \langle \mathbf{u}_h \cdot \mathbf{n}, 1\rangle_\sigma$ through a given side σ of an element K is a function of the unknowns p_L, sources, and boundary conditions on elements L sharing a vertex with this side. As shown in [47], this diffusive flux can be obtained by solution of *local* linear systems. Precise forms of these local linear systems and conditions on their solvability are discussed in [47].

The local flux expressions of [47] lead to the expression for the unknowns U as

$$U = \tilde{\mathbb{A}}^{-1}(F - \mathbb{B}^t P) - \mathbb{J}G, \tag{3.9}$$

where the advantage in contrast to (3.7) is that the matrices $\tilde{\mathbb{A}}^{-1}$ and $\mathbb{J}$ are *sparse* and can be easily and *locally constructed*. The second line of (3.6) then yields

$$\mathbb{B}\tilde{\mathbb{A}}^{-1}\mathbb{B}^t\, P = \mathbb{B}\tilde{\mathbb{A}}^{-1}F - (\mathbb{I} + \mathbb{B}\mathbb{J})G, \tag{3.10}$$

which is of the form (3.2) with

$$\begin{aligned} \mathbb{S} &= \mathbb{B}\tilde{\mathbb{A}}^{-1}\mathbb{B}^t, \\ H &= \mathbb{B}\tilde{\mathbb{A}}^{-1}F - (\mathbb{I} + \mathbb{B}\mathbb{J})G. \end{aligned}$$

The system matrix $\mathbb{B}\tilde{\mathbb{A}}^{-1}\mathbb{B}^t$ is in most cases, in dependence on $\mathcal{T}_h$ and $\mathbf{S}$, positive definite, although in general nonsymmetric. It has a wider stencil than the system matrix $\mathbb{S}$ in (3.2) determined by (3.1), (3.3). This matrix gets symmetric (in fact the same as that of (3.2) for (3.1), (3.3)) when $\mathbf{S} = \mathbb{I}$ and when the mesh consists of equilateral simplices. A full equivalence between the systems (3.10) and (3.2) with (3.1) and (3.3) however appears only when the source term f vanishes. Summarizing, as shown in [47], the scheme (3.5a)–(3.5b) is equivalent to a particular finite volume scheme, and both methods can be written in the same form (3.1)–(3.2).

3.4. Postprocessing. The finite volume or mixed finite element solution p_h is only piecewise constant. In order to derive energy a posteriori error estimates, we first construct a postprocessed approximation $\tilde{p}_h$ which has more regularity.

Let $\mathbf{u}_h \in \mathbf{RTN}(\mathcal{T}_h)$ be given by (3.5a)–(3.5b) in the mixed finite element method, or let $\mathbf{u}_h \in \mathbf{RTN}(\mathcal{T}_h)$ be prescribed by the fluxes $U_{K,\sigma}$ in the finite volume method, i.e., on each $K \in \mathcal{T}_h$ and $\sigma \in \mathcal{E}_K$, let $\mathbf{u}_h$ be such that

$$(\mathbf{u}_h \cdot \mathbf{n})|_\sigma := U_{K,\sigma}/|\sigma|. \tag{3.11}$$

We define a postprocessed approximation $\tilde{p}_h \in \mathbb{P}_2(\mathcal{T}_h)$ on each simplex in the following way:

$$-\mathbf{S}_K \nabla \tilde{p}_h|_K = \mathbf{u}_h|_K, \qquad \forall K \in \mathcal{T}_h, \tag{3.12a}$$

$$(1-\mu_K)\frac{(\tilde{p}_h,1)_K}{|K|} + \mu_K \tilde{p}_h(\mathbf{x}_K) = p_K, \qquad \forall K \in \mathcal{T}_h. \tag{3.12b}$$

Here $\mu_K = 0$ for mixed finite elements and $\mu_K = 0$ or 1 for finite volumes, depending on whether in the particular finite volume scheme (3.1) p_K represents the approximate mean value of p_h on $K \in \mathcal{T}_h$ or the approximate point value in $\mathbf{x}_K$, respectively. It is not difficult to show that such $\tilde{p}_h$ exists, is unique, but nonconforming (does not belong to $H^1(\Omega)$), see [48, Section 4.1] and [50, Section 3.2.1]. For mixed finite elements it is shown in [48] that $\tilde{p}_h \in W_0(\mathcal{T}_h)$, i.e., $\tilde{p}_h$ has continuous means of traces on interior sides $\mathcal{E}_h^{\mathrm{int}}$. The proof is simple: let $\sigma = \sigma_{K,L} \in \mathcal{E}_h^{\mathrm{int}}$ and take (3.5a) for the basis function $\mathbf{v}_\sigma$ associated with σ, i.e., $\mathbf{v}_\sigma \in \mathbf{RTN}(\mathcal{T}_h)$ such that $\langle \mathbf{v}_\sigma \cdot \mathbf{n}_\sigma, 1\rangle_\sigma = 1$ and $\langle \mathbf{v}_\sigma \cdot \mathbf{n}_\gamma, 1\rangle_\gamma = 0$ for all $\gamma \in \mathcal{E}_h$, $\gamma \neq \sigma$. Taking into account (3.12a)–(3.12b) and the fact that $g = 0$ on all $\gamma \in \mathcal{E}_h^{\mathrm{int}}$ and using the Green theorem, (3.5a) can be written as

$$-(\nabla \tilde{p}_h, \mathbf{v}_\sigma)_{K\cup L} - (\tilde{p}_h, \nabla\cdot\mathbf{v}_\sigma)_{K\cup L} = -\langle [\![\tilde{p}_h]\!], \mathbf{v}_\sigma \cdot \mathbf{n}_\sigma\rangle_\sigma = 0. \tag{3.13}$$

On the contrary, for the finite volume scheme (3.1), (3.3) it can be shown that $\tilde{p}_h \in W_0(\mathcal{T}_h)$ only if $f = 0$.

Under the condition that the finite volume scheme at hand satisfies some convergence properties it is shown in [50] that $\nabla \tilde{p}_h \to \nabla p$ and $\tilde{p}_h \to p$ in the $L^2(\Omega)$-norm for $h \to 0$ and that optimal a priori error estimates hold. This point is obvious in the mixed finite element case by (3.12a)–(3.12b) (see [51] for more comments). Note finally that the described postprocessing is local on each element and its cost is negligible.

3.5. Oswald interpolation operator. As the finite volume/mixed finite element approximation $\tilde{p}_h$ belongs to $H^1(\mathcal{T}_h)$ only, we will need in the following its $H^1(\Omega)$-conforming interpolation. For this purpose we adopt the Oswald interpolant considered, e.g., in [1], modified in such a way that it satisfies the prescribed boundary conditions, cf. [50].

For a given function $\varphi_h \in \mathbb{P}_k(\mathcal{T}_h)$, the Oswald interpolation operator $\mathcal{I}_{\mathrm{Os}}$ from $\mathbb{P}_k(\mathcal{T}_h)$ to $\mathbb{P}_k(\mathcal{T}_h) \cap H^1(\Omega)$ is defined as follows: let $\mathbf{x}$ be a Lagrangian node, i.e., a point where the Lagrangian degree of freedom for $\mathbb{P}_k(\mathcal{T}_h) \cap H^1(\Omega)$ is prescribed, see [11, Section 2.2]. If $\mathbf{x}$ lies in the interior of some $K \in \mathcal{T}_h$ or in the interior of some boundary side, $\mathcal{I}_{\mathrm{Os}}(\varphi_h)(\mathbf{x}) = \varphi_h(\mathbf{x})$. Otherwise, the value of $\mathcal{I}_{\mathrm{Os}}(\varphi_h)$ at $\mathbf{x}$ is defined by the average of the values of φ_h at this node from the neighboring elements, i.e.,

$$\mathcal{I}_{\mathrm{Os}}(\varphi_h)(\mathbf{x}) = \frac{1}{N_{\mathbf{x}}} \sum_{K\in\mathcal{T}_{\mathbf{x}}} \varphi_h|_K(\mathbf{x}),$$

where $\mathcal{T}_{\mathbf{x}} := \{K \in \mathcal{T}_h;\, \mathbf{x} \in K\}$ is the set of elements of $\mathcal{T}_h$ containing the node $\mathbf{x}$ and $N_{\mathbf{x}}$ denotes the number of elements contained in this set. Finally, let $\mathcal{I}_{\mathrm{Os}}^{\Gamma}(\varphi_h)$ be a modified Oswald interpolate differing from $\mathcal{I}_{\mathrm{Os}}(\varphi_h)$ only on such $K \in \mathcal{T}_h$ that contain a boundary side and such that

$$\mathcal{I}_{\mathrm{Os}}^{\Gamma}(\varphi_h)|_\Gamma = g \qquad \text{in the sense of traces.}$$

4. Inexact solution of systems of linear algebraic equations. In this section we introduce some notation related to the inexactly computed solutions of the systems of linear algebraic equations arising from the considered finite volume and mixed finite element schemes. For recent information about the corresponding linear algebraic solvers, we refer to [14].

Let P^{a} be an approximate solution of (3.2), i.e., $\mathbb{S}P^{\mathrm{a}} \approx H$. We then have the equation

$$\mathbb{S}P^{\mathrm{a}} = H - R, \tag{4.1}$$

where $R := H - \mathbb{S}P^{\mathrm{a}}$ is the algebraic residual vector associated with the approximation P^{a}. This means that an approximate solution P^{a} of problem (3.2) is the exact solution of the same problem with a perturbed right-hand side $H^{\mathrm{a}} := H - R$. Similarly in the mixed finite element case, a general inexact solution means that we have only U^{a}, P^{a} which solve

$$\begin{pmatrix} \mathbb{A} & \mathbb{B}^t \\ \mathbb{B} & 0 \end{pmatrix} \begin{pmatrix} U^{\mathrm{a}} \\ P^{\mathrm{a}} \end{pmatrix} = \begin{pmatrix} F^{\mathrm{a}} \\ G^{\mathrm{a}} \end{pmatrix} \tag{4.2}$$

with some perturbed F^{a}, G^{a}.

As we will see in Sections 4.1–4.3 below, in any of the considered cases we will get from an inexact algebraic solution a couple $p_h^{\mathrm{a}} \in \mathbb{P}_0(\mathcal{T}_h)$, $\mathbf{u}_h^{\mathrm{a}} \in \mathbf{RTN}(\mathcal{T}_h)$, where $\mathbf{u}_h^{\mathrm{a}}$ is such that

$$\langle \mathbf{u}_h^{\mathrm{a}} \cdot \mathbf{n}, 1 \rangle_{\partial K} = f_K |K| - \rho_K |K| \qquad \forall K \in \mathcal{T}_h \tag{4.3}$$

for $\rho_h \in \mathbb{P}_0(\mathcal{T}_h)$ to be specified. On this basis, we can build a postprocessed approximation $\tilde{p}_h^{\mathrm{a}} \in \mathbb{P}_2(\mathcal{T}_h)$ by

$$-\mathbf{S}_K \nabla \tilde{p}_h^{\mathrm{a}}|_K = \mathbf{u}_h^{\mathrm{a}}|_K, \qquad \forall K \in \mathcal{T}_h, \tag{4.4a}$$

$$(1 - \mu_K) \frac{(\tilde{p}_h^{\mathrm{a}}, 1)_K}{|K|} + \mu_K \tilde{p}_h^{\mathrm{a}}(\mathbf{x}_K) = p_K^{\mathrm{a}}, \qquad \forall K \in \mathcal{T}_h, \tag{4.4b}$$

as in Section 3.4. The backward error idea expressed by (4.1) and (4.2), together with the construction (4.3) and (4.4), will form a basis for our a posteriori error estimates, as we will see in Section 5 below. We now give the details on the different cases.

4.1. Finite volume method. We consider here a general finite volume scheme (3.1) where the fluxes $U_{K,\sigma}$ depend (linearly) on the values of p_h, on the Dirichlet boundary conditions given by the function g, and possibly also on the source term function f. Please notice that the mixed finite element discretized system (3.9)–(3.10) is in this way also included in the expression (3.2) for the discretized system of the finite volume method.

Defining $p_h^{\mathrm{a}} \in \mathbb{P}_0(\mathcal{T}_h)$ by $p_K^{\mathrm{a}} := (P^{\mathrm{a}})_{\ell(K)}$ and a residual function $\rho_h \in \mathbb{P}_0(\mathcal{T}_h)$ associated with the algebraic residual vector R by

$$\rho_K := \frac{(R)_{\ell(K)}}{|K|}, \qquad K \in \mathcal{T}_h, \tag{4.5}$$

equation (4.1) is equivalent to the set of conservation equations

$$\sum_{\sigma \in \mathcal{E}_K} U_{K,\sigma}^{\mathrm{a}} = f_K |K| - \rho_K |K| \qquad \forall K \in \mathcal{T}_h. \tag{4.6}$$

The fluxes $U^{\mathrm{a}}_{K,\sigma}$ are of the same form as $U_{K,\sigma}$, with the values of p_h replaced by p^{a}_h. In particular, they depend on the original source term function f, more precisely on $f_K|K|$, and not on the perturbed terms $f_K|K| - \rho_K|K|$. The fluxes are again continuous on the interior sides, $U^{\mathrm{a}}_{K,\sigma} = -U^{\mathrm{a}}_{L,\sigma}$ for all $\sigma = \sigma_{K,L} \in \mathcal{E}_h^{\mathrm{int}}$. For our specific example (3.3) we in particular get

$$U^{\mathrm{a}}_{K,\sigma} = -s_{K,L}\frac{|\sigma_{K,L}|}{d_{K,L}}(p^{\mathrm{a}}_L - p^{\mathrm{a}}_K) \quad \text{for } \sigma = \sigma_{K,L} \in \mathcal{E}_h^{\mathrm{int}},$$
$$U^{\mathrm{a}}_{K,\sigma} = -s_K\frac{|\sigma|}{d_{K,\sigma}}(g_\sigma - p^{\mathrm{a}}_K) \qquad \text{for } \sigma \in \mathcal{E}_K \cap \mathcal{E}_h^{\mathrm{ext}}.$$

Compared to (3.1), equation (4.6) contains an additional term on the right-hand side representing the error from the inexact solution of the algebraic system. We can now define $\mathbf{u}^{\mathrm{a}}_h \in \mathbf{RTN}(\mathcal{T}_h)$ by

$$(\mathbf{u}^{\mathrm{a}}_h \cdot \mathbf{n})|_\sigma := U^{\mathrm{a}}_{K,\sigma}/|\sigma|, \tag{4.7}$$

so that (4.3) follows from (4.6), with ρ_h given by (4.5).

Let us finally point out that the particular choice of fluxes in (3.1) and hence the particular form of the system matrix and the right-hand side vector in (3.2) is for our further analysis not important.

4.2. Mixed finite element method without means of traces continuity. Let U^{a}, P^{a} be an inexact solution of the mixed finite element discretized system (3.6) satisfying (4.2). Defining $p^{\mathrm{a}}_h \in \mathbb{P}_0(\mathcal{T}_h)$, $\mathbf{u}^{\mathrm{a}}_h \in \mathbf{RTN}(\mathcal{T}_h)$ by

$$p^{\mathrm{a}}_K := (P^{\mathrm{a}})_{\ell(K)}, \quad (\mathbf{u}^{\mathrm{a}}_h \cdot \mathbf{n})|_\sigma := (U^{\mathrm{a}})_{\wp(\sigma)}/|\sigma|, \tag{4.8}$$

(4.3) holds with

$$\rho_K := \frac{(G^{\mathrm{a}} - G)_{\ell(K)}}{|K|}, \qquad K \in \mathcal{T}_h. \tag{4.9}$$

If we form $\tilde{p}^{\mathrm{a}}_h$ by (4.4a)–(4.4b), then $\tilde{p}^{\mathrm{a}}_h \notin W_0(\mathcal{T}_h)$, as (3.13) no more holds true because of the perturbation of the right-hand side vector F in the first line of (4.2). This is the case whenever the analogy of (3.7) for the computed quantities U^{a} and P^{a} is satisfied with F^{a} different from F, i.e., $U^{\mathrm{a}} \neq \mathbb{A}^{-1}(F - \mathbb{B}^t P^{\mathrm{a}})$. All inexact mixed Schur complement methods, which are based on the inexact solution of (3.8), fall into this category. The most classical example is the inexact Uzawa algorithm (cf. Elman and Golub [13]).

The approach of Section 4.1 for the mixed finite element discretized system (3.9)–(3.10) suffers from the same trouble. Indeed, with the notation of Section 3.3, supposing

$$\mathbb{B}\tilde{\mathbb{A}}^{-1}\mathbb{B}^t P^{\mathrm{a}} = \mathbb{B}\tilde{\mathbb{A}}^{-1}F - (\mathbb{I} + \mathbb{B}\mathbb{J})G - R^{\mathrm{MFE}} \tag{4.10}$$

for some nonzero residual vector R^{MFE}, with the fluxes $\mathbf{u}^{\mathrm{a}}_h \in \mathbf{RTN}(\mathcal{T}_h)$ subsequently constructed through

$$U^{\mathrm{a}} = \tilde{\mathbb{A}}^{-1}(F - \mathbb{B}^t P^{\mathrm{a}}) - \mathbb{J}G,$$

the couple U^{a}, P^{a} is a solution of (4.2) with F^{a} generally different from F.

4.3. Mixed finite element method with means of traces continuity. If the inexact mixed finite element solution $U^{\rm a}$, $P^{\rm a}$ satisfies

$$\begin{pmatrix} \mathbb{A} & \mathbb{B}^t \\ \mathbb{B} & 0 \end{pmatrix} \begin{pmatrix} U^{\rm a} \\ P^{\rm a} \end{pmatrix} = \begin{pmatrix} F \\ G^{\rm a} \end{pmatrix}, \tag{4.11}$$

then defining $p_h^{\rm a}$ and $\mathbf{u}_h^{\rm a}$ by (4.8) and ρ_h by (4.9) immediately gives (4.3). Moreover, reconstructing $\tilde{p}_h^{\rm a}$ using (4.4a)–(4.4b) gives the means of traces continuity $\tilde{p}_h^{\rm a} \in W_0(\mathcal{T}_h)$, as (3.13) holds true. This happens whenever the analog of (3.7) is satisfied *exactly* for the computed quantities $U^{\rm a}$ and $P^{\rm a}$, i.e., $U^{\rm a} = \mathbb{A}^{-1}(F - \mathbb{B}^t P^{\rm a})$ and $F^{\rm a} = F$. We will see later that this case is particularly interesting when proving the (local) efficiency of our estimates.

Let $\mathbb{I} + \mathbb{B}\mathbb{J}$ in (3.10) or in (4.10) be an invertible matrix. Then we can find $G^{\rm a}$ such that

$$-(\mathbb{I} + \mathbb{B}\mathbb{J})G - R^{\rm MFE} = -(\mathbb{I} + \mathbb{B}\mathbb{J})G^{\rm a}.$$

Consequently, for $P^{\rm a}$ the solution of (4.10), which gives

$$\mathbb{B}\tilde{\mathbb{A}}^{-1}\mathbb{B}^t P^{\rm a} = \mathbb{B}\tilde{\mathbb{A}}^{-1}F - (\mathbb{I} + \mathbb{B}\mathbb{J})G^{\rm a}$$

using the above relation, and for $U^{\rm a}$ determined by

$$U^{\rm a} = \tilde{\mathbb{A}}^{-1}(F - \mathbb{B}^t P^{\rm a}) - \mathbb{J}G^{\rm a},$$

the couple $U^{\rm a}$, $P^{\rm a}$ is the solution of (4.11). Consequently, $\tilde{p}_h^{\rm a}$ determined by (4.4a)–(4.4b) with $p_h^{\rm a}$ and $\mathbf{u}_h^{\rm a}$ given by (4.8) yields $\tilde{p}_h^{\rm a} \in W_0(\mathcal{T}_h)$, i.e., $\tilde{p}_h^{\rm a}$ has continuous means of traces. The purpose is to redistribute the algebraic error given by $R^{\rm MFE}$ so that (4.11) holds true.

An important point here is that using a similar approach as in [47, equations (2.4)–(2.10)], $G^{\rm a}$ can be constructed by solving only *local* problems on patches of elements sharing a vertex. Denoting this vertex by V, the resulting local system can schematically be written as

$$-(\mathbb{I}_V + \mathbb{B}_V\mathbb{J}_V)G_V - R_V^{\rm MFE} = -(\mathbb{I}_V + \mathbb{B}_V\mathbb{J}_V)G_V^{\rm a}.$$

Note also that $G^{\rm a} - G = (\mathbb{I} + \mathbb{B}\mathbb{J})^{-1}R^{\rm MFE}$.

Finally, (4.3) holds with ρ_h given by (4.9).

5. A posteriori error estimates including the algebraic error. We derive in this section a posteriori error estimates which include the algebraic error. We first recall the following result proved as a part of [48, Lemma 7.1] (here $|||\cdot|||$ is the energy norm defined by (2.1)):

LEMMA 5.1 (Abstract a posteriori error estimation framework). *Consider arbitrary $p \in H^1_\Gamma(\Omega)$ and $\tilde{p} \in H^1(\mathcal{T}_h)$. Then*

$$|||p - \tilde{p}||| \le \inf_{s \in H^1_\Gamma(\Omega)} |||\tilde{p} - s||| + \sup_{\substack{\varphi \in H^1_0(\Omega) \\ |||\varphi|||=1}} \mathcal{B}(p - \tilde{p}, \varphi).$$

Before formulating the a posteriori error estimate, we recall the Poincaré inequality. It states that for a polygon/polyhedron $K \subset \mathbb{R}^d$ and $\varphi \in H^1(K)$,

$$\|\varphi - \varphi_K\|_K^2 \le C_{{\rm P},K} h_K^2 \|\nabla\varphi\|_K^2, \tag{5.1}$$

where $\varphi_K := (\varphi,1)_K/|K|$ is the mean of φ over K. For a convex K, which is the case of simplices, the constant $C_{\mathrm{P},K}$ can be evaluated as $1/\pi^2$, cf. [29, 7]. We also point out that $\mathcal{I}^\Gamma_{\mathrm{Os}}(\tilde p_h^{\mathrm{a}})$ used below is the modified Oswald interpolant of $\tilde p_h^{\mathrm{a}}$ described in Section 3.5.

Our a posteriori error estimates are based on the following theorem:

THEOREM 5.2 (A posteriori error estimate including the algebraic error). *Let p be the weak solution of (1.1) given by (2.2) with the data satisfying Assumption 2.1. Let a couple $p_h^{\mathrm{a}} \in \mathbb{P}_0(\mathcal{T}_h)$, $\mathbf{u}_h^{\mathrm{a}} \in \mathbf{RTN}(\mathcal{T}_h)$ be given, where $\mathbf{u}_h^{\mathrm{a}}$ satisfies (4.3) for some given function $\rho_h \in \mathbb{P}_0(\mathcal{T}_h)$. Finally, let $\tilde p_h^{\mathrm{a}} \in \mathbb{P}_2(\mathcal{T}_h)$ be the postprocessed approximation given by (4.4a)–(4.4b). Then*

$$|||p-\tilde p_h^{\mathrm{a}}||| \le \eta_{\mathrm{NC}} + \eta_{\mathrm{R}} + \eta_{\mathrm{AE}}, \tag{5.2}$$

where the global nonconformity *and* residual estimators *are given by*

$$\eta_{\mathrm{NC}} := \left\{\sum_{K\in\mathcal{T}_h} \eta_{\mathrm{NC},K}^2\right\}^{\frac12} \qquad \text{and} \qquad \eta_{\mathrm{R}} := \left\{\sum_{K\in\mathcal{T}_h} \eta_{\mathrm{R},K}^2\right\}^{\frac12},$$

respectively, and η_{AE} stands for the algebraic error estimator *defined by*

$$\eta_{\mathrm{AE}} := \inf_{\substack{\mathbf{r}_h\in\mathbf{RTN}(\mathcal{T}_h)\\ \nabla\cdot\mathbf{r}_h=\rho_h}} \sup_{\substack{\varphi\in H^1_0(\Omega)\\ |||\varphi|||=1}} (\mathbf{r}_h, \nabla\varphi). \tag{5.3}$$

The local nonconformity *and* residual estimators *are respectively given by*

$$\eta_{\mathrm{NC},K} := |||\tilde p_h^{\mathrm{a}} - \mathcal{I}^\Gamma_{\mathrm{Os}}(\tilde p_h^{\mathrm{a}})|||_K, \quad \eta_{\mathrm{R},K} := \sqrt{\frac{C_{\mathrm{P},K}}{c_{\mathbf{S},K}}}\, h_K \|f - f_K\|_K.$$

Proof. For any $s \in H^1_\Gamma(\Omega)$ we have from Lemma 5.1

$$\begin{aligned} |||p-\tilde p_h^{\mathrm{a}}||| &\le |||\tilde p_h^{\mathrm{a}} - s||| + \sup_{\substack{\varphi\in H^1_0(\Omega)\\ |||\varphi|||=1}} \mathcal{B}(p-\tilde p_h^{\mathrm{a}},\varphi)\\ &= |||\tilde p_h^{\mathrm{a}} - s||| + \sup_{\substack{\varphi\in H^1_0(\Omega)\\ |||\varphi|||=1}} [T_{\mathrm{R}}(\varphi) + T_{\mathrm{AE}}(\varphi)]\\ &\le |||\tilde p_h^{\mathrm{a}} - s||| + \sup_{\substack{\varphi\in H^1_0(\Omega)\\ |||\varphi|||=1}} T_{\mathrm{R}}(\varphi) + \sup_{\substack{\varphi\in H^1_0(\Omega)\\ |||\varphi|||=1}} T_{\mathrm{AE}}(\varphi), \end{aligned} \tag{5.4}$$

where $T_{\mathrm{R}}(\varphi) := \sum_{K\in\mathcal{T}_h} (\mathbf{S}\nabla(p-\tilde p_h^{\mathrm{a}}) + \mathbf{r}_h, \nabla\varphi)_K$ and $T_{\mathrm{AE}}(\varphi) := -(\mathbf{r}_h,\nabla\varphi)$ for an arbitrary $\mathbf{r}_h \in \mathbf{RTN}(\mathcal{T}_h)$ such that $\nabla\cdot\mathbf{r}_h = \rho_h$.

The term $T_{\mathrm{R}}(\varphi)$ can be expressed using the definition of the weak solution (2.2), (4.4a), and the Green theorem as (recall that $\mathbf{r}_h$, $\mathbf{u}_h^{\mathrm{a}} \in \mathbf{H}(\mathrm{div},\Omega)$ and $\varphi \in H^1_0(\Omega)$)

$$\begin{aligned} T_{\mathrm{R}}(\varphi) &= (f,\varphi) - \sum_{K\in\mathcal{T}_h} (\mathbf{S}\nabla\tilde p_h^{\mathrm{a}} - \mathbf{r}_h, \nabla\varphi)_K\\ &= (f,\varphi) + (\mathbf{r}_h + \mathbf{u}_h^{\mathrm{a}}, \nabla\varphi) = (f - \nabla\cdot(\mathbf{r}_h + \mathbf{u}_h^{\mathrm{a}}), \varphi). \end{aligned} \tag{5.5}$$

Since the divergence is piecewise constant for functions in $\mathbf{RTN}(\mathcal{T}_h)$, the Green theorem with (4.3) gives for any $K \in \mathcal{T}_h$

$$(\nabla\cdot\mathbf{u}_h^{\mathrm{a}})|_K|K| = (\nabla\cdot\mathbf{u}_h^{\mathrm{a}}, 1)_K = \langle \mathbf{u}_h^{\mathrm{a}}\cdot\mathbf{n}, 1\rangle_{\partial K} = f_K|K| - \rho_K|K|,$$

and, consequently,

$$(\nabla \cdot \mathbf{u}_h^{\mathrm{a}})|_K = f_K - \rho_K. \tag{5.6}$$

Thus, employing $\nabla \cdot \mathbf{r}_h|_K = \rho_K$,

$$f - \nabla \cdot (\mathbf{r}_h + \mathbf{u}_h^{\mathrm{a}}) = f - \rho_K - f_K + \rho_K = f - f_K \qquad \forall K \in \mathcal{T}_h.$$

Now let $\varphi_K := (\varphi, 1)_K/|K|$ be the mean value of φ over K. Using the above identities, we can rewrite (5.5) in the form

$$T_{\mathrm{R}}(\varphi) = \sum_{K\in\mathcal{T}_h} (f - f_K, \varphi - \varphi_K)_K$$

and from the Cauchy–Schwarz inequality, the Poincaré inequality (5.1), and definition (2.1) of the energy norm, we obtain the estimate

$$T_{\mathrm{R}}(\varphi) \le \sum_{K\in\mathcal{T}_h} \|f - f_K\|_K \|\varphi - \varphi_K\|_K \le \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{R},K} |||\varphi|||_K.$$

Using the Cauchy–Schwarz inequality once again together with $|||\varphi||| = 1$,

$$T_{\mathrm{R}}(\varphi) \le \left\{ \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{R},K}^2 \right\}^{\frac{1}{2}}.$$

With (5.4), putting $s = \mathcal{I}_{\mathrm{Os}}^{\Gamma}(\tilde{p}_h^{\mathrm{a}})$ and noticing that $\mathbf{r}_h \in \mathbf{RTN}(\mathcal{T}_h)$ such that $\nabla\cdot\mathbf{r}_h = \rho_h$ was chosen arbitrarily, the proof is finished. ∎

REMARK 5.3 (Form of the a posteriori error estimate). *Remark that by* (4.4a) *and by definition* (2.1) *of the energy norm, posing* $\mathbf{u} := -\mathbf{S}\nabla p$,

$$|||p - \tilde{p}_h^{\mathrm{a}}||| = \left\|\mathbf{S}^{-\frac{1}{2}}(\mathbf{u} - \mathbf{u}_h^{\mathrm{a}})\right\|,$$

so that the a posteriori error estimate of Theorem 5.2 equivalently controls the energy error in the flux.

The a posteriori error estimate given in Theorem 5.2 consists of three parts: the nonconformity estimator η_{NC} indicating the departure of the approximate solution $\tilde{p}_h^{\mathrm{a}}$ from the space $H^1(\Omega)$, the residual estimator η_{R} which measures the interpolation error in the right-hand side of problem (1.1), and the algebraic error estimator η_{AE} which counts for the error from the inexact solution of the algebraic system. Note that the nonconformity estimator depends on the actual approximation $\tilde{p}_h^{\mathrm{a}}$ of $\tilde{p}_h$ and thus implicitly also on ρ_h and *not only on the discretization error*, whereas the algebraic error estimator depends *only on the residual function* ρ_h. We discuss computable upper bounds on η_{AE} in Section 7 below. As it will turn out, in some approaches the function $\mathbf{r}_h$ does not need to be physically constructed. We will also present an approach which is based on constructing the function $\mathbf{r}_h$ locally. Finally, the residual estimator η_{R} depends only on the data from the continuous and discrete problems and is thus independent of the algebraic error. Moreover, whenever $f \in H^1(\mathcal{T}_h)$, this estimator is clearly superconvergent by the Poincaré inequality (5.1) (it converges as $O(h^2)$ for $h \to 0$) and its value is significant only on coarse grids or for highly varying $\mathbf{S}$. We shall give some more details in the next section.

The following remark follows from the freedom of choice of s and $\mathbf{r}_h$ in the proof of Theorem 5.2:

REMARK 5.4 (Abstract form of Theorem 5.2). *With the assumptions of Theorem 5.2,*

$$|||p - \tilde{p}_h^{\mathrm{a}}||| \le \eta_{\mathrm{NC}}^{\mathrm{A}} + \eta_{\mathrm{R}} + \eta_{\mathrm{AE}}^{\mathrm{A}}$$

with

$$\eta_{\mathrm{NC}}^{\mathrm{A}} := \inf_{s \in H^1_{\Gamma}(\Omega)} |||\tilde{p}_h^{\mathrm{a}} - s|||, \quad \eta_{\mathrm{AE}}^{\mathrm{A}} := \inf_{\substack{\mathbf{r} \in \mathbf{H}(\mathrm{div},\Omega) \\ \nabla\cdot\mathbf{r} = \rho_h}} \sup_{\substack{\varphi \in H^1_0(\Omega) \\ |||\varphi|||=1}} (\mathbf{r}, \nabla\varphi), \tag{5.7}$$

and η_{R} as in Theorem 5.2. Please note that

$$\eta_{\mathrm{NC}}^{\mathrm{A}} \le \eta_{\mathrm{NC}} \quad \textit{and} \quad \eta_{\mathrm{AE}}^{\mathrm{A}} \le \eta_{\mathrm{AE}}.$$

We now show that the abstract algebraic error estimator $\eta_{\mathrm{AE}}^{\mathrm{A}}$ given above is equal to the complementary energy of the flux of the solution of the original problem (1.1) with homogeneous Dirichlet boundary condition and the right-hand side replaced by the residual function ρ_h.

THEOREM 5.5 (Equivalence of the abstract algebraic error estimator and of the minimal complementary energy). *Consider an arbitrary $\rho_h \in \mathbb{P}_0(\mathcal{T}_h)$ and $\eta_{\mathrm{AE}}^{\mathrm{A}}$ given by (5.7). Then*

$$\eta_{\mathrm{AE}}^{\mathrm{A}} = \|\mathbf{S}^{-\frac{1}{2}}\mathbf{q}\|,$$

where $\mathbf{q} \in \mathbf{H}(\mathrm{div},\Omega)$, $\nabla\cdot\mathbf{q} = \rho_h$, is the unique minimizer of the complementary energy characterized by

$$\|\mathbf{S}^{-\frac{1}{2}}\mathbf{q}\| = \min_{\substack{\mathbf{r} \in \mathbf{H}(\mathrm{div},\Omega) \\ \nabla\cdot\mathbf{r} = \rho_h}} \|\mathbf{S}^{-\frac{1}{2}}\mathbf{r}\|, \tag{5.8}$$

or, equivalently, by $\mathbf{q} = -\mathbf{S}\nabla e$, where $e \in H^1_0(\Omega)$ is the unique weak solution of

$$-\nabla\cdot(\mathbf{S}\nabla e) = \rho_h \quad \textit{in} \quad \Omega, \qquad e = 0 \quad \textit{on} \quad \Gamma, \tag{5.9}$$

i.e.,

$$\mathcal{B}(e,\varphi) = (\rho_h, \varphi) \qquad \forall \varphi \in H^1_0(\varphi).$$

Proof. Using the Cauchy–Schwarz inequality,

$$\begin{aligned}\eta_{\mathrm{AE}}^{\mathrm{A}} &= \inf_{\substack{\mathbf{r} \in \mathbf{H}(\mathrm{div},\Omega) \\ \nabla\cdot\mathbf{r} = \rho_h}} \sup_{\substack{\varphi \in H^1_0(\Omega) \\ |||\varphi|||=1}} (\mathbf{r}, \nabla\varphi) \le \sup_{\substack{\varphi \in H^1_0(\Omega) \\ |||\varphi|||=1}} (\mathbf{q}, \nabla\varphi) \\ &= \sup_{\substack{\varphi \in H^1_0(\Omega) \\ |||\varphi|||=1}} (\mathbf{S}^{-\frac{1}{2}}\mathbf{q}, \mathbf{S}^{\frac{1}{2}}\nabla\varphi) \le \sup_{\substack{\varphi \in H^1_0(\Omega) \\ |||\varphi|||=1}} \big(\|\mathbf{S}^{-\frac{1}{2}}\mathbf{q}\| |||\varphi|||\big) = \|\mathbf{S}^{-\frac{1}{2}}\mathbf{q}\|.\end{aligned}$$

Before proceeding to the converse, let us recall that the problem of finding $\mathbf{q}$ as the minimizer of the complementary energy is equivalent to the problem of finding $\mathbf{q} \in \mathbf{H}(\mathrm{div},\Omega)$, $\nabla\cdot\mathbf{q} = \rho_h$, such that

$$(\mathbf{S}^{-1}\mathbf{q}, \mathbf{v}) = 0 \qquad \forall \mathbf{v} \in \mathbf{H}(\mathrm{div},\Omega);\ \nabla\cdot\mathbf{v} = 0, \tag{5.10}$$

see, e.g., [30, Theorem 7.1.1]. Let now $\mathbf{r} \in \mathbf{H}(\mathrm{div}, \Omega)$ such that $\nabla \cdot \mathbf{r} = \rho_h$ be arbitrary. Then, by (5.10), it holds $(\mathbf{S}^{-1}\mathbf{q}, \mathbf{q} - \mathbf{r}) = 0$, and using the fact that $\mathbf{q} = -\mathbf{S}\nabla e$, we get

$$\|\mathbf{S}^{-\frac{1}{2}}\mathbf{q}\|^2 = (\mathbf{S}^{-1}\mathbf{q}, \mathbf{q}) = (\mathbf{S}^{-1}\mathbf{q}, \mathbf{q} - \mathbf{r}) + (\mathbf{S}^{-1}\mathbf{q}, \mathbf{r}) = (-\nabla e, \mathbf{r}).$$

Hence

$$\|\mathbf{S}^{-\frac{1}{2}}\mathbf{q}\| = |||e||| = \left(\mathbf{r}, \frac{-\nabla e}{|||e|||}\right) \leq \sup_{\substack{\varphi \in H^1_0(\Omega) \\ |||\varphi|||=1}} (\mathbf{r}, \nabla\varphi),$$

which concludes the proof in virtue of the fact that $\mathbf{r} \in \mathbf{H}(\mathrm{div}, \Omega)$ such that $\nabla \cdot \mathbf{r} = \rho_h$ was chosen arbitrarily. ∎

6. Stopping criterion for iterative solvers and efficiency of the a posteriori error estimate. Using the obvious requirement for the efficiency of the PDE solver which states that the discretization and algebraic errors should be in balance, we derive in this section a stopping criterion for iterative solvers used to find an approximate solution of the discretized linear algebraic systems. Using this approach, we also prove global and local efficiency of our a posteriori error estimates in the sense that we show that the estimators also represent global and local lower bounds (up to a generic constant) for the error in the energy norm. Please note that all the results presented below still hold when η_{AE} is replaced by one of its computable upper bounds presented in Section 7 below.

The stopping criterion that we propose requires the value of the algebraic error estimator to be smaller than or comparable to the nonconformity part of the bound (5.2),

$$\eta_{\mathrm{AE}} \leq \gamma\, \eta_{\mathrm{NC}} \tag{6.1}$$

for some constant γ between 0 and 1, typically close to 1. This leads to the final upper bound

$$|||p - \tilde{p}_h^{\mathrm{a}}||| \leq (1+\gamma)\eta_{\mathrm{NC}} + \eta_{\mathrm{R}}.$$

In the further construction, we will also consider the case that η_{AE} admits local expressions $\eta_{\mathrm{AE},K}$ in all elements $K \in \mathcal{T}_h$ so that

$$\eta_{\mathrm{AE}} = \left\{ \sum_{K \in \mathcal{T}_h} \eta_{\mathrm{AE},K}^2 \right\}^{\frac{1}{2}}. \tag{6.2}$$

Under this assumption we will consider also a *local* stopping criterion of the form

$$\eta_{\mathrm{AE},K} \leq \gamma_K\, \eta_{\mathrm{NC},K} \qquad \forall K \in \mathcal{T}_h \tag{6.3}$$

for some constants γ_K between 0 and 1, typically close to 1.

In the rest of this section we will investigate the finite volume method and the mixed finite element method with and without the means of traces continuity (the three different cases considered in Section 4) separately. We will employ the notation $c_{\mathbf{S},\mathfrak{T}_K} := \min_{L \in \mathfrak{T}_K} c_{\mathbf{S},L}$, which is the lower bound on the eigenvalues of the diffusion tensor $\mathbf{S}$ on the patch of elements $\mathfrak{T}_K$ (see Assumption 2.1), and we will also make use of the following assumption:

ASSUMPTION 6.1 (Shape regularity of $\mathcal{T}$). *There exists a constant $\theta_{\mathcal{T}} > 0$ such that $\min_{K \in \mathcal{T}_h} h_K/\varrho_K \leq \theta_{\mathcal{T}}$ for all $\mathcal{T}_h \in \mathcal{T}$, where ϱ_K is the diameter of the largest ball inscribed in K.*

6.1. Mixed finite element method with means of traces continuity. Throughout this section, C, $\widetilde{C}$, and $\overline{C}$ will stand for generic constants dependent on the quantities specified below, possibly different at different occurrences. The following theorem shows that estimators for the mixed finite element case considered in Section 4.3, where $\tilde{p}_h^{\mathrm{a}} \in W_0(\mathcal{T}_h)$, represent under condition (6.1) a global lower bound for the energy error, up to a term penalizing the possible violation of the Dirichlet boundary condition. More precisely, we only have this result for nonconformity and algebraic error estimators η_{NC} and η_{AE}, which is standard and sufficient, as the residual estimator η_{R} represents only data oscillation and is generally of higher order. In the case of need, it can be included as shown in Theorem 6.3 below.

THEOREM 6.2 (Global efficiency of the a posteriori error estimate when $\tilde{p}_h^{\mathrm{a}} \in W_0(\mathcal{T}_h)$). *Let the assumptions of Theorem 5.2 and Assumption 6.1 be satisfied. Let $\tilde{p}_h^{\mathrm{a}} \in W_0(\mathcal{T}_h)$ and let (6.1) hold. Then*

$$\eta_{\mathrm{NC}} + \eta_{\mathrm{AE}} \le C(1+\gamma)(|||p - \tilde{p}_h^{\mathrm{a}}||| + |||\mathcal{I}_{\mathrm{Os}}(\tilde{p}_h^{\mathrm{a}}) - \mathcal{I}_{\mathrm{Os}}^{\Gamma}(\tilde{p}_h^{\mathrm{a}})|||),$$

or, more precisely,

$$\eta_{\mathrm{NC}} + \eta_{\mathrm{AE}} \le (1+\gamma)\sqrt{2}\left\{\sum_{K\in\mathcal{T}_h}\left(C\frac{C_{\mathbf{S},K}}{c_{\mathbf{S},\mathfrak{T}_K}}|||p - \tilde{p}_h^{\mathrm{a}}|||_{\mathfrak{T}_K}^2 + 2|||\mathcal{I}_{\mathrm{Os}}(\tilde{p}_h^{\mathrm{a}}) - \mathcal{I}_{\mathrm{Os}}^{\Gamma}(\tilde{p}_h^{\mathrm{a}})|||_K^2\right)\right\}^{\frac{1}{2}},$$

where the constant C in the last inequality depends only on the space dimension d and on the shape regularity parameter $\theta_{\mathcal{T}}$.

The proof of this theorem follows by squaring and summing over all $K \in \mathcal{T}_h$ from the proof of the following theorem. The following theorem itself shows that under the condition (6.3), the derived estimates also represent *local*, possibly up to the boundary term as above, lower bounds for the error. Consequently, they are suitable for adaptive mesh refinement.

THEOREM 6.3 (Local efficiency of the a posteriori error estimate when $\tilde{p}_h^{\mathrm{a}} \in W_0(\mathcal{T}_h)$). *Let the assumptions of Theorem 5.2 and Assumption 6.1 be satisfied. Let $\tilde{p}_h^{\mathrm{a}} \in W_0(\mathcal{T}_h)$ and let (6.2) hold together with (6.3). Then, for each $K \in \mathcal{T}_h$,*

$$\eta_{\mathrm{NC},K} + \eta_{\mathrm{AE},K} \le (1+\gamma_K)\left(C\sqrt{\frac{C_{\mathbf{S},K}}{c_{\mathbf{S},\mathfrak{T}_K}}}|||p - \tilde{p}_h^{\mathrm{a}}|||_{\mathfrak{T}_K} + |||\mathcal{I}_{\mathrm{Os}}(\tilde{p}_h^{\mathrm{a}}) - \mathcal{I}_{\mathrm{Os}}^{\Gamma}(\tilde{p}_h^{\mathrm{a}})|||_K\right).$$

If, moreover, the local algebraic error estimators are given by $\eta_{\mathrm{AE},K} = \|\mathbf{S}^{-\frac{1}{2}}\mathbf{r}_h\|_K$ for some $\mathbf{r}_h$ such that $\mathbf{r}_h \in \mathbf{RTN}(\mathcal{T}_h)$, $\nabla\cdot\mathbf{r}_h = \rho_h$, then

$$\begin{aligned}\eta_{\mathrm{NC},K} + \eta_{\mathrm{R},K} + \eta_{\mathrm{AE},K} \le{}& \left(1 + C\sqrt{\frac{C_{\mathbf{S},K}}{c_{\mathbf{S},K}}}\gamma_K\right)\left(\widetilde{C}\sqrt{\frac{C_{\mathbf{S},K}}{c_{\mathbf{S},\mathfrak{T}_K}}}|||p - \tilde{p}_h^{\mathrm{a}}|||_{\mathfrak{T}_K}\right.\\ &\left. + |||\mathcal{I}_{\mathrm{Os}}(\tilde{p}_h^{\mathrm{a}}) - \mathcal{I}_{\mathrm{Os}}^{\Gamma}(\tilde{p}_h^{\mathrm{a}})|||_K\right).\end{aligned} \tag{6.4}$$

Here the constant C depends only on the space dimension d and on the shape regularity parameter $\theta_{\mathcal{T}}$, and $\widetilde{C}$ depends in addition on the polynomial degree l of f (see Assumption 2.1).

Proof. It has been proved in [48, Theorem 4.4], [50, Theorem 4.2], and [51,

Theorem 6.15] that for any piecewise polynomial function $\tilde{p}_h^{\mathrm{a}} \in \mathbb{P}_m(\mathcal{T}_h) \cap W_0(\mathcal{T}_h)$,

$$\eta_{\mathrm{NC},K} \le C \sqrt{\frac{C_{\mathbf{S},K}}{c_{\mathbf{S},\mathfrak{T}_K}}} |||p - \tilde{p}_h^{\mathrm{a}}|||_{\mathfrak{T}_K} + |||\mathcal{I}_{\mathrm{Os}}(\tilde{p}_h^{\mathrm{a}}) - \mathcal{I}_{\mathrm{Os}}^{\Gamma}(\tilde{p}_h^{\mathrm{a}})|||_K, \tag{6.5a}$$

$$\sqrt{\frac{C_{\mathrm{P},K}}{c_{\mathbf{S},K}}}\, h_K \|f + \nabla\cdot(\mathbf{S}\nabla\tilde{p}_h^{\mathrm{a}})\|_K \le \overline{C} \sqrt{\frac{C_{\mathbf{S},K}}{c_{\mathbf{S},K}}} |||p - \tilde{p}_h^{\mathrm{a}}|||_K, \tag{6.5b}$$

where the constant C depends only on the space dimension d, on the shape regularity parameter $\theta_{\mathcal{T}}$, and on the polynomial degree m of $\tilde{p}_h^{\mathrm{a}}$, and $\overline{C}$ depends in addition on the polynomial degree l of f. The first assertion of the theorem is thus an immediate consequence of (6.5a) and of (6.3). For the second one, it remains to bound $\eta_{\mathrm{R},K}$. Using $f_K = (\nabla\cdot\mathbf{u}_h^{\mathrm{a}})|_K + \rho_K$ from (5.6), $\mathbf{u}_h^{\mathrm{a}}|_K = -\mathbf{S}_K\nabla\tilde{p}_h^{\mathrm{a}}|_K$ from (4.4a), the triangle inequality, and $\nabla\cdot\mathbf{r}_h = \rho_h$, we have

$$\eta_{\mathrm{R},K} = \sqrt{\frac{C_{\mathrm{P},K}}{c_{\mathbf{S},K}}}\, h_K \|f - f_K\|_K \le \sqrt{\frac{C_{\mathrm{P},K}}{c_{\mathbf{S},K}}}\, h_K (\|f + \nabla\cdot(\mathbf{S}\nabla\tilde{p}_h^{\mathrm{a}})\|_K + \|\nabla\cdot\mathbf{r}_h\|_K).$$

Whereas the first term on the right-hand side of this inequality is bounded by (6.5b), we have, using the inverse inequality (cf. [30, Proposition 6.3.2]) and the hypotheses on $\mathbf{S}$ from Assumption 2.1,

$$\|\nabla\cdot\mathbf{r}_h\|_K \le C h_K^{-1} \|\mathbf{r}_h\|_K \le C h_K^{-1} \sqrt{C_{\mathbf{S},K}} \|\mathbf{S}^{-\frac{1}{2}}\mathbf{r}_h\|_K$$

for some constant C only depending on d and $\theta_{\mathcal{T}}$. Thus

$$\eta_{\mathrm{R},K} \le \overline{C} \sqrt{\frac{C_{\mathbf{S},K}}{c_{\mathbf{S},K}}} |||p - \tilde{p}_h^{\mathrm{a}}|||_K + C\sqrt{C_{\mathrm{P},K}} \sqrt{\frac{C_{\mathbf{S},K}}{c_{\mathbf{S},K}}} \gamma_K \eta_{\mathrm{NC},K},$$

using also (6.3). ∎

We remark that the terms $|||\mathcal{I}_{\mathrm{Os}}(\tilde{p}_h^{\mathrm{a}}) - \mathcal{I}_{\mathrm{Os}}^{\Gamma}(\tilde{p}_h^{\mathrm{a}})|||_K$ in the above theorems penalize the possible violation of the Dirichlet boundary condition and they can be nonzero only for boundary simplices. They will disappear completely for, e.g., $g = 0$. Bound (6.4) of Theorem 6.3 is in particular relevant to the cases investigated in Sections 7.2 and 7.3 below, where the algebraic error estimators admit the desired form.

6.2. Finite volume method and mixed finite element method without means of traces continuity. When $\tilde{p}_h^{\mathrm{a}}$ is not contained in $W_0(\mathcal{T}_h)$, which is the case of Sections 4.1 and 4.2, (6.5b) still holds true but (6.5a) does not. In order to overcome this difficulty, one has to add to the right-hand side of (6.5a) the term

$$\widetilde{C}\sqrt{C_{\mathbf{S},K}} \sum_{\sigma\in\mathfrak{E}_K} \|p - \tilde{p}_h^{\mathrm{a}}\|_{\#,\sigma}\,,$$

where, for $\sigma = \sigma_{L,M} \in \mathcal{E}_h^{\mathrm{int}}$ and $\varphi \in H^1(\mathcal{T}_h)$,

$$\|\varphi\|_{\#,\sigma} := h_\sigma^{-\frac{1}{2}} \|\langle[\![\varphi]\!], 1\rangle_\sigma |\sigma|^{-1}\|_\sigma.$$

Note that $\langle[\![p]\!], 1\rangle_\sigma = 0$ for all $\sigma \in \mathcal{E}_h^{\mathrm{int}}$ and that $\langle[\![\tilde{p}_h^{\mathrm{a}}]\!], 1\rangle_\sigma = 0$ for all $\sigma \in \mathcal{E}_h^{\mathrm{int}}$ for $\tilde{p}_h^{\mathrm{a}}$ contained in $W_0(\mathcal{T}_h)$.

Let for a given set of sides $\mathcal{E}$

$$|||\varphi|||^2_{\#,\mathcal{E}} := c_{\mathbf{S},\mathcal{E}} \sum_{\sigma\in\mathcal{E}} \|\varphi\|^2_{\#,\sigma}\,,$$

where $c_{\mathbf{S},\mathcal{E}}$ is the minimum of the values $c_{\mathbf{S},K}$ over all $K \in \mathcal{T}_h$ which have at least one side in the set $\mathcal{E}$. With this definition we have the following counterpart of Theorem 6.2:

THEOREM 6.4 (Global efficiency of the a posteriori error estimate when $\tilde{p}_h^{\mathrm{a}} \notin W_0(\mathcal{T}_h)$). *Let the assumptions of Theorem 5.2 and Assumption 6.1 be satisfied. Let (6.1) hold. Then*

$$\eta_{\mathrm{NC}} + \eta_{\mathrm{AE}} \le C(1+\gamma)(|||p-\tilde{p}_h^{\mathrm{a}}||| + |||p-\tilde{p}_h^{\mathrm{a}}|||_{\#,\mathcal{E}_h^{\mathrm{int}}} + |||\mathcal{I}_{\mathrm{Os}}(\tilde{p}_h^{\mathrm{a}}) - \mathcal{I}^{\Gamma}_{\mathrm{Os}}(\tilde{p}_h^{\mathrm{a}})|||),$$

or, more precisely,

$$\eta_{\mathrm{NC}} + \eta_{\mathrm{AE}} \le (1+\gamma)\sqrt{2}\left\{\sum_{K\in\mathcal{T}_h}\left(C\frac{C_{\mathbf{S},K}}{c_{\mathbf{S},\mathfrak{T}_K}}\left(|||p-\tilde{p}_h^{\mathrm{a}}|||^2_{\mathfrak{T}_K} + |||p-\tilde{p}_h^{\mathrm{a}}|||^2_{\#,\mathfrak{E}_K}\right)\right.\right.$$
$$\left.\left. +2|||\mathcal{I}_{\mathrm{Os}}(\tilde{p}_h^{\mathrm{a}}) - \mathcal{I}^{\Gamma}_{\mathrm{Os}}(\tilde{p}_h^{\mathrm{a}})|||^2_K\right)\right\}^{\frac{1}{2}},$$

where the constant C in the last inequality depends only on the space dimension d and on the shape regularity parameter $\theta_{\mathcal{T}}$.

The proof of this statement follows from the proof of the following theorem. The following theorem itself states a local efficiency and its proof is fully analogous to Theorem 6.3. Please note that the bounds on the residual estimator $\eta_{\mathrm{R},K}$ could also be added as in (6.4):

THEOREM 6.5 (Local efficiency of the a posteriori error estimate when $\tilde{p}_h^{\mathrm{a}} \notin W_0(\mathcal{T}_h)$). *Let the assumptions of Theorem 5.2 and Assumption 6.1 be satisfied. Let (6.2) hold together with (6.3). Then, for each $K \in \mathcal{T}_h$,*

$$\eta_{\mathrm{NC},K} + \eta_{\mathrm{AE},K} \le (1+\gamma_K)\left(C\sqrt{\frac{C_{\mathbf{S},K}}{c_{\mathbf{S},\mathfrak{T}_K}}}\left(|||p-\tilde{p}_h^{\mathrm{a}}|||_{\mathfrak{T}_K} + |||p-\tilde{p}_h^{\mathrm{a}}|||_{\#,\mathfrak{E}_K}\right)\right.$$
$$\left. +|||\mathcal{I}_{\mathrm{Os}}(\tilde{p}_h^{\mathrm{a}}) - \mathcal{I}^{\Gamma}_{\mathrm{Os}}(\tilde{p}_h^{\mathrm{a}})|||_K\right),$$

where the constant C depends only on the space dimension d and on the shape regularity parameter $\theta_{\mathcal{T}}$.

We conclude this section by the following remark:

REMARK 6.6 (Both-sided estimates in the same norm when $\tilde{p}_h^{\mathrm{a}} \notin W_0(\mathcal{T}_h)$). *In Theorem 6.4 the lower bound is not given for the energy error $|||p-\tilde{p}_h^{\mathrm{a}}|||$, as is the case of the upper bound of Theorem 5.2, but for its augmented version $|||p-\tilde{p}_h^{\mathrm{a}}||| + |||p-\tilde{p}_h^{\mathrm{a}}|||_{\#,\mathcal{E}_h^{\mathrm{int}}}$. In order to give both-sided estimates in the same norm, it is sufficient to notice that $|||p-\tilde{p}_h^{\mathrm{a}}|||_{\#,\mathcal{E}_h^{\mathrm{int}}} = |||\tilde{p}_h^{\mathrm{a}}|||_{\#,\mathcal{E}_h^{\mathrm{int}}}$ and to estimate the error in $|||p-\tilde{p}_h^{\mathrm{a}}||| + |||p-\tilde{p}_h^{\mathrm{a}}|||_{\#,\mathcal{E}_h^{\mathrm{int}}}$ by $\eta_{\mathrm{NC}} + \eta_{\mathrm{R}} + \eta_{\mathrm{AE}} + |||\tilde{p}_h^{\mathrm{a}}|||_{\#,\mathcal{E}_h^{\mathrm{int}}}$. Similar technique applies to Theorem 6.5. For some remarks on such approaches, we refer to [15, Section 6].*

6.3. An alternative algebraic error estimator for the mixed finite element method without means of traces continuity. When considering solution of the linear algebraic system arising from the mixed finite element discretization such that there is no means of traces continuity of $\tilde{p}_h^{\mathrm{a}}$, see Section 4.2, the nonconformity estimator $\eta_{\mathrm{NC},K}$ in fact contains a part of the algebraic error corresponding to the fact that the first equation in (4.2) is not satisfied with the right-hand side F but only with some perturbation F^{a}. It is clear that $F - F^{\mathrm{a}}$ reflects the corresponding algebraic error and it gives an indication on how much the means of traces of $\tilde{p}_h^{\mathrm{a}}$ are discontinuous. We could prescribe $\tilde{p}_h^{\#} \in \mathbb{P}_d(\mathcal{T}_h) \cap W_0(\mathcal{T}_h)$ (recall that d is the space dimension) for each $K \in \mathcal{T}_h$ at each Lagrangian node of $\mathbb{P}_d(K)$, except of those lying at the barycentres of the sides, by the value of $\tilde{p}_h^{\mathrm{a}}$ at this node. The values at the barycentres of the sides could then be established so that the mean value of $\tilde{p}_h^{\#}$ over a side $\sigma_{K,L} \in \mathcal{E}_h^{\mathrm{int}}$ is given by the arithmetic average of the mean value of $\tilde{p}_h^{\mathrm{a}}|_K$ and that of $\tilde{p}_h^{\mathrm{a}}|_L$ on $\sigma_{K,L}$. The reason for using $\mathbb{P}_3(K)$ in three space dimensions is that the space $\mathbb{P}_2(K)$ in this case does not have Lagrangian nodes at the side barycentres, cf., e.g., [11, Section 2.2]. Considering $s = \mathcal{I}_{\mathrm{Os}}^{\Gamma}(\tilde{p}_h^{\#})$ in Theorem 5.2 and using the triangle inequality,

$$|||\tilde{p}_h^{\mathrm{a}} - s||| \le |||\tilde{p}_h^{\mathrm{a}} - \tilde{p}_h^{\#}||| + |||\tilde{p}_h^{\#} - \mathcal{I}_{\mathrm{Os}}^{\Gamma}(\tilde{p}_h^{\#})|||.$$

Consequently, we could replace (5.2) by

$$|||p - \tilde{p}_h^{\mathrm{a}}||| \le \eta_{\mathrm{NC}}^{\#} + \eta_{\mathrm{R}} + \eta_{\mathrm{AE}} + \eta_{\mathrm{AE}}^{\#},$$

where

$$\eta_{\mathrm{NC}}^{\#} := |||\tilde{p}_h^{\#} - \mathcal{I}_{\mathrm{Os}}^{\Gamma}(\tilde{p}_h^{\#})|||, \quad \eta_{\mathrm{AE}}^{\#} := |||\tilde{p}_h^{\mathrm{a}} - \tilde{p}_h^{\#}|||.$$

Then the stopping criterion (6.1) for the algebraic solver would be modified to

$$\eta_{\mathrm{AE}} + \eta_{\mathrm{AE}}^{\#} \le \gamma\, \eta_{\mathrm{NC}}^{\#}.$$

Unlike in the previous cases, this approach takes into account the algebraic error in the first equation of (4.2) separately.

7. Evaluation of the algebraic error estimator. The algebraic error estimator η_{AE} of Section 5 was defined in a general way without specification of the techniques for computing it. In this section we discuss three different approaches giving a computable upper bound on η_{AE}.

First, let us define, for the function $\rho_h \in \mathbb{P}_0(\mathcal{T}_h)$ from (4.3) whose construction is specified in Sections 4.1–4.3, a vector R by

$$(R)_{\ell(K)} := \rho_K |K|, \qquad K \in \mathcal{T}_h. \tag{7.1}$$

Note that the above vector R is the residual vector R from the finite volume method of Section 4.1, whereas $R = G^{\mathrm{a}} - G$ in the mixed finite element method of Sections 4.2 and 4.3.

7.1. Simple bound using the algebraic residual vector. A simple bound on the abstract algebraic error estimator η_{AE} which does not need an explicit construction of a vector function $\mathbf{r}_h$ such that $\mathbf{r}_h \in \mathbf{RTN}(\mathcal{T}_h)$, $\nabla \cdot \mathbf{r}_h = \rho_h$, can be based directly on the residual function ρ_h, or, equivalently, on the vector R of (7.1). The Euclidean

norm of R still prevails as a measure of the error and as a stopping criterion in iterative methods in practical computations, though it is well-known that it can be misleading, see Section 1. The following theorem shows that in order to get a guaranteed upper bound on the algebraic error, *an appropriately weighted* Euclidean norm of R has to be used. The weights depend on the element measures, the diameter of Ω, and the minimum of $c_{\mathbf{S},K}$ over all $K \in \mathcal{T}_h$. They cause, due to the fact that they do not fully reflect the correlation between the data in the problem setting, a huge overestimation of the error. This is true in particular when adaptive mesh refinement is applied and/or when the values of the coefficients exhibit significant variations. For a supportive simple algebraic reasoning see, e.g., [21, Section 17.5]. Illustrative numerical experiments are presented in Section 8 below. The result is summarized in the following lemma:

LEMMA 7.1 (Algebraic error estimator using the algebraic residual vector). *The algebraic error estimator* η_{AE} *from Theorem 5.2 can be bounded as*

$$\eta_{\mathrm{AE}} \le \eta_{\mathrm{AE}}^{(1)} := \sqrt{\frac{C_{\mathrm{F},\Omega}}{c_{\mathbf{S},\Omega}}}\, h_\Omega \left\{ \sum_{K\in\mathcal{T}_h} \rho_K^2 |K| \right\}^{\frac{1}{2}}, \tag{7.2}$$

where $c_{\mathbf{S},\Omega} := \min_{K\in\mathcal{T}_h} c_{\mathbf{S},K}$ *and* h_Ω *is the diameter of the domain* Ω.

Proof. Using the Green theorem and the Cauchy–Schwarz inequality, we have

$$\begin{aligned} (\mathbf{r}_h, \nabla\varphi) = -(\nabla\cdot\mathbf{r}_h, \varphi) &= -\sum_{K\in\mathcal{T}_h} (\rho_K, \varphi)_K \\ &\le \sum_{K\in\mathcal{T}_h} \rho_K |K|^{\frac{1}{2}} \|\varphi\|_K \le \left\{ \sum_{K\in\mathcal{T}_h} \rho_K^2 |K| \right\}^{\frac{1}{2}} \|\varphi\|. \end{aligned} \tag{7.3}$$

As $\varphi \in H_0^1(\Omega)$, we can now relate the norm $\|\varphi\|$ to the energy norm $|||\varphi|||$ using the Friedrichs inequality by

$$\|\varphi\| \le \sqrt{C_{\mathrm{F},\Omega}}\, h_\Omega \|\nabla\varphi\| \le \sqrt{\frac{C_{\mathrm{F},\Omega}}{c_{\mathbf{S},\Omega}}}\, h_\Omega |||\varphi|||\,. \tag{7.4}$$

Considering $|||\varphi||| = 1$ and combining (7.3) and (7.4) proves the statement. As for the value of $C_{\mathrm{F},\Omega}$, we refer to, e.g., Nečas [27, Section 1.2] or Rektorys [31, Chapter 30]. In dependence on the domain Ω, it in general ranges between 1 and $1/\pi^2$. Note as well that h_Ω may be replaced by the infimum over the thicknesses of Ω in the given direction, cf., e.g., [46]. ∎

We point out that (7.2) can be rewritten in the algebraic form as

$$\eta_{\mathrm{AE}}^{(1)} = C\sqrt{R^t \mathbb{D}^{-1} R} = C\|R\|_{\mathbb{D}^{-1}},$$

where $\mathbb{D} := \mathrm{diag}(|\ell^{-1}(k)|)_{k=1}^N$ is a finite volume-type mass matrix, ℓ represents the enumeration of elements in $\mathcal{T}_h$ defined in Section 3.1, and the constant C is given by $C := (C_{\mathrm{F},\Omega}/c_{\mathbf{S},\Omega})^{1/2} h_\Omega$.

7.2. Bound using the minimal discrete complementary energy and its equivalence to the algebraic energy error of mixed finite elements. In this section we relate the algebraic error estimator η_{AE} from Theorem 5.2 to the minimal discrete complementary energy, giving a discrete analogue of Theorem 5.5. The value

of the minimal discrete complementary energy will turn out to be given by the lowest-order mixed finite element approximation of problem (5.9), independently of which method was used to approximate the original problem (1.1). Finally, we will show also the equivalence of the minimal discrete complementary energy to the algebraic energy error induced by the Schur complement matrix $\mathbb{B}\mathbb{A}^{-1}\mathbb{B}^t$ in the mixed finite element method of Section 4.3.

Consider the lowest-order mixed finite element approximation of (5.9). It consists in finding $\mathbf{q}_h \in \mathbf{RTN}(\mathcal{T}_h)$ and $e_h \in \mathbb{P}_0(\mathcal{T}_h)$ such that

$$(\mathbf{S}^{-1}\mathbf{q}_h, \mathbf{v}_h) - (e_h, \nabla\cdot\mathbf{v}_h) = 0 \qquad \forall\, \mathbf{v}_h \in \mathbf{RTN}(\mathcal{T}_h), \tag{7.5a}$$

$$-(\nabla\cdot\mathbf{q}_h, 1)_K = -\rho_K|K| \qquad \forall K \in \mathcal{T}_h, \tag{7.5b}$$

and, in matrix form, to

$$\begin{pmatrix} \mathbb{A} & \mathbb{B}^t \\ \mathbb{B} & 0 \end{pmatrix}\begin{pmatrix} Q \\ E \end{pmatrix} = \begin{pmatrix} 0 \\ -R \end{pmatrix}. \tag{7.6}$$

Recall that $\rho_h \in \mathbb{P}_0(\mathcal{T}_h)$ is the function from (4.3) and R is prescribed by (7.1). From (7.6) we also get

$$\mathbb{S}E = R, \tag{7.7}$$

where $\mathbb{S} = \mathbb{B}\mathbb{A}^{-1}\mathbb{B}^t$ is the Schur complement matrix. Recall that the matrix $\mathbb{S}$ is symmetric and positive definite, so that it induces an algebraic energy norm by $\|X\|_{\mathbb{S}}^2 := X^t\mathbb{S}X$.

The following theorem relates the algebraic error estimator η_{AE}, the discrete complementary energy of $\mathbf{q}_h$, and the quantity $\|E\|_{\mathbb{S}}$. Its importance is rather theoretical, since the bound is not computable without solving (7.5a)–(7.5b).

THEOREM 7.2 (Algebraic error estimator using the minimal discrete complementary energy). *Let $\mathbf{q}_h$ be the solution of (7.5a)–(7.5b). Then the algebraic error estimator η_{AE} from Theorem 5.2 can be bounded by*

$$\eta_{\mathrm{AE}} \le \eta_{\mathrm{AE}}^{(2)} := \|\mathbf{S}^{-\frac{1}{2}}\mathbf{q}_h\|.$$

Moreover, this bound is equal to the minimal discrete complementary energy, i.e.,

$$\eta_{\mathrm{AE}}^{(2)} = \inf_{\substack{\mathbf{r}_h\in\mathbf{RTN}(\mathcal{T}_h)\\ \nabla\cdot\mathbf{r}_h=\rho_h}} \|\mathbf{S}^{-\frac{1}{2}}\mathbf{r}_h\|. \tag{7.8}$$

Finally, this bound admits an equivalent form

$$\eta_{\mathrm{AE}}^{(2)} = \|E\|_{\mathbb{S}}.$$

Proof. The first statement follows directly from definition (5.3) of η_{AE} and the Cauchy–Schwarz inequality. For the second one, note that $\nabla\cdot\mathbf{q}_h = \rho_h$ by (7.5b) and that $(\mathbf{S}^{-1}\mathbf{q}_h, \mathbf{v}_h) = 0$ for all $\mathbf{v}_h \in \mathbf{RTN}(\mathcal{T}_h)$ such that $\nabla\cdot\mathbf{v}_h = 0$ by (7.5a), which is equivalent to (7.8) as in the proof of Theorem 5.5. Finally,

$$\|E\|_{\mathbb{S}}^2 = E^t\mathbb{S}E = E^tR = \sum_{K\in\mathcal{T}_h} e_K\rho_K|K|,$$

and, putting $\mathbf{v}_h = \mathbf{q}_h$ in (7.5a) and using (7.5b),

$$\|\mathbf{S}^{-\frac{1}{2}}\mathbf{q}_h\|^2 = (\mathbf{S}^{-1}\mathbf{q}_h, \mathbf{q}_h) = (e_h, \nabla\cdot\mathbf{q}_h) = \sum_{K\in\mathcal{T}_h} e_K \rho_K |K|,$$

which proves the third statement and finishes the proof. □

REMARK 7.3 (Equivalence of $\eta_{\mathrm{AE}}^{(2)}$ to the algebraic energy error of the mixed finite element method with means of traces continuity). *Consider the mixed finite element problem* (3.6)*, with the unknowns* U*,* P*, and its inexact version* (4.11)*, with the unknowns* U^{a}*,* P^{a} *(please recall that* (4.11) *is equal to* (4.2) *with* $F = F^{\mathrm{a}}$*). Denote the error in the dual variable by* $Q := U - U^{\mathrm{a}}$ *and the error in the primal variable by* $E := P - P^{\mathrm{a}}$*. Then it is obvious to see that system* (7.6) *can now be obtained by subtracting the problems* (3.6) *and* (4.11)*. Hence* (7.7) *represents in this case the error equation of the mixed finite element method. Note that it is important that the inexact solution leads to* (4.11)*, see the discussion in Section 4. Consequently, the last statement of Theorem 7.2 says that* $\eta_{\mathrm{AE}}^{(2)}$ *is nothing but the algebraic energy error of the mixed finite element method induced be the Schur complement matrix* $\mathbb{S}$*.*

REMARK 7.4 (Guaranteed upper bound on the algebraic energy error for the mixed finite element method with means of traces continuity). *We remark that Theorem 7.2 and Remark 7.3, in combination with the results of Section 7.3 below, give a fully and easily computable upper bound not only for the algebraic error estimator* $\eta_{\mathrm{AE}}^{(2)}$ *but, at the same time, also for the algebraic energy error in the mixed finite element method when* $F = F^{\mathrm{a}}$*.*

REMARK 7.5 (Relation of $\eta_{\mathrm{AE}}^{(2)}$ to the algebraic energy error of the finite volume method). *Consider the approximation of* (5.9) *by the finite volume scheme given in Section 3.1. It consists in finding* $e_h \in \mathbb{P}_0(\mathcal{T}_h)$ *such that*

$$\sum_{\sigma\in\mathcal{E}_K} U_{K,\sigma} = \rho_K|K| \qquad \forall K \in \mathcal{T}_h, \tag{7.9}$$

where $U_{K,\sigma}$ *are the prescribed fluxes (which depend linearly on the values of* e_h*). In matrix form, this leads to*

$$\mathbb{S}E = R, \tag{7.10}$$

where $\mathbb{S}$ *is the matrix from* (3.2)*. In the above equations, once again,* $\rho_h \in \mathbb{P}_0(\mathcal{T}_h)$ *is the function from* (4.3) *and* R *is prescribed by* (7.1)*. As in the mixed finite element case, the matrix* $\mathbb{S}$ *is symmetric and positive definite, so that it induces an algebraic energy norm* $\|\cdot\|_{\mathbb{S}}$*. We now shed some light on the relationship between* η_{AE} *and* $\|E\|_{\mathbb{S}}$*.*

Let us construct a postprocessed error $\tilde{e}_h \in \mathbb{P}_2(\mathcal{T}_h)$ *from* e_h *as described in Section 3.4 and put* $\mathbf{q}_h := -\mathbf{S}\nabla\tilde{e}_h$*. Then* $\mathbf{q}_h \in \mathbf{RTN}(\mathcal{T}_h)$ *and* $\nabla\cdot\mathbf{q}_h = \rho_h$ *by* (7.9)*, so that*

$$\eta_{\mathrm{AE}} \le \|\mathbf{S}^{-\frac{1}{2}}\mathbf{q}_h\|$$

follows directly from definition (5.3) *of* η_{AE} *and the Cauchy–Schwarz inequality. Sup-*

pose for the moment that $\tilde{e}_h \in W_0(\mathcal{T}_h)$. *Under this condition,*

$$\begin{aligned}\|\mathbf{S}^{-\frac{1}{2}}\mathbf{q}_h\|^2 &= \sum_{K\in\mathcal{T}_h} (\mathbf{S}\nabla\tilde{e}_h, \nabla\tilde{e}_h)_K \\ &= \sum_{K\in\mathcal{T}_h} \left\{(-\nabla\cdot(\mathbf{S}\nabla\tilde{e}_h), \tilde{e}_h)_K + \langle \mathbf{S}\nabla\tilde{e}_h|_K\cdot\mathbf{n}, \tilde{e}_h\rangle_{\partial K}\right\} \\ &= \sum_{K\in\mathcal{T}_h} (-\nabla\cdot(\mathbf{S}\nabla\tilde{e}_h), \tilde{e}_h)_K = \sum_{K\in\mathcal{T}_h} e_K\rho_K|K| = \|E\|_{\mathbb{S}}^2\end{aligned} \tag{7.11}$$

by the Green theorem. Note that the term $\sum_{K\in\mathcal{T}_h}\langle \mathbf{S}\nabla\tilde{e}_h|_K\cdot\mathbf{n}, \tilde{e}_h\rangle_{\partial K}$ *is equal to zero by the facts that* $\mathbf{S}\nabla\tilde{e}_h\cdot\mathbf{n}$ *is sidewise constant as* $\mathbf{S}\nabla\tilde{e}_h \in \mathbf{RTN}(\mathcal{T}_h)$ *and that* $\tilde{e}_h \in W_0(\mathcal{T}_h)$. *Unfortunately, as discussed in Section 3.4,* $\tilde{e}_h$ *does not in general belong to the space* $W_0(\mathcal{T}_h)$, *although numerical experiments show that the violation of the means of traces continuity is only very slight. Thus* (7.11) *holds only approximately in the finite volume case. As however demonstrated in Section 8 below, the approximate bound*

$$\eta_{\mathrm{AE}} \lesssim \|E\|_{\mathbb{S}},$$

in combination with the following remark gives a powerful a posteriori algebraic error estimate and an excellent stopping criterion.

REMARK 7.6 (Approximate upper bound on the algebraic energy error in the conjugate gradient method). *Let a system of the form* (3.2) *be given, let* P^{a} *be its approximate solution by the conjugate gradient method [20] so that* (4.1) *holds, and let* $E := P - P^{\mathrm{a}}$ *so that* (7.10) *holds. Then the algebraic energy error* $\|E\|_{\mathbb{S}}$ *can be estimated using the techniques from [41, 42]. Though the estimates from [41, 42] do not give guaranteed upper bounds on* $\|E\|_{\mathbb{S}}$, *numerical evidence shows that they can be very useful in practical computations.*

In particular, if we consider the conjugate gradient method for solving a system of the form (3.2), *then, for the algebraic energy error at the* n*-th iteration* $E_n^{\mathrm{CG}} := P - P_n^{\mathrm{CG}}$, *it holds, considering* ν *additional conjugate gradients iterations,*

$$\|E_n^{\mathrm{CG}}\|_{\mathbb{S}}^2 = \sum_{j=n}^{n+\nu} \mu_j^{\mathrm{CG}}\|R_j^{\mathrm{CG}}\|^2 + \|P - P_{n+\nu}^{\mathrm{CG}}\|_{\mathbb{S}}^2, \tag{7.12a}$$

$$\|E_n^{\mathrm{CG}}\|_{\mathbb{S}}^2 \approx \hat{\eta}_{\mathrm{AE}}^{(2)} := \sum_{j=n}^{n+\nu} \mu_j^{\mathrm{CG}}\|R_j^{\mathrm{CG}}\|^2. \tag{7.12b}$$

Here P_j^{CG} *stands for the approximate solution of* (3.2) *computed in the* j*-th iteration step of the conjugate gradient method and* R_j^{CG} *is the corresponding algebraic residual,* $R_j^{\mathrm{CG}} := H - \mathbb{S}P_j^{\mathrm{CG}}$. *The inaccuracy of the approximation of* (7.12a) *by* (7.12b) *is given by the squared size of the algebraic energy error at the* $(n+\nu)$*-th step. The values* μ_j^{CG} *and* $\|R_j^{\mathrm{CG}}\|^2$ *are available from the conjugate gradient iterations, see [41, 42], and also the summary in [26, Section 5.3]. Please notice that here we need* ν *additional iterations of the CG method.*

For the relationship to the classical results on moments and Gauss–Christoffel quadrature as well as for other related approaches for estimation of errors we refer to [17, 18, 19, 25, 26, 39] and to the references therein.

7.3. Explicit construction of $\mathbf{r}_h$. In this section we present an approach which is based on an explicit construction of the vector function $\mathbf{r}_h$ in the algebraic error estimator η_{AE} from Theorem 5.2. First, the following corollary is an immediate consequence of Theorem 7.2:

COROLLARY 7.7 (Algebraic error estimator based on an explicitly constructed $\mathbf{r}_h$). *Consider an arbitrary $\mathbf{r}_h \in \mathbf{RTN}(\mathcal{T}_h)$ such that $\nabla \cdot \mathbf{r}_h = \rho_h$. Then the algebraic error estimator η_{AE} from Theorem 5.2 and $\eta_{\mathrm{AE}}^{(2)}$ from Theorem 7.2 can be bounded from above by*

$$\eta_{\mathrm{AE}} \le \eta_{\mathrm{AE}}^{(2)} \le \eta_{\mathrm{AE}}^{(3)}(\mathbf{r}_h) := \|\mathbf{S}^{-\frac{1}{2}}\mathbf{r}_h\|.$$

Note that the fact that $\eta_{\mathrm{AE}} \le \eta_{\mathrm{AE}}^{(3)}(\mathbf{r}_h)$ follows immediately from (5.3) by the Cauchy–Schwarz inequality.

We now present a simple algorithm with a linear complexity in the number of mesh elements which finds a convenient function $\mathbf{r}_h$, without having to solve (7.5a)–(7.5b) or any other global problem. The first step is to find an enumeration of the elements of $\mathcal{T}_h$ such that for each K_i, there is a side $\sigma \in \mathcal{E}_{K_i}$ which does not lie on the boundary of $\cup_{j=1}^{i-1} K_j$. Such an enumeration of the elements of $\mathcal{T}_h$ can be always found for meshes consisting of simplices using, e.g., the standard depth-first search in the graph associated with the partition $\mathcal{T}_h$. The algorithm is described as follows: set $\mathcal{T} := \mathcal{T}_h$, $i := N$, and while $i \ge 2$ do:

1. find $K \in \mathcal{T}$ such that there is a side $\sigma \in K$ which lies on the boundary of $\mathcal{T}$;
2. set $K_i := K$, $\mathcal{T} := \mathcal{T} \setminus K$, $i := i - 1$.

Finally denote as K_1 the last element.

With such an enumeration, we find $\mathbf{r}_h$ separately on each element of $\mathcal{T}_h$ while proceeding sequentially for $i = 1, 2, \dots, N$ in the following steps:

1. find $\mathbf{r}_i \in \mathbf{RTN}(K_i)$ such that

$$\mathbf{r}_i = \arg \min_{\widetilde{\mathbf{r}} \in \widetilde{\mathbf{RTN}}(K_i)} \|\mathbf{S}^{-\frac{1}{2}}\widetilde{\mathbf{r}}\|_{K_i},$$

where $\widetilde{\mathbf{RTN}}(K_i)$ are functions of $\mathbf{RTN}(K_i)$ such that

$$\nabla \cdot \widetilde{\mathbf{r}}_i = \rho_{K_i}, \quad \widetilde{\mathbf{r}}_i \cdot \mathbf{n}_\sigma = \mathbf{r}_h \cdot \mathbf{n}_\sigma \text{ on all } \sigma \in \mathcal{E}_{K_i} \cap \mathcal{E}_{K_j},\ j < i;$$

2. set $\mathbf{r}_h|_{K_i} := \mathbf{r}_i$.

The $\mathbf{r}_h$ constructed in this way is not optimal (it replaces the global minimization (7.8) by a local one), but, as shown in the experiments, it is a good candidate for giving a useful estimate.

8. Numerical experiments. In this section we illustrate on model problems with both homogeneous and inhomogeneous diffusion tensors that our a posteriori error estimates give tight overall error bounds and reliable stopping criteria for iterative solvers. We will consider two examples.

EXAMPLE 8.1 (Laplace equation). *We consider the Laplace equation $-\Delta p = 0$ in $\Omega = (-1, 1) \times (-1, 1)$, i.e., $\mathbf{S} = \mathbb{I}$ and $f = 0$ in (1.1). Let*

$$p(x, y) = \exp\left(\frac{x}{10}\right) \cos\left(\frac{y}{10}\right)$$

and let g in (1.1) be defined by the values of this p on the boundary Γ of Ω. Then p is the (weak as well as classical) solution of problem (1.1).

EXAMPLE 8.2 (Problem with an inhomogeneous diffusion tensor). *We consider the diffusion equation $-\nabla\cdot(\mathbf{S}\nabla p)=0$ and suppose that $\Omega=(-1,1)\times(-1,1)$ is divided into four subdomains Ω_i corresponding to the axis quadrants (the first quadrant $\{(x,y)\in\mathbb{R}^2;\ x>0,\ y>0\}\cap\Omega$ is denoted by Ω_1 and the subsequent numbering is done counterclockwise). Let $\mathbf{S}$ be piecewise constant and equal to $s_i\mathbb{I}$ in Ω_i. We consider two choices of the diffusion tensor $\mathbf{S}$, listed in Table 8.1. Then with the coefficients α, a_i, and b_i, also listed in Table 8.1, and with the Dirichlet boundary condition imposed accordingly, the analytical solution in each subdomain Ω_i has in polar coordinates (ϱ,ϑ) the form*

$$p(\varrho,\vartheta)|_{\Omega_i}=\varrho^{\alpha}(a_i\sin(\alpha\vartheta)+b_i\cos(\alpha\vartheta)), \tag{8.1}$$

see [34]. Note that p only belongs to $H^{1+\alpha}(\Omega)$ and exhibits a singularity at the origin. It is continuous but only the normal component of its flux $-\mathbf{S}\nabla p$ is continuous across the interfaces.

$s_1=s_3=5,\ s_2=s_4=1$			
$\alpha=$	0.53544095		
$a_1=$	0.44721360	$b_1=$	1.00000000
$a_2=$	−0.74535599	$b_2=$	2.33333333
$a_3=$	−0.94411759	$b_3=$	0.55555556
$a_4=$	−2.40170264	$b_4=$	−0.48148148
$s_1=s_3=100,\ s_2=s_4=1$			
$\alpha=$	0.12690207		
$a_1=$	0.10000000	$b_1=$	1.00000000
$a_2=$	−9.60396040	$b_2=$	2.96039604
$a_3=$	−0.48035487	$b_3=$	−0.88275659
$a_4=$	7.70156488	$b_4=$	−6.45646175

TABLE 8.1
The values of the coefficients in (8.1) for the two choices of the diffusion tensor $\mathbf{S}$.

In our experiments we use the finite volume scheme (3.1), (3.3), which we extend from triangular grids admissible in the sense of [16, Definition 9.1] to strictly Delaunay triangular meshes, cf. [16, Example 9.1]. For the diffusion tensor the harmonic averaging is employed and modified by taking into account the distances of the circumcenters $\mathbf{x}_K$, $K\in\mathcal{T}_h$, from the sides of K; for details, we refer to [50]. In such a setting, the results of the mixed finite element and finite volume methods are very close, as demonstrated by the results presented in [48, 50]. The advantage of the finite volume method in this case is that it yields symmetric positive definite matrices with at most four nonzero elements in each row.

We start our computations with an unstructured mesh $\mathcal{T}_h$ of Ω consisting of 112 elements. In Example 8.1 the mesh is refined uniformly, i.e., each triangular element in $\mathcal{T}_h$ is subdivided into four elements. In Example 8.2 it is refined adaptively. The adaptive mesh refinement strategy is described in detail in [50]; the essential point is to equilibrate the estimated local errors while keeping the mesh strictly Delaunay. The refinement process is stopped when the number of elements in $\mathcal{T}_h$ exceeds the number 1700, which results in all cases in algebraic systems of similar size. This relatively

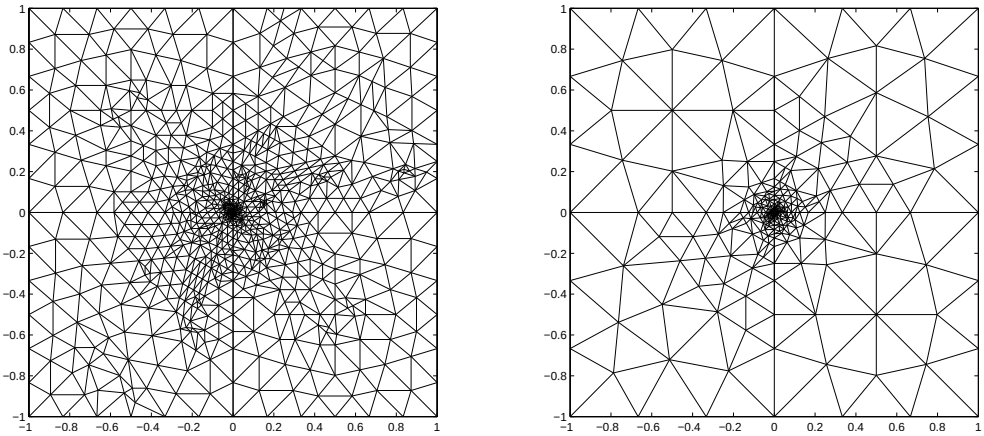

FIG. 8.1. *The adaptively refined mesh with 1812 elements for the model problem in Example 8.2 with* $s_1 = s_3 = 5$, $s_2 = s_4 = 1$ *(*$\alpha = 0.53544095$*) (left part) and with 1736 elements for the problem with* $s_1 = s_3 = 100$, $s_2 = s_4 = 1$ *(*$\alpha = 0.12690207$*) (right part).*

small number of elements was chosen because of the second choice of coefficients in Example 8.2. The singularity is here so significant that for around 2000 triangles, the diameter of the smallest triangles near the origin is 10^{-15}, which approaches machine precision. The final mesh in Example 8.1 consists of 1792 elements, in the first case of Example 8.2 of 1812 elements, and in the second case of Example 8.2 of 1736 elements. The last two meshes are shown in Figure 8.1. Recall that the matrix size N in the finite volume method is equal to the number of mesh elements.

The arising algebraic systems (3.2) are solved approximately using the conjugate gradient method preconditioned by the incomplete Cholesky factorization with no fill-in (IC(0)), see [24]. For illustrative purposes, we use for all meshes the zero initial guess. In practical computations, the approximate solution from the previous refinement level should be interpolated onto the current mesh and used as a starting vector. For each approximate solution P^{a} of (3.2) computed by the number of conjugate gradients steps determined by the stopping criteria specified below, we evaluate the estimator η_{NC} defined in Theorem 5.2 as $|||\tilde{p}_h^{\mathrm{a}} - \mathcal{I}_{\mathrm{Os}}(\tilde{p}_h^{\mathrm{a}})|||$ (we consider the additional error from the inhomogeneous boundary condition as negligible). Then we compute the algebraic error estimators described in Section 7. Note that η_{R} is zero since $f = 0$ in both examples. In order to illustrate the behavior of the nonconforming and algebraic error estimators, the conjugate gradient method for a given mesh is stopped when the local stopping criterion (6.3) based on the estimator $\eta^{(3)}_{\mathrm{AE}}(\mathbf{r}_h)$ is satisfied, i.e., when

$$\eta^{(3)}_{\mathrm{AE},K}(\mathbf{r}_h) := \|\mathbf{S}^{-\frac{1}{2}}\mathbf{r}_h\|_K \leq \gamma\, \eta_{\mathrm{NC},K} \qquad \forall K \in \mathcal{T}_h$$

with $\gamma = 10^{-3}$. In practical computations, it is advisable to use a value of γ much closer to one, in dependence on the given problem.

Results for meshes obtained at the last stage of the uniform or adaptive mesh refinement process are illustrated in Figures 8.2–8.4. In all figures the iteration number represents the iteration n of the conjugate gradient method for solving the algebraic system (3.2). The results for the Laplace equation in Example 8.1 are plotted in Figure 8.2. The results for Example 8.2 with the inhomogeneous $\mathbf{S}$ with s_i given in

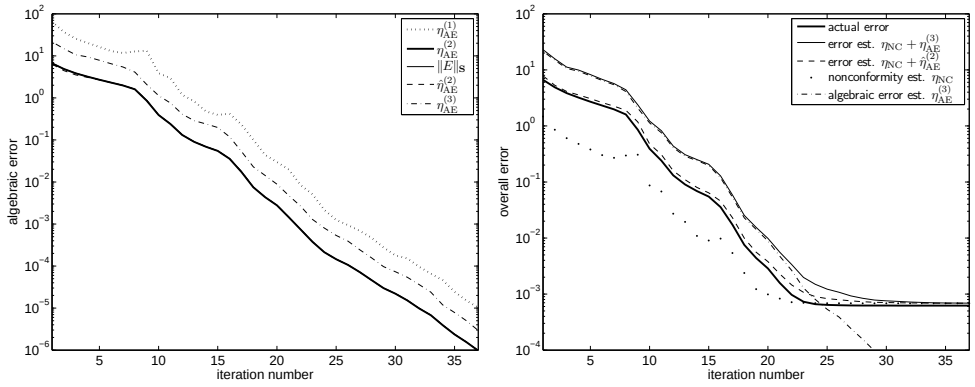

FIG. 8.2. *Example 8.1, uniformly refined mesh with* 1792 *elements. Left part: algebraic error estimators* $\eta^{(1)}_{\rm AE}$ *(dotted line),* $\eta^{(2)}_{\rm AE}$ *(bold solid line),* $\|E\|_{\mathbb{S}}$ *(solid line),* $\hat\eta^{(2)}_{\rm AE}$ *(dashed line), and* $\eta^{(3)}_{\rm AE}$ *(dash-dotted line); right part: actual error (bold solid line), error estimate* $\eta_{\rm NC}+\eta^{(3)}_{\rm AE}$ *(solid line), error estimate* $\eta_{\rm NC}+\hat\eta^{(2)}_{\rm AE}$ *(dashed line), nonconformity estimator* $\eta_{\rm NC}$ *(dots), and algebraic error estimator* $\eta^{(3)}_{\rm AE}$ *(dash-dotted line).*

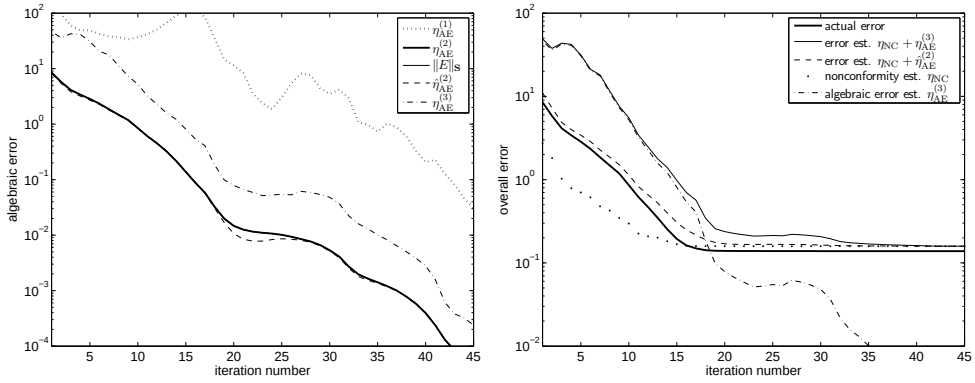

FIG. 8.3. *Example 8.2 with* $s_1=s_3=5$, $s_2=s_4=1$ *(*$\alpha=0.53544095$*), adaptively refined mesh with* 1812 *elements. Left part: algebraic error estimators* $\eta^{(1)}_{\rm AE}$ *(dotted line),* $\eta^{(2)}_{\rm AE}$ *(bold solid line),* $\|E\|_{\mathbb{S}}$ *(solid line),* $\hat\eta^{(2)}_{\rm AE}$ *(dashed line), and* $\eta^{(3)}_{\rm AE}$ *(dash-dotted line); right part: actual error (bold solid line), error estimate* $\eta_{\rm NC}+\eta^{(3)}_{\rm AE}$ *(solid line), error estimate* $\eta_{\rm NC}+\hat\eta^{(2)}_{\rm AE}$ *(dashed line), nonconformity estimator* $\eta_{\rm NC}$ *(dots), and algebraic error estimator* $\eta^{(3)}_{\rm AE}$ *(dash-dotted line).*

the upper part of Table 8.1 are plotted in Figure 8.3 and the results for $\mathbf{S}$ with s_i as given in the lower part of Table 8.1 are plotted in Figure 8.4.

Left parts of Figures 8.2–8.4 show the values of the algebraic error estimators proposed in Section 7. The estimator $\eta^{(1)}_{\rm AE}$ based on the weighted norm of the algebraic residual vector, see Lemma 7.1, is plotted by dotted lines. As expected, it provides the worst information among all considered measures of the algebraic error. This is in particular evident in Example 8.2 where the adaptive mesh refinement is employed, see Figures 8.3 and 8.4 (in Figure 8.4 it is out of the scale for almost all iteration steps displayed). The algebraic error estimator $\eta^{(2)}_{\rm AE}$ defined in Theorem 7.2 is plotted by bold solid lines. We evaluate $\eta^{(2)}_{\rm AE}$ by solving the saddle-point system (7.6) using a

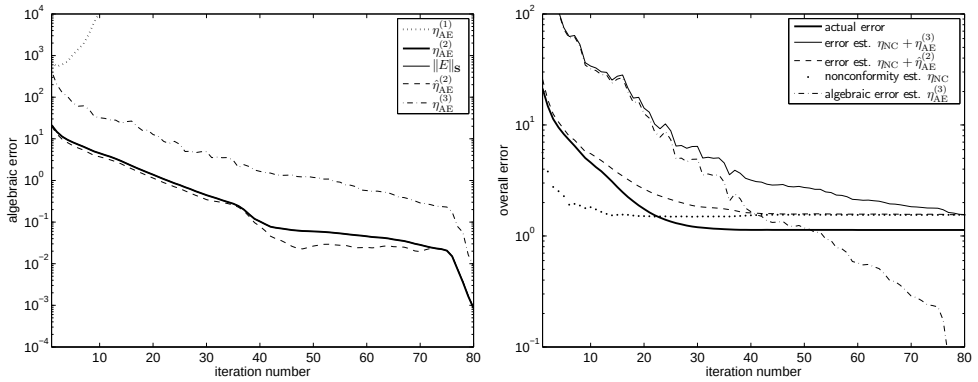

FIG. 8.4. *Example 8.2 with* $s_1 = s_3 = 100$, $s_2 = s_4 = 1$ *(*$\alpha = 0.12690207$*), adaptively refined mesh with* 1736 *elements. Left part: algebraic error estimators* $\eta_{\rm AE}^{(1)}$ *(dotted line),* $\eta_{\rm AE}^{(2)}$ *(bold solid line),* $\|E\|_{\mathbb{S}}$ *(solid line),* $\hat\eta_{\rm AE}^{(2)}$ *(dashed line), and* $\eta_{\rm AE}^{(3)}$ *(dash-dotted line); right part: actual error (bold solid line), error estimate* $\eta_{\rm NC}+\eta_{\rm AE}^{(3)}$ *(solid line), error estimate* $\eta_{\rm NC}+\hat\eta_{\rm AE}^{(2)}$ *(dashed line), nonconformity estimator* $\eta_{\rm NC}$ *(dots), and algebraic error estimator* $\eta_{\rm AE}^{(3)}$ *(dash-dotted line).*

direct algebraic solver and computing $\sqrt{Q^t\mathbb{A}Q}$, as $\eta_{\rm AE}^{(2)} = \|\mathbf{S}^{-\frac{1}{2}}\mathbf{q}_h\| = \sqrt{Q^t\mathbb{A}Q}$. The algebraic energy errors $\|E_n^{\rm CG}\|_{\mathbb{S}}$ of conjugate gradients (see Remark 7.6) are evaluated by solving $\mathbb{S}E_n^{\rm CG} = R_n^{\rm CG}$ using a direct solver. Here $\mathbb{S}$ is the finite volume matrix and $R_n^{\rm CG} := H - \mathbb{S}P_n^{\rm CG}$. The values $\|E_n^{\rm CG}\|_{\mathbb{S}}$ are plotted by solid lines. Note that as suggested in Remark 7.5, the values of $\|E_n^{\rm CG}\|_{\mathbb{S}}$ (we will use the simpler notation $\|E\|_{\mathbb{S}}$ from now on) and $\eta_{\rm AE}^{(2)}$ are very close. In fact, it is not possible to distinguish between them in Figures 8.2–8.4. The quantity $\hat\eta_{\rm AE}^{(2)}$ plotted by dashed lines stands for the *estimate* of $\|E\|_{\mathbb{S}}$ described in Remark 7.6, where we have used $\nu = 5$. Finally, the estimator $\eta_{\rm AE}^{(3)}$ of Section 7.3 is plotted by dash-dotted lines.

The estimate $\hat\eta_{\rm AE}^{(2)}$ is close to $\|E\|_{\mathbb{S}}$ (and, as noted above, also to the algebraic error estimator $\eta_{\rm AE}^{(2)}$), even though there are visible underestimations in Figures 8.3 and 8.4. This is due to the rather slow convergence of the conjugate gradient method in this case, cf. [41, 42]. The estimator $\eta_{\rm AE}^{(3)}$ on the contrary represents a *guaranteed upper bound* for the algebraic error estimator $\eta_{\rm AE}^{(2)}$. Please notice that this is a conceptual difference between $\eta_{\rm AE}^{(3)}$ and $\hat\eta_{\rm AE}^{(2)}$. We see that $\eta_{\rm AE}^{(3)}$ only slightly overestimates the algebraic error estimator $\eta_{\rm AE}^{(2)}$. This holds true also in Example 8.2 where the presence of the singularity and using the adaptive mesh refinement complicate the task.

On right parts of Figures 8.2–8.4, we present the actual energy norm of the overall error $|||p - \tilde p_h^{\rm a}|||$ (bold solid lines). We compute it in each triangle by the 7-point quadrature formula, see, e.g., [53, Section 9.10] (we consider the additional quadrature error negligible). The guaranteed upper bound $\eta_{\rm NC} + \eta_{\rm AE}^{(3)}$ on $|||p - \tilde p_h^{\rm a}|||$ is represented by solid lines, while its components, the nonconformity estimator $\eta_{\rm NC}$ and the algebraic error estimator $\eta_{\rm AE}^{(3)}$, are plotted by dots and dash-dotted lines, respectively. For comparison, we also include the estimate $\eta_{\rm NC} + \hat\eta_{\rm AE}^{(2)}$ plotted by dashed lines.

Figures 8.2–8.4 show that for small number of iterations the algebraic part of the error dominates. As the number of iterations of the conjugate gradient method grows, the algebraic part of the error drops to the level of the discretization error,

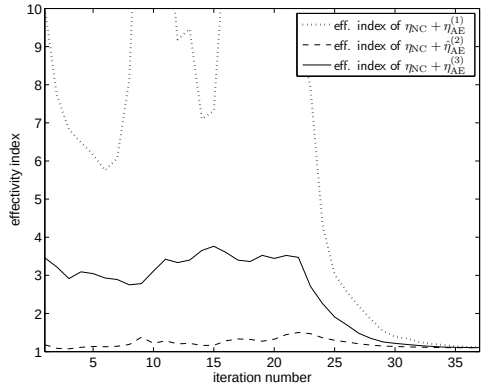

FIG. 8.5. *Example 8.1: effectivity indices of the error estimates* $\eta_{\mathrm{NC}} + \eta_{\mathrm{AE}}^{(1)}$ *(dotted line),* $\eta_{\mathrm{NC}} + \hat\eta_{\mathrm{AE}}^{(2)}$ *(dashed line), and* $\eta_{\mathrm{NC}} + \eta_{\mathrm{AE}}^{(3)}$ *(solid line).*

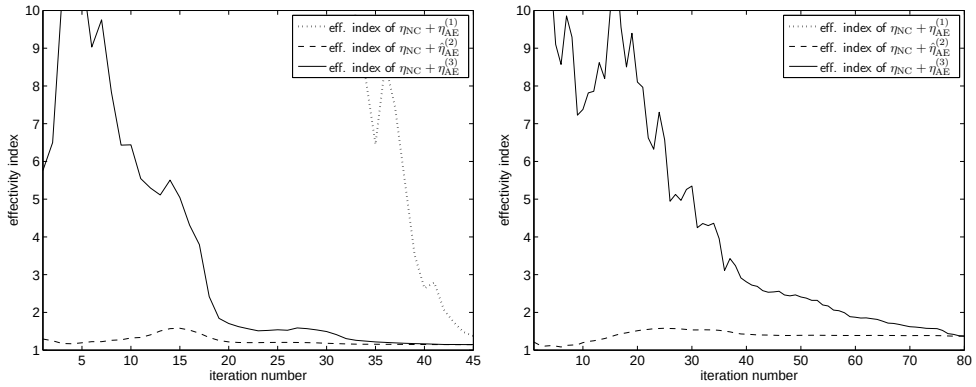

FIG. 8.6. *Example 8.2 with* $s_1 = s_3 = 5$, $s_2 = s_4 = 1$ *(*$\alpha = 0.53544095$*) (left part) and Example 8.2 with* $s_1 = s_3 = 100$, $s_2 = s_4 = 1$ *(*$\alpha = 0.12690207$*) (right part): effectivity indices of the error estimates* $\eta_{\mathrm{NC}} + \eta_{\mathrm{AE}}^{(1)}$ *(dotted line),* $\eta_{\mathrm{NC}} + \hat\eta_{\mathrm{AE}}^{(2)}$ *(dashed line), and* $\eta_{\mathrm{NC}} + \eta_{\mathrm{AE}}^{(3)}$ *(solid line). The dotted line is essentially out of the scale of the figure*

which is reflected by the fact that the curves of η_{NC} and $\eta_{\mathrm{AE}}^{(3)}$ intersect. Then η_{NC} starts to stagnate, while the estimate on the algebraic error $\eta_{\mathrm{AE}}^{(3)}$ further decreases and it ultimately gets negligible in comparison with the discretization error. Our stopping criteria for iterative solvers (6.1) and (6.3) essentially state that it is meaningless to continue in iterations when solving the linear algebraic system after the iteration number for which $\eta_{\mathrm{AE},K}^{(3)}(\mathbf{r}_h) \approx \gamma\, \eta_{\mathrm{NC},K}$ is reached.

The quality of the estimates on the overall error using all the algebraic error estimators discussed in this paper, i.e., the quantities $\eta_{\mathrm{NC}}+\eta_{\mathrm{AE}}^{(1)}$, $\eta_{\mathrm{NC}}+\hat\eta_{\mathrm{AE}}^{(2)}$, and $\eta_{\mathrm{NC}}+\eta_{\mathrm{AE}}^{(3)}$, is illustrated in Figures 8.5 and 8.6. Here we have plotted the effectivity indices, i.e., the ratios of the estimated and the actual overall error. As it can be expected, $\eta_{\mathrm{AE}}^{(1)}$ gives a large overestimation of the actual algebraic error and the corresponding effectivity index is very poor (in the right part of Figure 8.6 it is completely out of scale). Recall that the estimate $\eta_{\mathrm{NC}} + \eta_{\mathrm{AE}}^{(3)}$ gives a guaranteed upper bound. Its effectivity index is very reasonable even in the first conjugate gradients iterations in

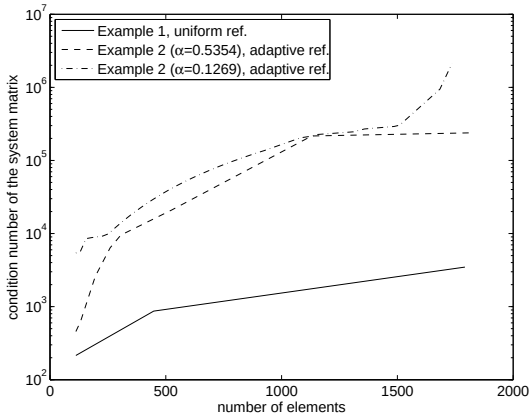

FIG. 8.7. *Condition numbers of system matrices with respect to the number of elements in Examples 8.1 and 8.2*

the second case of Example 8.2 and it gets close to the optimal value of one quickly. Finally, even though $\hat{\eta}_{\mathrm{AE}}^{(2)}$ does not represent a guaranteed upper bound for $\eta_{\mathrm{AE}}^{(2)}$, the estimate $\eta_{\mathrm{NC}} + \hat{\eta}_{\mathrm{AE}}^{(2)}$ gives in all our experiments very tight estimates for the overall error. The effectivity index is here in all cases remarkably close to one.

An interesting issue connected with solving the linear algebraic systems arising from problems with mesh refinement is the growth of the condition number of the system matrix as the refinement proceeds. This is of particular importance when adaptive mesh refinement is applied in the presence of singularity. Without taking into consideration the algebraic part of the error it is sometimes claimed in the literature that adaptive mesh refinement can provide *an arbitrary accurate numerical solution*. Similar claims should however be carefully examined and revisited. Adaptive discretization in the presence of singularity can lead to highly ill-conditioned systems of linear algebraic equations. This can have two main effects:

- the iterative solvers can become slow and the computation of the numerical solution can become more expensive;
- the maximum attainable accuracy of the (direct as well as iterative) linear algebraic solvers can for very highly ill-conditioned systems become very poor, which can prevent reaching the desired accuracy of the numerical solution of the original problem regardless how small the discretization error becomes.

Figure 8.7 shows for our examples the dependence of the spectral condition number of the system matrix $\mathbb{S}$ on the number of elements in the mesh. In the case of the homogeneous diffusion tensor and the uniform mesh refinement of Example 8.1, the condition number of $\mathbb{S}$ is growing according to the well-known theoretical result as $O(N^2)$. In Example 8.2 with inhomogeneous diffusion coefficients, the solutions exhibit a singularity at the origin. The adaptive mesh refinement compensates for the effect of the singularity which causes the nonuniform distribution of the error. This has an unfavorable effect on the growth of the condition number of the system matrix $\mathbb{S}$, see Figure 8.7. If we proceed with the refinement, the condition number of $\mathbb{S}$ will soon reach reach the value of the inverse of machine precision, which will make algebraic computations practically meaningless. Though a more detailed discussion of this problem is beyond the scope of this paper, we believe that its role can be substantial and it will have to be investigated in a near future.

9. Concluding remarks. Deriving tight a posteriori estimates under the assumption that the associated systems of linear equations are solved exactly is mathematically much easier than without this assumption. It however precludes the efficient use of such estimates in practical large scale computations, where the linear systems, solved by iterative algebraic solvers, are never solved exactly, and should even be *solved inexactly on purpose*.

Efficient using of iterative algebraic solvers requires balancing the algebraic and discretization errors. For practical purposes, the message is that it is useless to make an extensive number of algebraic solver iterations after the algebraic error drops significantly below the discretization error.

Adaptive mesh refinement can lead in the presence of singularity to pathologically ill-conditioned linear algebraic systems. Getting an arbitrary numerical accuracy by an over limit refinements of meshes is illusive. The resulting linear algebraic problem can get so ill-conditioned that it can eventually prevent obtaining of a single digit of accuracy of practically computed numerical solution. As pointed out in a schematic way in [40], modeling, discretization, and computation form interconnected stages of a *single solution process*. The *errors on the different stages should be in balance*. Considering the numerical analysis and the discretization stages separately from computations is philosophically wrong. Similar approaches will lead in solving difficult problems to dead ends.

Acknowledgments. This work was initiated during the summer school CEMRACS organized by the laboratory of the third author in summer 2007 in Luminy/Marseille, France and the authors gratefully acknowledge all the support. The second author thanks for the support during his visit of the Jacques-Louis Lions laboratory in September 2008.

REFERENCES

[1] Y. Achdou, C. Bernardi, and F. Coquel, *A priori and a posteriori analysis of finite volume discretizations of Darcy's equations*, Numer. Math., 96 (2003), pp. 17–42.
[2] M. Ainsworth and J. T. Oden, *A posteriori error estimation in finite element analysis*, Pure and Applied Mathematics (New York), Wiley-Interscience [John Wiley & Sons], New York, 2000.
[3] M. Arioli, *A stopping criterion for the conjugate gradient algorithm in a finite element method framework*, Numer. Math., 97 (2004), pp. 1–24.
[4] M. Arioli, D. Loghin, and A. J. Wathen, *Stopping criteria for iterations in finite element methods*, Numer. Math., 99 (2005), pp. 381–410.
[5] I. Babuška, *Numerical stability in problems of linear algebra*, SIAM J. Numer. Anal., 9 (1972), pp. 53–77.
[6] I. Babuška and W. C. Rheinboldt, *Error estimates for adaptive finite element computations*, SIAM J. Numer. Anal., 15 (1978), pp. 736–754.
[7] M. Bebendorf, *A note on the Poincaré inequality for convex domains*, Z. Anal. Anwendungen, 22 (2003), pp. 751–756.
[8] R. Becker, C. Johnson, and R. Rannacher, *Adaptive error control for multigrid finite element methods*, Computing, 55 (1995), pp. 271–288.
[9] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, vol. 15 of Springer Series in Computational Mathematics, Springer-Verlag, New York, 1991.
[10] C. Burstedde and A. Kunoth, *Fast iterative solution of elliptic control problems in wavelet discretization*, J. Comput. Appl. Math., 196 (2006), pp. 299–319.
[11] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, vol. 4 of Studies in Mathematics and its Applications, North-Holland, Amsterdam, 1978.
[12] P. Deuflhard, *Cascadic conjugate gradient methods for elliptic partial differential equations: algorithm and numerical results*, in Domain decomposition methods in scientific and engi-

neering computing (University Park, PA, 1993), vol. 180 of Contemp. Math., Amer. Math. Soc., Providence, RI, 1994, pp. 29–42.
[13] H. C. Elman and G. H. Golub, *Inexact and preconditioned Uzawa algorithms for saddle point problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1645–1661.
[14] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*, Oxford University Press, 2005.
[15] A. Ern, A. F. Stephansen, and M. Vohralík, *Guaranteed and robust a posteriori error estimation based on flux reconstruction for discontinuous Galerkin methods*. Preprint R07050, Laboratoire Jacques-Louis Lions, submitted for publication, 2007.
[16] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handbook of Numerical Analysis, Vol. VII, North-Holland, Amsterdam, 2000, pp. 713–1020.
[17] 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.
[18] G. H. Golub and G. Meurant, *Matrices, moments and quadrature II: how to compute the norm of the error in iterative methods*, BIT, 37 (1997), pp. 687–705.
[19] G. H. Golub and Z. Strakoš, *Estimates in quadratic formulas*, Numer. Algorithms, 8 (1994), pp. 241–268.
[20] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Natl. Bur. Stand., 49 (1952), pp. 409–436.
[21] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second ed., 2002.
[22] S. Korotov, *Two-sided a posteriori error estimates for linear elliptic problems with mixed boundary conditions*, Appl. Math., 52 (2007), pp. 235–249.
[23] Y. Maday and A. T. Patera, *Numerical analysis of a posteriori finite element bounds for linear functional outputs*, Math. Models Methods Appl. Sci., 10 (2000), pp. 785–799.
[24] J. A. Meijerink and H. A. van der Vorst, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comp., 31 (1977), pp. 148–162.
[25] G. Meurant, *The computation of bounds for the norm of the error in the conjugate gradient algorithm*, Numer. Algorithms, 16 (1997), pp. 77–87.
[26] G. Meurant and Z. Strakoš, *The Lanczos and conjugate gradient algorithms in finite precision arithmetic*, Acta Numer., 15 (2006), pp. 471–542.
[27] J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Masson, Paris, 1967.
[28] 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.
[29] L. E. Payne and H. F. Weinberger, *An optimal Poincaré inequality for convex domains*, Arch. Ration. Mech. Anal., 5 (1960), pp. 286–292.
[30] A. Quarteroni and A. Valli, *Numerical approximation of partial differential equations*, vol. 23 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 1994.
[31] K. Rektorys, *Variational Methods in Mathematics, Science, and Engineering*, Kluwer, Dordrecht, 1982.
[32] S. Repin, *A posteriori error estimation for nonlinear variational problems by duality theory*, Zapiski Nauchnykh Seminarov, 243 (1997), pp. 201–214.
[33] S. I. Repin and A. Smolianski, *Functional-type a posteriori error estimates for mixed finite element methods*, Russian J. Numer. Anal. Math. Modelling, 20 (2005), pp. 365–382.
[34] B. Rivière, M. F. Wheeler, and K. Banas, *Part II. Discontinuous Galerkin method applied to single phase flow in porous media*, Comput. Geosci., 4 (2000), pp. 337–349.
[35] U. Rüde, *Error estimators based on stable splittings*, in Proceedings of the 7th International Conference on Domain Decomposition in Science and Engineering Computing, Pennsylvania State University, D. Keyes, ed., vol. 180, Providence: American Mathematical Society, 1994, pp. 111–118.
[36] Y. Saad, *Iterative Methods for Sparse Linear Systems*, SIAM, 2nd ed., 2003.
[37] Y. Saad and M. H. Schultz, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[38] V. Shaidurov and L. Tobiska, *The convergence of the cascadic conjugate-gradient method applied to elliptic problems in domains with re-entrant corners*, Math. Comp., 69 (2000), pp. 501–520.
[39] Z. Strakoš, *Model reduction using the Vorobyev moment problem*, Numer. Algorithms, to appear (2008).
[40] Z. Strakoš and J. Liesen, *On numerical stability in large scale linear algebraic computations*,

ZAMM Z. Angew. Math. Mech., 85 (2005), pp. 307–325.
[41] 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.
[42] Z. Strakoš and P. Tichý, *Error estimation in preconditioned conjugate gradients*, BIT, 45 (2005), pp. 789–817.
[43] H. A. van der Vorst, *Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of non-symmetric linear systems*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 631–644.
[44] R. S. Varga, *Matrix Iterative Analysis*, Springer, Berlin, 2 ed., 1992.
[45] R. Verfürth, *A review of a posteriori error estimation and adaptive mesh-refinement techniques*, Teubner-Wiley, Stuttgart, 1996.
[46] M. Vohralík, *On the discrete Poincaré–Friedrichs inequalities for nonconforming approximations of the Sobolev space H^1*, Numer. Funct. Anal. Optim., 26 (2005), pp. 925–952.
[47] ———, *Equivalence between lowest-order mixed finite element and multi-point finite volume methods on simplicial meshes*, M2AN Math. Model. Numer. Anal., 40 (2006), pp. 367–391.
[48] ———, *A posteriori error estimates for lowest-order mixed finite element discretizations of convection-diffusion-reaction equations*, SIAM J. Numer. Anal., 45 (2007), pp. 1570–1599.
[49] ———, *Guaranteed and fully robust a posteriori error estimates for conforming discretizations of diffusion problems with discontinuous coefficients*. Preprint R08009, Laboratoire Jacques-Louis Lions, submitted for publication, 2008.
[50] ———, *Residual flux-based a posteriori error estimates for finite volume and related locally conservative methods*, Numer. Math., DOI 10.1007/s00211-008-0168-4 (electronic), (2008).
[51] ———, *Unified primal formulation-based a priori and a posteriori error analysis of mixed finite element methods*. Preprint R08036, Laboratoire Jacques-Louis Lions, submitted for publication, 2008.
[52] 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.
[53] O. C. Zienkiewicz and R. L. Taylor, *The Finite Element Method. Volume I: The Basis*, Butterworth-Heinemann, Oxford, 5th ed., 2000.