ESAIM: M2AN 57 (2023) 1–27
https://doi.org/10.1051/m2an/2022078

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

A POSTERIORI ERROR ESTIMATES FOR MIXED FINITE ELEMENT DISCRETIZATIONS OF THE NEUTRON DIFFUSION EQUATIONS

Patrick Ciarlet[1], Minh Hieu Do[2] and François Madiot[2,*]

Abstract. We analyse *a posteriori* error estimates for the discretization of the neutron diffusion equations with mixed finite elements. We provide guaranteed and locally efficient estimators on a base block equation, the one-group neutron diffusion equation. We pay particular attention to AMR strategies on Cartesian meshes, since such structures are common for nuclear reactor core applications. We exhibit a robust marker strategy for this specific constraint, the *direction marker* strategy.

Mathematics Subject Classification. 65J10, 65N15, 65N30, 65N50.

Received June 4, 2020. Accepted September 13, 2022.

1. Introduction

The diffusion equation can model different physical phenomena, for instance Darcy's law, Fick's law or the neutron diffusion. Among models that are used in the nuclear industry, the multigroup neutron diffusion equation plays a central role [1]. The base block is the one-group neutron diffusion equation. In [2,3], the first author and co-authors carried out the numerical analysis of this one-group neutron diffusion equation with a source term, discretized with mixed finite elements. The analysis included in particular the case of low-regularity solutions. *A priori estimates* were derived in the process. A natural question is then the *a posteriori* analysis of the method, to further optimize the cost of the numerical method. This is the main topic we address in this paper.

A posteriori analysis for mixed finite elements has been extensively studied, see [4–7] and references therein for the Poisson equation [8, 9], for the diffusion-reaction equation (one-group neutron diffusion equation), and [10] for the convection-diffusion-reaction equation.

Nuclear reactor cores often have a Cartesian geometry. Indeed, in the models, the base brick, which is called a cell, is a rectangular cuboid of $\mathbb{R}^3$. The global layout is a set of cells that are distributed on a 3D grid, so that the global domain of the reactor core is represented by a rectangular cuboid of $\mathbb{R}^3$. Each cell can be made of fuel, absorbing or reflector material. To account for the different materials, the coefficients in the models are *piecewise polynomials* (possibly piecewise constant) with respect to the position, *i.e.* their restriction to each cell is a polynomial [1, 11, 12]. In practice the coefficients characterizing the materials may differ from one cell to another by a factor of order 10 or more.

Keywords and phrases. Neutronics, diffusion equation, mixed formulation, low regularity solution, *a posteriori* error estimates, mesh refinement.

[1] POEMS, CNRS, INRIA, ENSTA Paris, Institut Polytechnique de Paris, 91120 Palaiseau, France.

[2] Université Paris-Saclay, CEA, Service d'Études des Réacteurs et de Mathématiques Appliquées, 91191 Gif-sur-Yvette, France.
*Corresponding author: `francois.madiot@cea.fr`

The outline is as follows.

In Sections 2 and 3, we introduce some notations and our model problem. Then in Section 4, we recall how it can be solved in a mixed setting. To that aim we build the standard equivalent variational formulation, and provide the existing *a priori* numerical analysis results that allow one to compare the discrete solution to the exact one. For the discretization, we choose the well-known Raviart–Thomas–Nédélec finite element RTN_k, where $k \geq 0$ denotes the order.

In Section 5, we propose the *a posteriori* analysis of the model. We begin by the reconstruction of the solution (*via* post-processing), which can be devised in at least two ways: one is specific to the lowest-order, and the second one can be applied to any order. We also mention an averaging approach for the reconstruction. In Section 6, we propose some numerical experiments to compare the resulting strategies. For that, we focus on a specific discretization, based on Cartesian meshes. This kind of discretization is of particular importance for nuclear core simulations.

2. Notations

We choose the same notations as in [2, 3]. Throughout the paper, C is used to denote a generic positive constant which is independent of the mesh size, the mesh and the quantities/fields of interest. We also use the shorthand notation $A \lesssim B$ for the inequality $A \leq CB$, where A and B are two scalar quantities, and C is a generic constant.

Vector-valued (resp. tensor-valued) function spaces are written in boldface character (resp. blackboard characters); for the latter, the index *sym* indicates symmetric fields. Given an open set $\mathcal{O} \subset \mathbb{R}^d$, $d = 1, 2, 3$, we use the notation $(\cdot,\cdot)_{0,\mathcal{O}}$ (respectively $\|\cdot\|_{0,\mathcal{O}}$) for the $L^2(\mathcal{O})$ and $\mathbf{L}^2(\mathcal{O}) = (L^2(\mathcal{O}))^d$ scalar products (resp. norms). More generally, $(\cdot,\cdot)_{s,\mathcal{O}}$ and $\|\cdot\|_{s,\mathcal{O}}$ (respectively $|\cdot|_{s,\mathcal{O}}$) denote the scalar product and norm (resp. semi-norm) of the Sobolev spaces $H^s(\mathcal{O})$ and $\mathbf{H}^s(\mathcal{O}) = (H^s(\mathcal{O}))^d$ for $s \in \mathbb{R}$ (resp. for $s > 0$).

If moreover the boundary $\partial\mathcal{O}$ is Lipschitz, $\mathbf{n}$ denotes the unit outward normal vector field to $\partial\mathcal{O}$. Finally, it is assumed that the reader is familiar with vector-valued function spaces related to the diffusion equation, such as $\mathbf{H}(\mathrm{div}\,;\mathcal{O})$, $\mathbf{H}_0(\mathrm{div}\,;\mathcal{O})$ etc.

Specifically, we let Ω be a bounded, connected and open subset of $\mathbb{R}^d$ for $d = 2, 3$, having a Lipschitz boundary which is piecewise smooth. We split Ω into N connected open disjoints parts $\{\Omega_i\}_{1\leq i\leq N}$ with Lipschitz, piecewise smooth boundaries: $\overline{\Omega} = \cup_{1\leq i\leq N}\overline{\Omega_i}$ and the set $\{\Omega_i\}_{1\leq i\leq N}$ is called a partition of Ω. For a field v defined over Ω, we shall use the notations $v_i = v_{|\Omega_i}$, for $1 \leq i \leq N$.

Given a partition $\{\Omega_i\}_{1\leq i\leq N}$ of Ω, we introduce a function space with piecewise regular elements:

$$\mathcal{P}W^{1,\infty}(\Omega) = \big\{ D \in L^\infty(\Omega) \,|\, D_i \in W^{1,\infty}(\Omega_i),\ 1 \leq i \leq N \big\}.$$

To measure $\psi \in \mathcal{P}W^{1,\infty}(\Omega)$, we use the natural norm $\|\psi\|_{\mathcal{P}W^{1,\infty}(\Omega)} = \max_{i=1,N} \|\psi_i\|_{W^{1,\infty}(\Omega_i)}$.

3. The model

Given a source term $S_f \in L^2(\Omega)$, we consider the following neutron diffusion equation, with vanishing Dirichlet boundary condition. In its primal form, it is written:

$$\begin{cases} \text{Find } \phi \in H_0^1(\Omega) \text{ such that} \\ -\mathrm{div}\,\mathbb{D}\,\mathbf{grad}\,\phi + \Sigma_a\,\phi = S_f \text{ in } \Omega, \end{cases} \tag{3.1}$$

where ϕ, $\mathbb{D}$, and Σ_a denote respectively the neutron flux, the diffusion coefficient and the macroscopic absorption cross section. Finally, S_f denotes the fission source. When solving the neutron diffusion equation, $\mathbb{D}$ is scalar-valued. We choose to consider more generally that $\mathbb{D}$ is a (symmetric) tensor-valued coefficient. The coefficients defining Problem (3.1) satisfy the assumptions:

$$\begin{cases} (\mathbb{D}, \Sigma_a) \in \mathbb{L}^\infty_{sym}(\Omega) \times L^\infty(\Omega), & \\ \exists D_*, D^* > 0,\ \forall \boldsymbol{z} \in \mathbb{R}^d,\ D_*\|\boldsymbol{z}\|^2 \leq (\mathbb{D}\,\boldsymbol{z}, \boldsymbol{z}) \leq D^*\|\boldsymbol{z}\|^2 & \text{a.e. in } \Omega, \\ \exists (\Sigma_a)_*, (\Sigma_a)^* > 0,\ 0 < (\Sigma_a)_* \leq \Sigma_a \leq (\Sigma_a)^* & \text{a.e. in } \Omega. \end{cases} \tag{3.2}$$

Classically, Problem (3.1) is equivalent to the following variational formulation:

$$\begin{cases} \text{Find } \phi \in H_0^1(\Omega) \text{ such that} \\ \forall \psi \in H_0^1(\Omega), \quad (\mathbb{D}\,\mathbf{grad}\,\phi, \mathbf{grad}\,\psi)_{0,\Omega} + (\Sigma_a \phi, \psi)_{0,\Omega} = (S_f, \psi)_{0,\Omega}. \end{cases} \tag{3.3}$$

Under the assumptions (3.2) on the coefficients, the primal problem (3.1) is well-posed, in the sense that for all $S_f \in L^2(\Omega)$, there exists one and only one solution $\phi \in H_0^1(\Omega)$ that solves (3.1), with the bound $\|\phi\|_{1,\Omega} \lesssim \|S_f\|_{0,\Omega}$. Provided that the coefficient $\mathbb{D}$ is piecewise smooth, the solution has extra smoothness (see *e.g.* Prop. 1 in [2]). Instead of imposing a Dirichlet boundary condition on $\partial\Omega$, one can consider a Neumann or Fourier boundary condition $\mu_F \phi + (\mathbb{D}\,\mathbf{grad}\,\phi) \cdot \mathbf{n} = 0$, with $\mu_F \geq 0$. Results are similar. Throughout the paper, we add remarks on the extension to the situation where $\Sigma_a \geq 0$ may vanish. In particular, the analysis we propose covers both the pure diffusion case, and the diffusion-reaction case.

4. Variational formulation and discretization

Let us introduce the function spaces:

$$\mathcal{H} = \{ \xi = (\mathbf{q}, \psi) \in \mathbf{L}^2(\Omega) \times L^2(\Omega) \}, \ \|\xi\|_{\mathcal{H}} = \left(\|\mathbf{q}\|_{0,\Omega}^2 + \|\psi\|_{0,\Omega}^2 \right)^{1/2};$$
$$\mathcal{X} = \{ \xi = (\mathbf{q}, \psi) \in \mathbf{H}(\mathrm{div}\,, \Omega) \times L^2(\Omega) \}, \ \|\xi\|_{\mathcal{X}} = \left(\|\mathbf{q}\|_{\mathbf{H}(\mathrm{div}\,,\Omega)}^2 + \|\psi\|_{0,\Omega}^2 \right)^{1/2}.$$

From now on, we use the notations: $\zeta = (\mathbf{p}, \phi)$ and $\xi = (\mathbf{q}, \psi)$.

4.1. Mixed variational formulation

The solution ϕ to (3.1) belongs to $H^1(\Omega)$, so if one lets $\mathbf{p} = -\mathbb{D}\,\mathbf{grad}\,\phi \in \mathbf{L}^2(\Omega)$, the neutron diffusion problem may be written as:

$$\begin{cases} \text{Find } (\mathbf{p}, \phi) \in \mathbf{H}(\mathrm{div}\,, \Omega) \times H_0^1(\Omega) \text{ such that} \\ -\mathbb{D}^{-1}\,\mathbf{p} - \mathbf{grad}\,\phi = 0 \text{ in } \Omega, \\ \mathrm{div}\,\mathbf{p} + \Sigma_a \phi = S_f \text{ in } \Omega. \end{cases} \tag{4.1}$$

Solving the mixed problem (4.1) is equivalent to solving (3.1).

Proposition 4.1. *Let $\mathbb{D}, \Sigma_a$ satisfy (3.2). The solution $(\mathbf{p}, \phi) \in \mathbf{H}(\mathrm{div}\,, \Omega) \times H_0^1(\Omega)$ to (4.1) is such that ϕ is a solution to (3.1) with the same data. Conversely, the solution $\phi \in H_0^1(\Omega)$ to (3.1) is such that $(-\mathbb{D}\,\mathbf{grad}\,\phi, \phi) \in \mathbf{H}(\mathrm{div}\,, \Omega) \times H_0^1(\Omega)$ is a solution to (4.1) with the same data.*

To obtain the variational formulation for the mixed problem (4.1), let $\mathbf{q} \in \mathbf{H}(\mathrm{div}\,, \Omega)$ and $\psi \in L^2(\Omega)$, multiply the first equation of (4.1) by $\mathbf{q}$, the second equation of (4.1) by $\psi \in L^2(\Omega)$, and integrate over Ω. Adding up the contributions, one finds that:

$$-(\mathbb{D}^{-1}\,\mathbf{p}, \mathbf{q})_{0,\Omega} - (\mathbf{grad}\,\phi, \mathbf{q})_{0,\Omega} + (\psi, \mathrm{div}\,\mathbf{p})_{0,\Omega} + (\Sigma_a \phi, \psi)_{0,\Omega} = (S_f, \psi)_{0,\Omega}. \tag{4.2}$$

One may integrate by parts the second term in the left-hand side, which yields: $-(\mathbf{grad}\,\phi, \mathbf{q})_{0,\Omega} = (\phi, \mathrm{div}\,\mathbf{q})_{0,\Omega}$. We conclude that the solution to (4.1) also solves:

$$\begin{cases} \text{Find } (\mathbf{p}, \phi) \in \mathcal{X} \text{ such that} \\ \forall (\mathbf{q}, \psi) \in \mathcal{X}, \quad -(\mathbb{D}^{-1}\,\mathbf{p}, \mathbf{q})_{0,\Omega} + (\phi, \mathrm{div}\,\mathbf{q})_{0,\Omega} + (\psi, \mathrm{div}\,\mathbf{p})_{0,\Omega} + (\Sigma_a\,\phi, \psi)_{0,\Omega} = (S_f, \psi)_{0,\Omega}. \end{cases} \tag{4.3}$$

Because $\mathbb{D}$ is a symmetric tensor field, the form

$$c \ : \ ((\mathbf{p}, \phi), (\mathbf{q}, \psi)) \mapsto -(\mathbb{D}^{-1}\,\mathbf{p}, \mathbf{q})_{0,\Omega} + (\phi, \mathrm{div}\,\mathbf{q})_{0,\Omega} + (\psi, \mathrm{div}\,\mathbf{p})_{0,\Omega} + (\Sigma_a\,\phi, \psi)_{0,\Omega} \tag{4.4}$$

is continuous, bilinear and symmetric on $\mathbf{H}(\mathrm{div}\,, \Omega) \times L^2(\Omega)$.

We may rewrite the variational formulation (4.3) as:

$$\begin{cases} \text{Find } (\mathbf{p}, \phi) \in \mathbf{H}(\mathrm{div}\,, \Omega) \times L^2(\Omega) \text{ such that} \\ \forall (\mathbf{q}, \psi) \in \mathbf{H}(\mathrm{div}\,, \Omega) \times L^2(\Omega), \quad c((\mathbf{p}, \phi), (\mathbf{q}, \psi)) = (S_f, \psi)_{0,\Omega}. \end{cases} \tag{4.5}$$

Proposition 4.2. *The solution $\zeta = (\mathbf{p}, \phi)$ to (4.5) satisfies (4.1). Hence, problems (4.5) and (4.1) are equivalent.*

One may prove that the mixed formulation (4.5) is well-posed, see Theorem 4.4 in [3]. As a matter of fact, the result is obtained by proving an inf-sup condition in $\mathcal{X}$, which we recall here.

Theorem 4.1. *Let $\mathbb{D}$ and Σ_a satisfy (3.2). Then, the bilinear symmetric form c fulfills an inf-sup condition:*

$$\exists \eta > 0, \quad \inf_{\zeta \in \mathcal{X}} \sup_{\xi \in \mathcal{X}} \frac{c(\zeta, \xi)}{\|\zeta\|_{\mathcal{X}} \, \|\xi\|_{\mathcal{X}}} \geq \eta. \tag{4.6}$$

4.2. Discretization and *a priori* error analysis

We study conforming discretizations of (4.5). Let $(\mathcal{T}_h)_h$ be a family of meshes, made for instance of simplices, or of rectangles ($d = 2$), resp. cuboids ($d = 3$), indexed by a parameter h equal to the largest diameter of elements of a given mesh. We introduce discrete, finite-dimensional, spaces indexed by h as follows:

$$\mathbf{Q}_h \subset \mathbf{H}(\mathrm{div}\,, \Omega), \text{ and } L_h \subset L^2(\Omega).$$

The conforming discretization of the variational formulation (4.5) is then:

$$\begin{cases} \text{Find } (\mathbf{p}_h, \phi_h) \in \mathbf{Q}_h \times L_h \text{ such that} \\ \forall (\mathbf{q}_h, \psi_h) \in \mathbf{Q}_h \times L_h, \quad c((\mathbf{p}_h, \phi_h), (\mathbf{q}_h, \psi_h)) = (S_f, \psi_h)_{0,\Omega}. \end{cases} \tag{4.7}$$

Following Definition 2.14 in [13], we assume that $(\mathbf{Q}_h)_h$, resp. $(L_h)_h$ have the *approximability property* in the sense that

$$\begin{array}{l} \forall \mathbf{q} \in \mathbf{H}(\mathrm{div}\,, \Omega), \ \lim_{h \to 0} \big(\inf_{\mathbf{q}_h \in \mathbf{Q}_h} \|\mathbf{q} - \mathbf{q}_h\|_{\mathbf{H}(\mathrm{div}\,, \Omega)} \big) = 0, \\ \forall \psi \in L^2(\Omega), \ \lim_{h \to 0} (\inf_{\psi_h \in L_h} \|\psi - \psi_h\|_{0,\Omega}) = 0. \end{array} \tag{4.8}$$

We also impose that the space L_h^0 of piecewise constant fields on the mesh is included in L_h, and that $\mathrm{div}\, \mathbf{Q}_h \subset L_h$. We finally define:

$$\mathcal{X}_h = \{ \, \xi_h = (\mathbf{q}_h, \psi_h) \in \mathbf{Q}_h \times L_h \}, \text{ endowed with } \| \cdot \|_{\mathcal{X}}.$$

Remark 4.1. At some point, the discrete spaces are considered locally, *i.e.* restricted to one element of the mesh. So, one introduces the local spaces $\mathbf{Q}_h(K)$, $L_h(K)$, $\mathcal{X}_h(K)$ for every $K \in \mathcal{T}_h$.

Provided the above conditions are fulfilled, one may derive a uniform discrete inf-sup condition under the same assumptions as in Theorem 4.1 (*cf.* [3], Thm. 4.5).

Theorem 4.2. *Let $\mathbb{D} \in \mathcal{P}\mathbb{W}^{1,\infty}(\Omega)$, resp. $\Sigma_a \in \mathcal{P}W^{1,\infty}(\Omega)$, satisfy (3.2). Assume that $(\mathbf{Q}_h)_h$, $(L_h)_h$ fulfill (4.8), $L_h^0 \subset L_h$ and $\mathrm{div}\, \mathbf{Q}_h \subset L_h$ for all h. Then the bilinear form c fulfills a uniform discrete inf-sup condition in $\mathcal{X}_h$.*

$$\exists \, \eta' > 0, \quad \forall h, \quad \inf_{\zeta_h \in \mathcal{X}_h} \sup_{\xi_h \in \mathcal{X}_h} \frac{c(\zeta_h, \xi_h)}{\|\zeta_h\|_{\mathcal{X}} \, \|\xi_h\|_{\mathcal{X}}} \geq \eta'. \tag{4.9}$$

The classical *a priori* error analysis follows. Let $\zeta_h = (\mathbf{p}_h, \phi_h)$ be the solution to (4.7).

Corollary 4.1. *Under the assumptions of Theorem 4.2, there holds:*

$$\exists \, C > 0, \quad \forall h, \quad \|\zeta - \zeta_h\|_{\mathcal{X}_h} \leq C \inf_{\xi_h \in \mathbf{x}_h} \|\zeta - \xi_h\|_{\mathcal{X}_h}. \tag{4.10}$$

Explicit *a priori* error estimates may be derived, see *e.g.* [3].

In this paper, we focus on the Raviart–Thomas–Nédélec (RTN) Finite Element [14, 15].

For *simplicial meshes*, that is meshes made of simplices, the finite element spaces RTN_k can be described as follows, where $k \geq 0$ is the order of the discretization for the scalar fields of L_h, see *e.g.* [16].

The boundary of a simplex $K \in \mathcal{T}_h$ is made of the union of $(d-1)$-simplices, called *facets* from now on, and denoted by $(F_e^K)_{1\leq e\leq d+1}$. We let $\mathbb{P}_k(K)$ be the space of polynomials of maximal degree k on K, resp. $\mathbb{P}_k(F_e^K)$ the space of polynomials of maximal degree k on F_e^K. The definition is

$$\mathrm{RTN}_k(K) = \big\{\mathbf{q} \in \mathbf{L}^2(K) \,|\, \exists \mathbf{a} \in (\mathbb{P}_k(K))^d,\ \exists b \in \mathbb{P}_k(K),\ \forall \mathbf{x} \in K,\ \mathbf{q}(\mathbf{x}) = \mathbf{a} + b\mathbf{x}\big\}.$$

Observe that for all $\mathbf{q} \in \mathrm{RTN}_k(K)$, for all $e \in \{1, \cdots, d+1\}$, $(\mathbf{q}\cdot\mathbf{n})_{|F_e^K} \in \mathbb{P}_k(F_e^K)$. The definitions of the finite element spaces RTN_k are then

$$\mathbf{Q}_h = \big\{\mathbf{q}_h \in \mathbf{H}(\mathrm{div}\,, \Omega) \mid \forall K \in \mathcal{T}_h,\ \mathbf{q}_{h|K} \in \mathrm{RTN}_k(K)\big\},\ L_h = \big\{\psi_h \in L^2(\Omega) \mid \forall K \in \mathcal{T}_h,\ \psi_{h|K} \in \mathbb{P}_k(K)\big\}.$$

For *rectangular or Cartesian meshes*, a description of the Raviart–Thomas–Nédélec (RTN) finite element spaces can be found for instance in Section 4.2 of [12]. We consider those meshes explicitly for the numerical examples, see Section 6.

5. *A POSTERIORI* STUDIES FOR A MIXED FINITE ELEMENT DISCRETIZATION

To develop the study of *a posteriori* estimates, we use the so-called reconstruction of the discrete solution ζ_h. In what follows, we denote by $\tilde{\zeta}_h := \tilde{\zeta}_h(\zeta_h)$ a reconstruction, and by $\eta := \eta(\tilde{\zeta}_h)$ an estimator. Classically, our aim is to obtain *reliable* and *efficient* estimators for the reconstructed error $\zeta - \tilde{\zeta}_h$, meaning that:

$$\begin{aligned} \left\|\zeta - \tilde{\zeta}_h\right\| &\leq \mathtt{C}\,\eta \text{ (reliability)} \\ \eta &\leq \mathtt{c}\left\|\zeta - \tilde{\zeta}_h\right\| \text{ (efficiency)} \end{aligned}$$

where $\mathtt{C}$ and $\mathtt{c}$ are generic constants, and $\|\cdot\|$ is some norm to measure the error. To that aim, we consider that

$$V = H_0^1(\Omega), \text{the original space of solutions, see (3.1)},$$

is the *default* space of scalar reconstructed fields. We also introduce the *broken spaces*

$$H^1(\mathcal{T}_h) = \big\{\psi \in L^2(\Omega) \mid \psi \in H^1(K), \forall K \in \mathcal{T}_h\big\}, \quad \mathbf{H}(\mathrm{div}\,; \mathcal{T}_h) = \big\{\mathbf{q} \in \mathbf{L}^2(\Omega) \mid \mathbf{q} \in \mathbf{H}(\mathrm{div}\,; K), \forall K \in \mathcal{T}_h\big\}.$$

A first approach has been suggested in [17], Chapter 8. The reconstruction $\tilde{\zeta}_h = (\tilde{\mathbf{p}}_h, \tilde{\phi}_h)$ is defined as

$$\begin{aligned} &\tilde{\mathbf{p}}_h = \mathbf{p}_h \in \mathbf{Q}_h \subset \mathbf{H}(\mathrm{div}\,; \Omega), \\ &\tilde{\phi}_h \in V. \end{aligned}$$

Remark 5.1. For other boundary conditions, *i.e.* for a Neumann or Fourier boundary condition, the *default* space V of scalar reconstructed fields would be equal to $H^1(\Omega)$.

In Section 5.1, we recall some reconstruction approaches for RTN finite element spaces. Section 5.2 is devoted to the derivation of *a posteriori* estimates.

5.1. Reconstruction of the discrete solution

In this section, we investigate some approaches to devise a reconstruction of the discrete solution $(\mathbf{p}_h, \phi_h)$, here obtained with the RTN_k finite element discretization, for $k \geq 0$.

For illustrative purposes, we consider simplicial meshes (see Rem. 5.2). Let us introduce some further notations, given such a mesh $\mathcal{T}_h$. The set of facets of $\mathcal{T}_h$ is denoted $\mathcal{F}_h$, and it is split as $\mathcal{F}_h = \mathcal{F}_h^i \cup \mathcal{F}_h^e$, with $\mathcal{F}_h^e$ (resp. $\mathcal{F}_h^i$) being the set of boundary facets (resp. interior facets). We denote by $\mathbb{P}_k(\mathcal{T}_h)$ the space of piecewise polynomials of maximal degree k on each simplex $K \in \mathcal{T}_h$. We let $\mathcal{V}_h^k$ be the set of interpolation points (or nodes) where the degrees of freedom of the V-conforming Lagrange Finite Element space of order k are defined. And, for a node $a \in \mathcal{V}_h^k$, we denote by $\mathcal{T}_a$ the set of simplices K such that $a \in K$.

We recall the definition of the (original) Oswald interpolation operator [18] $\mathcal{I}_{\mathrm{Os}} : \mathbb{P}_k(\mathcal{T}_h) \to \mathbb{P}_k(\mathcal{T}_h) \cap V$ such that

$$\forall \phi_h \in \mathbb{P}_k(\mathcal{T}_h),\ \forall a \in \mathcal{V}_h^k, \quad \mathcal{I}_{\mathrm{Os}}(\phi_h)(a) = \frac{1}{|\mathcal{T}_a|} \sum_{K \in \mathcal{T}_a} \phi_{h|K}(a).$$

A second, modified Oswald operator is defined in [10] as follows. Let

$$W_0(\mathcal{T}_h) = \left\{ \psi_h \in L^2(\mathcal{T}_h) \mid \forall K \in \mathcal{T}_h, \psi_{h|K} \in H^1(K); \forall F \in \mathcal{F}_h^i, \int_F [\psi_h] = 0; \forall F \in \mathcal{F}_h^e, \int_F \psi_h = 0 \right\},$$

where $[\psi_h]|_F = \psi_h|_{K_1}\mathbf{n}_{K_1} + \psi_h|_{K_2}\mathbf{n}_{K_2}$ denotes the jump of ψ_h on the facet $F \in \mathcal{F}_h^i$ shared by elements K_1 and K_2 and $\mathbf{n}_{K_{1,2}}$ is the unit outer normal of the mesh element $K_{1,2} \in \mathcal{T}_h$. Then, the modified Oswald operator[1] $\mathcal{I}_{\mathrm{MO}} : \mathbb{P}_2(\mathcal{T}_h) \cap W_0(\mathcal{T}_h) \to \mathbb{P}_d(\mathcal{T}_h) \cap V$ is such that

$$\begin{aligned} &\forall \phi_h \in \mathbb{P}_2(\mathcal{T}_h) \cap W_0(\mathcal{T}_h),\ \forall a \in \mathcal{V}_h^d, \\ &\mathcal{I}_{\mathrm{MO}}(\phi_h)(a) = \begin{cases} \mathcal{I}_{\mathrm{Os}}(\phi_h)(a) & \text{if } a \text{ is not located at a barycenter of a facet} \\ m(\phi_h, a) & \text{else.} \end{cases} \end{aligned}$$

Above, the values $(m(\phi_h, a_F))_{F \in \mathcal{F}_h}$ at the barycenters of the facets are then defined so that the mean value of $\mathcal{I}_{\mathrm{MO}}(\phi_h)$ on every facet is equal to the mean value of ϕ_h on the same facet.

Remark 5.2. Observe that the results presented in this section can be extended to the case of rectangular or cuboid meshes [7].

5.1.1. *Averaging operator*

We introduce the averaging operator of the neutron flux $\mathcal{I}_{av} : \mathbb{P}_k(\mathcal{T}_h) \to \mathbb{P}_{k+1}(\mathcal{T}_h) \cap V$ such that

$$\forall \phi_h \in \mathbb{P}_k(\mathcal{T}_h),\ \forall a \in \mathcal{V}_h^{k+1}, \quad \mathcal{I}_{av}(\phi_h)(a) = \frac{1}{|\mathcal{T}_a|} \sum_{K \in \mathcal{T}_a} \phi_{h|K}(a).$$

The *average reconstruction* is

$$\tilde{\zeta}_{av,h} = (\mathbf{p}_h, \mathcal{I}_{av}(\phi_h)). \tag{5.1}$$

5.1.2. *Post-processing approaches*

In order to recover the relation $\mathbf{p} = -\mathbb{D}\,\mathbf{grad}\,\phi$ at the discrete level, some post-processing techniques have been introduced for mixed finite element method [7, 10]. The first one is specific to a discretization with the RTN_0 finite element, whereas the second one can be applied to any discretization with a RTN_k finite element, *i.e.* k can be any integer, possibly equal to 0.

[1] Recall that $d = 2$ or $d = 3$.

For $k = 0$, the author in [10] chooses one post-processed scalar variable $\mathcal{I}_{pp}(\mathbf{p}_h, \phi_h) = \widehat{\phi}_h \in \mathbb{P}_2(\mathcal{T}_h)$, which is such that

$$\forall K \in \mathcal{T}_h, \quad -\mathbb{D}_K\Big(\mathbf{grad}\,\widehat{\phi}_h\Big)_{|K} = \mathbf{p}_{h|K}, \; \frac{\Big(\widehat{\phi}_h, 1\Big)_{0,K}}{|K|} = \phi_{h|K}. \tag{5.2}$$

Problems (5.2) are local and independent on each element $K \in \mathcal{T}_h$. We define the RTN_0 post-processing by $\mathcal{I}_{\mathrm{Os}} \circ \mathcal{I}_{pp}$. The *reconstruction* associated to the RTN_0 post-processing is

$$\tilde{\zeta}_{pp,h} = (\mathbf{p}_h, \mathcal{I}_{\mathrm{Os}} \circ \mathcal{I}_{pp}(\mathbf{p}_h, \phi_h)). \tag{5.3}$$

On the other hand, for $k \geq 1$, there exists no solution to Problem (5.2). We present here the approach proposed in [19], valid for $k \geq 0$. It is shown there that the solution to (4.7), $\zeta_h = (\mathbf{p}_h, \phi_h) \in \mathcal{X}_h$, is also equal to the first argument of the solution of a hybrid formulation (see (5.4) below), where the constraint on the continuity of the normal trace of $\mathbf{p}_h$ is relaxed. Let

$$\Lambda_h = \big\{\lambda_h \in L^2\big(\mathcal{F}_h^i\big) \mid \exists \mathbf{q}_h \in \mathbf{Q}_h,\, \lambda_{h|F} = \mathbf{q}_h \cdot \mathbf{n}_{|F},\, \forall F \in \mathcal{F}_h^i\big\},$$

be the space of the Lagrange multipliers and let $\tilde{\mathcal{X}}_h = \Pi_{K \in \mathcal{T}_h} \mathcal{X}_h(K)$ be the unconstrained approximation space with the RTN_k local finite element spaces. By definition, $\mathcal{X}_h$ is a strict subset of $\tilde{\mathcal{X}}_h$.

The hybrid formulation is:

$$\begin{cases} \text{Find } (\zeta_h, \lambda_h) \in \tilde{\mathcal{X}}_h \times \Lambda_h \text{ such that} \\ \forall (\xi_h, \mu_h) \in \tilde{\mathcal{X}}_h \times \Lambda_h, \quad c(\zeta_h, \xi_h) - \sum_{F \in \mathcal{F}_h^i} \int_F \lambda_h [\mathbf{q}_h \cdot \mathbf{n}] + \sum_{F \in \mathcal{F}_h^i} \int_F \mu_h [\mathbf{p}_h \cdot \mathbf{n}] = (S_f, \psi_h)_{0,\Omega}. \end{cases} \tag{5.4}$$

Let $\Pi_{M_h} : \tilde{\mathcal{X}}_h \times \Lambda_h \to M_h$ be the projection onto an appropriate space[2] such that, given $(\zeta_h, \lambda_h) \in \tilde{\mathcal{X}}_h \times \Lambda_h$, its projection $\widehat{\phi}_h = \Pi_{M_h}(\zeta_h, \lambda_h)$ is governed by

$$\forall (\psi_h, \mu_h) \in L_h \times \Lambda_h, \quad \Big(\Sigma_a \widehat{\phi}_h, \psi_h\Big)_{0,\Omega} + \sum_{F \in \mathcal{F}_h^i} \int_F \widehat{\phi}_h \mu_h = (\Sigma_a \phi_h, \psi_h)_{0,\Omega} + \sum_{F \in \mathcal{F}_h^i} \int_F \lambda_h \mu_h.$$

Remark 5.3. In the situation where $\Sigma_a \geq 0$ may vanish, the projection $\widehat{\phi}_h = \Pi_{M_h}(\zeta_h, \lambda_h)$ is defined by

$$\forall (\psi_h, \mu_h) \in L_h \times \Lambda_h, \quad \Big(\Sigma_\star \widehat{\phi}_h, \psi_h\Big)_{0,\Omega} + \sum_{F \in \mathcal{F}_h^i} \int_F \widehat{\phi}_h \mu_h = (\Sigma_\star \phi_h, \psi_h)_{0,\Omega} + \sum_{F \in \mathcal{F}_h^i} \int_F \lambda_h \mu_h,$$

where for all $K \in \mathcal{T}_h$,

$$\Sigma_\star|_K = \begin{cases} \Sigma_a & \text{if } \inf_K \Sigma_a > 0, \\ \sup_K \Sigma_a & \text{if } \inf_K \Sigma_a = 0 \text{ and } \sup_K \Sigma_a > 0, \\ 1 & \text{otherwise.} \end{cases} \tag{5.5}$$

[2] The space M_h is defined here as $M_h = \Pi_{K \in \mathcal{T}_h} M_h(K)$, with for all $K \in \mathcal{T}_h$,

$$M_h(K) = \begin{cases} \Big\{\psi_h \in \mathbb{P}_{k+3}(K) : \psi_h|_{F_e^K} \in \mathbb{P}_{k+1}\big(F_e^K\big) \text{ for } 1 \leq e \leq d+1\Big\} & \text{if } k \text{ is even,} \\ \Big\{\psi_h \in \mathbb{P}_{k+3}(K) : \psi_h|_{F_e^K} \in \mathbb{P}_k\big(F_e^K\big) \oplus \tilde{\mathbb{P}}_{k+2}(F) \text{ for } 1 \leq e \leq d+1\Big\} & \text{if } k \text{ is odd,} \end{cases}$$

where $\tilde{\mathbb{P}}_{k+2}(F)$ denotes the $L^2(F)$-orthogonal complement of $\mathbb{P}_{k+1}(F)$ in $\mathbb{P}_{k+2}(F)$ for any facet $F \in \mathcal{F}_h$. We refer to [19] for the definition of *ad hoc* finite-dimensional spaces M_h for various families and types of elements.

Finally, we refer to [20] for an application of this technique in the field of neutronics. The RTN post-processing is defined here by $\mathcal{I}^2_{\rm RTN} : \tilde{\mathcal{X}}_h \times \Lambda_h \to \mathbb{P}_{k+2}(\mathcal{T}_h) \cap V$ such that

$$\forall(\zeta_h,\lambda_h) \in \tilde{\mathcal{X}}_h \times \Lambda_h,\ \forall a \in \mathcal{V}_h^{k+2},\quad \mathcal{I}^2_{\rm RTN}(\zeta_h,\lambda_h)(a) = \frac{1}{|\mathcal{T}_a|}\sum_{K\in\mathcal{T}_a} (\Pi_{M_h}(\zeta_h,\lambda_h))_{|K}(a).$$

The *reconstruction* associated to the RTN post-processing is

$$\tilde{\zeta}_{{\rm RTN},h} = \big(\mathbf{p}_h, \mathcal{I}^2_{\rm RTN}(\zeta_h,\lambda_h)\big). \tag{5.6}$$

5.1.3. Adding bubbles functions

This section details a possible correction of a reconstruction $\tilde{\zeta}_h = (\mathbf{p}_h, \tilde{\phi}_h)$ to enforce the conservation of local averages, such as (5.14) below. It consists in adding bubble functions ([21], Sect. 3.2.2). The resulting reconstruction with bubble correction is defined as

$$\tilde{\zeta}_{h,\rm bubbles} = \left(\mathbf{p}_h, \tilde{\phi}_h + \sum_{K\in\mathcal{T}_h} \frac{\left(\Sigma_a\left(\phi_h - \tilde{\phi}_h\right), 1\right)_{0,K}}{(\Sigma_a\, b_K, 1)_{0,K}} b_K\right), \tag{5.7}$$

where b_K is the bubble function on $K \in \mathcal{T}_h$ defined as the product of the barycentric coordinates of K.

5.2. *A posteriori* error estimates

We now detail the derivation of *a posteriori* estimates. We define

$$\begin{aligned} d_S(\zeta,\xi) &= \left(\mathbb{D}^{-1}\,\mathbf{p},\mathbf{q}\right)_{0,\Omega} + (\Sigma_a\phi,\psi)_{0,\Omega}\\ d(\zeta,\xi) &= d_S(\zeta,\xi) + (\psi, \operatorname{div}\mathbf{p})_{0,\Omega} - (\phi,\operatorname{div}\mathbf{q})_{0,\Omega} = c(\zeta,(-\mathbf{q},\psi)).\end{aligned}$$

The definition is extended to piecewise smooth fields on $\mathcal{T}_h$ by replacing $\int_\Omega$ by $\sum_{K\in\mathcal{T}_h}\int_K$.

Given $K \in \mathcal{T}_h$, we also define π_0^K the $L^2(K)$-orthogonal projection on the space $L_h^0(K)$, and

$$\begin{aligned} \mathbb{D}_K^{\max} &= \sup_{\mathbf{q}\in\mathbf{L}^2(K)\setminus\{0\}} \frac{(\mathbb{D}\,\mathbf{q},\mathbf{q})_{0,K}}{\|\mathbf{q}\|^2_{0,K}}, \quad \mathbb{D}_K^{\min} = \inf_{\mathbf{q}\in\mathbf{L}^2(K)\setminus\{0\}} \frac{(\mathbb{D}\,\mathbf{q},\mathbf{q})_{0,K}}{\|\mathbf{q}\|^2_{0,K}},\\ \Sigma_{a,K}^{\max} &= \sup_{\psi\in L^2(K)\setminus\{0\}} \frac{(\Sigma_a\psi,\psi)_{0,K}}{\|\psi\|^2_{0,K}}, \quad \Sigma_{a,K}^{\min} = \inf_{\psi\in L^2(K)\setminus\{0\}} \frac{(\Sigma_a\psi,\psi)_{0,K}}{\|\psi\|^2_{0,K}}.\end{aligned}$$

In order to state the estimates, at some point we will use the following assumptions.

Assumption 5.1. *The coefficients $\mathbb{D}$, Σ_a are piecewise constant on $\mathcal{T}_h$, and $S_f \in L_h$.*

Assumption 5.2. *The coefficients $\mathbb{D}^{-1}$, Σ_a are piecewise polynomials on $\mathcal{T}_h$, and $S_f \in L_h$.*

Finally, we recall that there exists $C_{P,d} > 0$, the so-called Poincaré constant (see *e.g.* Eq. (2.1) in [10]), such that, for all h, for all $K \in \mathcal{T}_h$ and for all $\varphi \in H^1(K)$, it holds that

$$\left\|\varphi - \pi_0^K\varphi\right\|_{0,K} \le C_{P,d}\, h_K\, \|\mathbf{grad}\,\varphi\|_{0,K}. \tag{5.8}$$

Note that $C_{P,d} = \frac{1}{\pi}$ in the case where the considered mesh elements are convex; *cf.* [22,23].

In Section 5.2.1, we recall a classical *a posteriori* error framework in the primal setting (unknown ϕ), where the error is measured in $H^1(\mathcal{T}_h)$ norm. We propose two alternatives in the mixed setting (unknown $(\mathbf{p},\phi)$): for the first one we measure the error with respect to the $\mathcal{H}$ norm, while for the second one we use a weighted $\mathbf{H}(\operatorname{div};\mathcal{T}_h) \times L^2(\Omega)$ norm. Both approaches are respectively developed in Sections 5.2.2 and 5.2.3.

5.2.1. Estimates in H^1 norm

In this section, we aim to briefly recall the *a posteriori* error framework introduced in [10]. The energy norm associated to the primal form is

$$|||\phi|||_p^2 = \left\|\mathbb{D}^{1/2}\mathbf{grad}\,\phi\right\|_{0,\Omega}^2 + \left\|\Sigma_a^{1/2}\phi\right\|_{0,\Omega}^2.$$

Therefore, we define the broken norm

$$|||\psi|||_{p,\mathcal{T}_h}^2 = \sum_{K\in\mathcal{T}_h} |||\psi|||_{p,K}^2, \text{ where } |||\psi|||_{p,K}^2 = \left\|\mathbb{D}^{1/2}\mathbf{grad}\,\psi\right\|_{0,K}^2 + \left\|\Sigma_a^{1/2}\psi\right\|_{0,K}^2. \tag{5.9}$$

We recall the following *a posteriori* error estimate ([10], Thm. 4.2, p. 1578).

Theorem 5.1. *Let ζ and ζ_h be respectively the solution to (4.5) and (4.7) with RTN_0 finite elements, and let $\widehat{\phi}_h = \mathcal{I}_{pp}(\phi_h)$. For all $K\in\mathcal{T}_h$, we define the residual estimator*

$$\tilde{\eta}_{r,K} = m_K\left\|S_f + \operatorname{div}\left(\mathbb{D}\,\mathbf{grad}\,\widehat{\phi}_h\right) - \Sigma_a\widehat{\phi}_h\right\|_{0,K}, \quad \text{with } m_K = \min\left\{\frac{C_{P,d}\,h_K}{\left(\mathbb{D}_K^{\min}\right)^{1/2}}, \frac{1}{\left(\Sigma_{a,K}\right)^{1/2}}\right\}, \tag{5.10}$$

and the nonconformity estimator

$$\tilde{\eta}_{nc,K} = \left|\left|\left|\widehat{\phi}_h - \mathcal{I}_{\mathrm{MO}}\left(\widehat{\phi}_h\right)\right|\right|\right|_{p,K}.$$

Then, under Assumption 5.1, one has the reliability estimate

$$\left|\left|\left|\phi - \widehat{\phi}_h\right|\right|\right|_{p,\mathcal{T}_h} \le \left\{\sum_{K\in\mathcal{T}_h}\tilde{\eta}_{nc,K}^2\right\}^{1/2} + \left\{\sum_{K\in\mathcal{T}_h}\tilde{\eta}_{r,K}^2\right\}^{1/2}. \tag{5.11}$$

The following theorem states the local efficiency of the residual estimator ([10], Thm. 4.4, pp. 1578, 1579).

Theorem 5.2 (Local efficiency of the *a posteriori* error estimators)**.** *Let ζ and ζ_h be respectively the solution to (4.5) and (4.7) with RTN_0 finite elements, and let $\widehat{\phi}_h = \mathcal{I}_{pp}(\phi_h)$. Under Assumption 5.1, there holds on every $K\in\mathcal{T}_h$*

$$\tilde{\eta}_{r,K} \le \mathtt{C}\left(\frac{\mathbb{D}_K^{\max}}{\mathbb{D}_K^{\min}}\right)^{1/2}\left|\left|\left|\phi - \widehat{\phi}_h\right|\right|\right|_{p,K}, \tag{5.12}$$

with the constant $\mathtt{C}$ depending only on the polynomial degree k of S_f, the space dimension d, and the shape-regularity parameter $\kappa_K = |K|/h_K^d$.

5.2.2. Estimates in $\mathcal{H}$ norm

In this section, we use the broken norm associated to the bilinear form d_S, *i.e.*

$$|||\xi|||_{\mathcal{T}_h}^2 = \sum_{K\in\mathcal{T}_h}|||\xi|||_K^2, \text{ where } |||\xi|||_K^2 = \left\|\mathbb{D}^{-1/2}\mathbf{q}\right\|_{0,K}^2 + \left\|\Sigma_a^{1/2}\psi\right\|_{0,K}^2. \tag{5.13}$$

We note that, according to assumption (3.2) on $\mathbb{D}$ and Σ_a, $|||\cdot|||_{\mathcal{T}_h}$ and $\|\cdot\|_{\mathcal{H}}$ define equivalent norms on $\mathcal{H}$.

Lemma 5.1. *Let $\xi,\chi,\zeta\in\mathcal{H}$, we have the following estimate*

$$|||\zeta-\xi|||_{\mathcal{T}_h} \le |||\xi-\chi|||_{\mathcal{T}_h} + \left|d_S\left(\zeta-\xi, \frac{\zeta-\chi}{|||\zeta-\chi|||_{\mathcal{T}_h}}\right)\right|.$$

Proof. We follow the proof given in [7], Theorem 3.1. We first assume that $|||\zeta-\xi|||_{\mathcal{T}_h} \le |||\zeta-\chi|||_{\mathcal{T}_h}$. We have that

$$\begin{aligned}|||\zeta-\chi|||^2_{\mathcal{T}_h} &= d_S(\zeta-\chi,\zeta-\chi) = d_S(\zeta-\xi,\zeta-\chi) + d_S(\xi-\chi,\zeta-\chi)\\ &\le |||\zeta-\chi|||_{\mathcal{T}_h} d_S\left(\zeta-\xi, \frac{\zeta-\chi}{|||\zeta-\chi|||_{\mathcal{T}_h}}\right) + |||\xi-\chi|||_{\mathcal{T}_h}|||\zeta-\chi|||_{\mathcal{T}_h}.\end{aligned}$$

Using the assumption, we infer the estimate. Second, we assume that $|||\zeta-\chi|||_{\mathcal{T}_h} \le |||\zeta-\xi|||_{\mathcal{T}_h}$. We then have,

$$\begin{aligned}|||\zeta-\xi|||^2_{\mathcal{T}_h} &= d_S(\zeta-\xi,\zeta-\xi) = d_S(\zeta-\xi,\zeta-\chi) + d_S(\zeta-\xi,\chi-\xi)\\ &\le |||\zeta-\chi|||_{\mathcal{T}_h} d_S\left(\zeta-\xi, \frac{\zeta-\chi}{|||\zeta-\chi|||_{\mathcal{T}_h}}\right) + |||\xi-\chi|||_{\mathcal{T}_h}|||\zeta-\xi|||_{\mathcal{T}_h}\\ &\le |||\zeta-\xi|||_{\mathcal{T}_h} d_S\left(\zeta-\xi, \frac{\zeta-\chi}{|||\zeta-\chi|||_{\mathcal{T}_h}}\right) + |||\xi-\chi|||_{\mathcal{T}_h}|||\zeta-\xi|||_{\mathcal{T}_h}.\end{aligned}$$

This concludes the proof. $\square$

Theorem 5.3. *Let ζ and ζ_h be respectively the solution to (4.5) and (4.7). Let $\tilde{\zeta}_h = (\mathbf{p}_h, \tilde{\phi}_h) \in \mathbf{Q}_h \times V$ be a reconstruction of $\zeta_h = (\mathbf{p}_h, \phi_h)$ such that, for all $K \in \mathcal{T}_h$,*

$$\left(\Sigma_a\tilde{\phi}_h, 1\right)_{0,K} = (\Sigma_a\phi_h, 1)_{0,K}. \tag{5.14}$$

For any $K \in \mathcal{T}_h$, we define the residual estimator

$$\overline{\eta}_{r,K} = \overline{m}_K\, \eta_{r,K}, \tag{5.15}$$

where

$$\eta_{r,K} = \left\|\Sigma_a^{-1/2}\left(S_f - \operatorname{div}\mathbf{p}_h - \Sigma_a\tilde{\phi}_h\right)\right\|_{0,K} \quad \textit{and} \quad \overline{m}_K = \min\left\{1, \frac{C_{P,d}h_K\left(\Sigma_{a,K}^{\max}\right)^{1/2}}{\left(\mathbb{D}_K^{\min}\right)^{1/2}}\right\}, \tag{5.16}$$

and the flux estimator

$$\eta_{f,K} = \left\|\mathbb{D}^{1/2}\left(\mathbb{D}^{-1}\mathbf{p}_h + \mathbf{grad}\,\tilde{\phi}_h\right)\right\|_{0,K}. \tag{5.17}$$

One has the reliability estimate

$$\left|\left|\left|\zeta-\tilde{\zeta}_h\right|\right|\right|_{\mathcal{T}_h} \le \left(\sum_{K\in\mathcal{T}_h}\overline{\eta}^2_{r,K}\right)^{1/2} + \left(\sum_{K\in\mathcal{T}_h}\eta^2_{f,K}\right)^{1/2}. \tag{5.18}$$

Proof. Using Lemma 5.1, we have, for all $\chi \in \mathcal{H}$,

$$\left|\left|\left|\zeta-\tilde{\zeta}_h\right|\right|\right|_{\mathcal{T}_h} \le \left|\left|\left|\tilde{\zeta}_h-\chi\right|\right|\right|_{\mathcal{T}_h} + \left|d_S\left(\zeta-\tilde{\zeta}_h, \frac{\zeta-\chi}{|||\zeta-\chi|||_{\mathcal{T}_h}}\right)\right|. \tag{5.19}$$

For any $\psi \in V$, let $\chi = (-\mathbb{D}\,\mathbf{grad}\,\psi, \psi) \in \mathcal{H}$. One has, by symmetry of $\mathbb{D}$ and according to (3.3), that

$$\begin{aligned}d_S(\zeta,\zeta-\chi) &= \left(\mathbb{D}^{-1}\mathbf{p}, \mathbf{p}+\mathbb{D}\,\mathbf{grad}\,\psi\right)_{0,\Omega} + (\Sigma_a\phi, \phi-\psi)_{0,\Omega}\\ &= (\mathbf{grad}\,\phi, \mathbb{D}\,\mathbf{grad}\,(\phi-\psi))_{0,\Omega} + (\Sigma_a\phi,\phi-\psi)_{0,\Omega}\\ &= (S_f, \phi-\psi)_{0,\Omega}.\end{aligned}$$

So it follows that

$$\begin{aligned} d_S\Big(\zeta-\tilde{\zeta}_h,\zeta-\chi\Big) &= \sum_{K\in\mathcal{T}_h}(S_f,\phi-\psi)_{0,K} - d_S\Big(\tilde{\zeta}_h,\zeta-\chi\Big) \\ &= \sum_{K\in\mathcal{T}_h}(S_f,\phi-\psi)_{0,K} - \big(\mathbb{D}^{-1}\mathbf{p}_h,-\mathbb{D}\,\mathbf{grad}\,(\phi-\psi)\big)_{0,K} - \Big(\Sigma_a\tilde{\phi}_h,\phi-\psi\Big)_{0,K}. \end{aligned}$$

Owing to the fact that $\phi-\psi\in V$ and the symmetry of $\mathbb{D}$, we can integrate by parts the second term

$$\begin{aligned} d_S\Big(\zeta-\tilde{\zeta}_h,\zeta-\chi\Big) &= \sum_{K\in\mathcal{T}_h}\Big(S_f-\operatorname{div}\mathbf{p}_h-\Sigma_a\tilde{\phi}_h,\phi-\psi\Big)_{0,K}; \\ \text{so } \Big|d_S\Big(\zeta-\tilde{\zeta}_h,\zeta-\chi\Big)\Big| &\le \sum_{K\in\mathcal{T}_h}\eta_{r,K}|||\zeta-\chi|||_K. \end{aligned} \tag{5.20}$$

We also obtain

$$\begin{aligned} d_S\Big(\zeta-\tilde{\zeta}_h,\zeta-\chi\Big) &= \sum_{K\in\mathcal{T}_h}\Big(S_f-\operatorname{div}\mathbf{p}_h-\Sigma_a\tilde{\phi}_h,(\phi-\psi)\Big)_{0,K} \\ &= \sum_{K\in\mathcal{T}_h}\Big(S_f-\operatorname{div}\mathbf{p}_h-\Sigma_a\tilde{\phi}_h,(\phi-\psi)-\pi_0^K(\phi-\psi)\Big)_{0,K}; \\ \text{so } \Big|d_S\Big(\zeta-\tilde{\zeta}_h,\zeta-\chi\Big)\Big| &\le \sum_{K\in\mathcal{T}_h}\Big\|S_f-\operatorname{div}\mathbf{p}_h-\Sigma_a\tilde{\phi}_h\Big\|_{0,K}C_{P,d}h_K||\mathbf{grad}\,(\phi-\psi)||_{0,K} \\ &\le \sum_{K\in\mathcal{T}_h}\eta_{r,K}\frac{(\Sigma_{a,K}^{\max})^{1/2}C_{P,d}h_K}{\big(\mathbb{D}_K^{\min}\big)^{1/2}}|||\zeta-\chi|||_K, \end{aligned} \tag{5.21}$$

where we used (5.14) and (4.7) in the first line, and we applied the Poincaré inequality (5.8) in the second line. Collecting (5.20) and (5.21) and using the definition of $\overline{\eta}_{r,K}$, we find

$$\Big|d_S\Big(\zeta-\tilde{\zeta}_h,\zeta-\chi\Big)\Big| \le \sum_{K\in\mathcal{T}_h}\overline{\eta}_{r,K}|||\zeta-\chi|||_K.$$

Next, using (5.19) and the Cauchy–Schwarz inequality, we get

$$\Big|\Big|\Big|\zeta-\tilde{\zeta}_h\Big|\Big|\Big|_{\mathcal{T}_h} \le \Big|\Big|\Big|\tilde{\zeta}_h-\chi\Big|\Big|\Big|_{\mathcal{T}_h} + \left(\sum_{K\in\mathcal{T}_h}\overline{\eta}_{r,K}^2\right)^{1/2}.$$

We conclude the proof by choosing $\chi=(-\mathbb{D}\,\mathbf{grad}\,\tilde{\phi}_h,\tilde{\phi}_h)\in\mathcal{H}$ and using the definition of $\eta_{f,K}$. $\square$

Remark 5.4. We recall the illuminating Prager–Synge theorem [24] which states that, given $\phi\in V$ the weak solution of (3.3), $\tilde{\phi}_h\in V$ and $\mathbf{p}_h\in\mathbf{H}(\operatorname{div},\Omega)$ with $\operatorname{div}\mathbf{p}_h+\Sigma_a\,\tilde{\phi}_h=S_f$ arbitrary, then one has the equality

$$\begin{aligned} &\Big\|\mathbb{D}^{1/2}\mathbf{grad}\,\Big(\phi-\tilde{\phi}_h\Big)\Big\|_{0,\Omega}^2 + 2\Big\|\Sigma_a^{1/2}\Big(\phi-\tilde{\phi}_h\Big)\Big\|_{0,\Omega}^2 + \Big\|\mathbb{D}^{1/2}\mathbf{grad}\,\phi+\mathbb{D}^{-1/2}\mathbf{p}_h\Big\|_{0,\Omega}^2 \\ &\quad = \Big\|\mathbb{D}^{1/2}\mathbf{grad}\,\tilde{\phi}_h+\mathbb{D}^{-1/2}\mathbf{p}_h\Big\|_{0,\Omega}^2. \end{aligned}$$

This result yields that, if one chooses a reconstruction that is $\mathbf{H}(\operatorname{div},\Omega)\times V$-conforming, one can derive simultaneous reliability and efficiency in a straightforward manner.

This approach may also be applied to the *primal* energy norm. In order to state the following theorem, we introduce the bilinear form associated to the broken norm $|||\cdot|||_{p,\mathcal{T}_h}$ (see (5.9)),

$$d_p(\phi,\psi) = \sum_{K\in\mathcal{T}_h} (\mathbb{D}\,\mathbf{grad}\,\phi, \mathbf{grad}\,\psi)_{0,K} + (\Sigma_a\phi,\psi)_{0,K}.$$

Theorem 5.4. *Under the assumptions of Theorem 5.3, one has the reliability estimate*

$$\left|\left|\left|\phi-\tilde{\phi}_h\right|\right|\right|_{p,\mathcal{T}_h} \leq \left(\sum_{K\in\mathcal{T}_h}\left(\overline{\eta}_{r,K}+\eta_{f,K}\right)^2\right)^{1/2}. \tag{5.22}$$

Proof. Recall that $\tilde{\phi}_h \in V$. Since ϕ solves (3.3), one has that

$$\begin{aligned} d_p\left(\phi,\phi-\tilde{\phi}_h\right) &= \left(\mathbb{D}\,\mathbf{grad}\,\phi, \mathbf{grad}\left(\phi-\tilde{\phi}_h\right)\right)_{0,\Omega} + \left(\Sigma_a\phi,\phi-\tilde{\phi}_h\right)_{0,\Omega} \\ &= \left(S_f,\phi-\tilde{\phi}_h\right)_{0,\Omega}. \end{aligned}$$

Then, we have

$$\begin{aligned} \left|\left|\left|\phi-\tilde{\phi}_h\right|\right|\right|^2_{p,\mathcal{T}_h} &= d_p\left(\phi-\tilde{\phi}_h,\phi-\tilde{\phi}_h\right) \\ &= \sum_{K\in\mathcal{T}_h}\left(S_f,\phi-\tilde{\phi}_h\right)_{0,K} - d_p\left(\tilde{\phi}_h,\phi-\tilde{\phi}_h\right) \\ &= \sum_{K\in\mathcal{T}_h}\left(S_f,\phi-\tilde{\phi}_h\right)_{0,K} - \left(\mathbb{D}\,\mathbf{grad}\,\tilde{\phi}_h,\mathbf{grad}\left(\phi-\tilde{\phi}_h\right)\right)_{0,K} - \left(\Sigma_a\tilde{\phi}_h,\phi-\tilde{\phi}_h\right)_{0,K}. \\ &= \sum_{K\in\mathcal{T}_h}\left(S_f,\phi-\tilde{\phi}_h\right)_{0,K} + \left(\mathbf{p}_h,\mathbf{grad}\left(\phi-\tilde{\phi}_h\right)\right)_{0,K} - \left(\Sigma_a\tilde{\phi}_h,\phi-\tilde{\phi}_h\right)_{0,K} \\ &\quad - \left(\mathbf{p}_h+\mathbb{D}\,\mathbf{grad}\,\tilde{\phi}_h,\mathbf{grad}\left(\phi-\tilde{\phi}_h\right)\right)_{0,K} \\ &= \sum_{K\in\mathcal{T}_h}\left(S_f-\operatorname{div}\mathbf{p}_h-\Sigma_a\tilde{\phi}_h,\phi-\tilde{\phi}_h\right)_{0,K} - \left(\mathbf{p}_h+\mathbb{D}\,\mathbf{grad}\,\tilde{\phi}_h,\mathbf{grad}\left(\phi-\tilde{\phi}_h\right)\right)_{0,K}. \end{aligned}$$

To reach the last line, we integrated by parts the second term.

By the symmetry of $\mathbb{D}$ and the Cauchy–Schwarz inequality, we find for the second term

$$\left|\sum_{K\in\mathcal{T}_h}\left(\mathbf{p}_h+\mathbb{D}\,\mathbf{grad}\,\tilde{\phi}_h,\mathbf{grad}\left(\phi-\tilde{\phi}_h\right)\right)_{0,K}\right| \leq \sum_{K\in\mathcal{T}_h}\eta_{f,K}\left\|\mathbb{D}^{1/2}\mathbf{grad}\left(\phi-\tilde{\phi}_h\right)\right\|_{0,K}. \tag{5.23}$$

On the other hand, for the first term, one has by the Cauchy–Schwarz inequality

$$\left|\sum_{K\in\mathcal{T}_h}\left(S_f-\operatorname{div}\mathbf{p}_h-\Sigma_a\tilde{\phi}_h,\phi-\tilde{\phi}_h\right)_{0,K}\right| \leq \sum_{K\in\mathcal{T}_h}\eta_{r,K}\left\|\Sigma_a^{1/2}\left(\phi-\tilde{\phi}_h\right)\right\|_{0,K}. \tag{5.24}$$

We may also use (5.14) to write

$$\sum_{K\in\mathcal{T}_h}\left(S_f-\operatorname{div}\mathbf{p}_h-\Sigma_a\tilde{\phi}_h,\phi-\tilde{\phi}_h\right)_{0,K} = \sum_{K\in\mathcal{T}_h}\left(S_f-\operatorname{div}\mathbf{p}_h-\Sigma_a\tilde{\phi}_h,\left(\phi-\tilde{\phi}_h\right)-\pi_0^K\left(\phi-\tilde{\phi}_h\right)\right)_{0,K};$$

so $$\left|\sum_{K\in\mathcal{T}_h}\left(S_f - \operatorname{div}\mathbf{p}_h - \Sigma_a\tilde{\phi}_h, \phi - \tilde{\phi}_h\right)_{0,K}\right| \le \sum_{K\in\mathcal{T}_h}\left\|S_f - \operatorname{div}\mathbf{p}_h - \Sigma_a\tilde{\phi}_h\right\|_{0,K} C_{P,d}h_K\left\|\mathbf{grad}\left(\phi-\tilde{\phi}_h\right)\right\|_{0,K}$$
$$\le \sum_{K\in\mathcal{T}_h}\eta_{r,K}\frac{\left(\Sigma_{a,K}^{\max}\right)^{1/2}C_{P,d}h_K}{\left(\mathbb{D}_K^{\min}\right)^{1/2}}\left\|\mathbb{D}^{1/2}\mathbf{grad}\left(\phi-\tilde{\phi}_h\right)\right\|_{0,K}. \tag{5.25}$$

By definition (see (5.9)), we know that $\|\Sigma_a^{1/2}(\phi-\tilde{\phi}_h)\|_{0,K} \le |||\phi-\tilde{\phi}_h|||_{p,K}$ and $\|\mathbb{D}^{1/2}\mathbf{grad}\,(\phi-\tilde{\phi}_h)\|_{0,K} \le |||\phi-\tilde{\phi}_h|||_{p,K}$. For the first term, collecting (5.24) and (5.25), we find that

$$\left|\sum_{K\in\mathcal{T}_h}\left(S_f - \operatorname{div}\mathbf{p}_h - \Sigma_a\tilde{\phi}_h, \phi - \tilde{\phi}_h\right)_{0,K}\right| \le \sum_{K\in\mathcal{T}_h}\overline{\eta}_{r,K}\left|\left|\left|\phi-\tilde{\phi}_h\right|\right|\right|_{p,K}. \tag{5.26}$$

And, with the help of (5.23), we get

$$\left|\left|\left|\phi-\tilde{\phi}_h\right|\right|\right|_{p,\mathcal{T}_h}^2 \le \sum_{K\in\mathcal{T}_h}\left(\overline{\eta}_{r,K}+\eta_{f,K}\right)\left|\left|\left|\phi-\tilde{\phi}_h\right|\right|\right|_{p,K},$$

which leads to the conclusion (5.22) by applying the Cauchy–Schwarz inequality one last time. □

Theorem 5.5 (Local efficiency of the *a posteriori* error estimators)**.** *Let Assumption 5.2 be fulfilled. For $K\in\mathcal{T}_h$, let $\overline{\eta}_{r,K}$ and $\eta_{f,K}$ be the residual and flux estimators respectively given by (5.15), and (5.17). In addition, we suppose that $\tilde{\phi}_h$ is piecewise polynomial on $\mathcal{T}_h$. The following estimates hold true*

$$\overline{\eta}_{r,K} \le \left(\frac{\Sigma_{a,K}^{\max}}{\Sigma_{a,K}^{\min}}\right)^{1/2}\left(\mathtt{c}\frac{\mathbb{D}_K^{\max}}{\mathbb{D}_K^{\min}}+\mathtt{C}\right)^{1/2}\left|\left|\left|\zeta-\tilde{\zeta}_h\right|\right|\right|_K \tag{5.27}$$
$$\eta_{f,K} \le \left|\left|\left|\zeta-\tilde{\zeta}_h\right|\right|\right|_K + \left|\left|\left|\phi-\tilde{\phi}_h\right|\right|\right|_{p,K}, \tag{5.28}$$

where $\mathtt{c}$ and $\mathtt{C}$ are constants which depend only on the polynomial degree of S_f, Σ_a and $\tilde{\phi}_h$, d, and on the shape-regularity parameter κ_K.

Proof. The proof follows that given in [10], Lemma 7.6. Let ψ_K be the bubble function on K, given as the product of the $d+1$ linear functions that take the value 1 at one vertex of K and vanish at the other vertices. Let $\psi_r = (S_f - \operatorname{div}\mathbf{p}_h - \Sigma_a\tilde{\phi}_h)$. Note that ψ_r is a polynomial in K, because each term appearing in its definition is a polynomial (thanks to Assumption 5.2 for S_f and Σ_a). Then the equivalence of norms on finite-dimensional spaces, the definition of ψ_K and the inverse inequality (*cf.*, *e.g.* [25], Thm. 3.2.6) respectively give

$$c\|\psi_r\|_{0,K}^2 \le (\psi_r, \psi_K\psi_r)_{0,K}, \tag{5.29}$$
$$\|\psi_K\psi_r\|_{0,K} \le \|\psi_r\|_{0,K}, \tag{5.30}$$
$$\|\mathbf{grad}\,(\psi_K\psi_r)\|_{0,K} \le C\,h_K^{-1}\|\psi_K\psi_r\|_{0,K}, \tag{5.31}$$

with the constants c and C depending only on the polynomial degree of S_f, Σ_a and $\tilde{\phi}_h$, d, and κ_K.

Let $\xi_{r,K} = (0, \psi_K\psi_r)$ in K, and 0 elsewhere: by construction, $\xi_{r,K}\in\mathcal{X}$. Then we have, by the definition of the bilinear form d and of ζ,

$$d\left(\zeta-\tilde{\zeta}_h, \xi_{r,K}\right) = d(\zeta, \xi_{r,K}) - d\left(\tilde{\zeta}_h, \xi_{r,K}\right)$$
$$= (S_f, \psi_K\psi_r)_{0,\Omega} - \left(\Sigma_a\tilde{\phi}_h, \psi_K\psi_r\right)_{0,\Omega} - (\operatorname{div}\mathbf{p}_h, \psi_K\psi_r)_{0,\Omega}$$

$$= (\psi_r, \psi_K \psi_r)_{0,K}.$$

On the other hand,

$$\begin{aligned}
d\Big(\zeta - \tilde{\zeta}_h, \xi_{r,K}\Big) &= \Big(\Sigma_a\Big(\phi - \tilde{\phi}_h\Big), \psi_K\psi_r\Big)_{0,\Omega} + (\operatorname{div}(\mathbf{p}-\mathbf{p}_h), \psi_K\psi_r)_{0,\Omega} \\
&= \Big(\Sigma_a\Big(\phi - \tilde{\phi}_h\Big), \psi_K\psi_r\Big)_{0,\Omega} - (\mathbf{p}-\mathbf{p}_h, \mathbf{grad}\,(\psi_K\psi_r))_{0,K} \\
&\le \Big|\Big|\Big|\zeta - \tilde{\zeta}_h\Big|\Big|\Big|_K \left(\Big\|\mathbb{D}^{1/2}\mathbf{grad}\,(\psi_K\psi_r)\Big\|_{0,K}^2 + \Big\|\Sigma_a^{1/2}\psi_K\psi_r\Big\|_{0,K}^2\right)^{1/2},
\end{aligned} \tag{5.32}$$

where we integrated by parts the second term in the second line. Combining (5.29)–(5.32), one comes to

$$c\|\psi_r\|_{0,K}^2 \le \Big|\Big|\Big|\zeta - \tilde{\zeta}_h\Big|\Big|\Big|_K \|\psi_r\|_{0,K}\big(C^2\mathbb{D}_K^{\max}h_K^{-2} + \Sigma_{a,K}^{\max}\big)^{1/2}.$$

Considering the definition (5.15) of $\overline{\eta}_{r,K}$, we infer that

$$\overline{\eta}_{r,K} \le \overline{m}_K\big(\Sigma_{a,K}^{\min}\big)^{-1/2}\|\psi_r\|_{0,K} \le \Big|\Big|\Big|\zeta - \tilde{\zeta}_h\Big|\Big|\Big|_K c^{-1}\left(\frac{\Sigma_{a,K}^{\max}}{\Sigma_{a,K}^{\min}}\right)^{1/2}\overline{m}_K\Big(C^2\mathbb{D}_K^{\max}h_K^{-2}\big(\Sigma_{a,K}^{\max}\big)^{-1} + 1\Big)^{1/2}.$$

If $1 < \dfrac{C_{P,d}h_K(\Sigma_{a,K}^{\max})^{1/2}}{\big(\mathbb{D}_K^{\min}\big)^{1/2}}$, we have $\overline{m}_K = 1$ and

$$\overline{m}_K\big(C^2\mathbb{D}_K^{\max}h_K^{-2}(\Sigma_{a,K}^{\max})^{-1} + 1\big)^{1/2} \le \left(C^2C_{P,d}^2\frac{\mathbb{D}_K^{\max}}{\mathbb{D}_K^{\min}} + 1\right)^{1/2}.$$

Otherwise, we have $\overline{m}_K = \dfrac{C_{P,d}h_K(\Sigma_{a,K}^{\max})^{1/2}}{(\mathbb{D}_K^{\min})^{1/2}}$ and

$$\begin{aligned}
\overline{m}_K\big(C^2\mathbb{D}_K^{\max}h_K^{-2}(\Sigma_{a,K}^{\max})^{-1} + 1\big)^{1/2} &= \frac{C_{P,d}h_K(\Sigma_{a,K}^{\max})^{1/2}}{\big(\mathbb{D}_K^{\min}\big)^{1/2}}\Big(C^2\mathbb{D}_K^{\max}h_K^{-2}\big(\Sigma_{a,K}^{\max}\big)^{-1} + 1\Big)^{1/2} \\
&= \left(C^2C_{P,d}^2\frac{\mathbb{D}_K^{\max}}{\mathbb{D}_K^{\min}} + \frac{C_{P,d}^2h_K^2(\Sigma_{a,K}^{\max})}{\big(\mathbb{D}_K^{\min}\big)}\right)^{1/2} \\
&\le \left(C^2C_{P,d}^2\frac{\mathbb{D}_K^{\max}}{\mathbb{D}_K^{\min}} + 1\right)^{1/2}.
\end{aligned}$$

This concludes the proof of (5.27).

We now proceed with the triangle inequality for the second estimate,

$$\begin{aligned}
\Big\|\mathbb{D}^{1/2}\Big(\mathbb{D}^{-1}\mathbf{p}_h + \mathbf{grad}\,\tilde{\phi}_h\Big)\Big\|_{0,K} &\le \Big\|\mathbb{D}^{-1/2}\mathbf{p}_h - \mathbb{D}^{-1/2}\mathbf{p}\Big\|_{0,K} + \Big\|\mathbb{D}^{-1/2}\mathbf{p} + \mathbb{D}^{1/2}\mathbf{grad}\,\tilde{\phi}_h\Big\|_{0,K} \\
&\le \Big\|\mathbb{D}^{-1/2}(\mathbf{p}_h - \mathbf{p})\Big\|_{0,K} + \Big\|\mathbb{D}^{1/2}\Big(\mathbf{grad}\,\phi - \mathbf{grad}\,\tilde{\phi}_h\Big)\Big\|_{0,K}.
\end{aligned}$$

Considering the definition of $\eta_{f,K}$ by (5.17) concludes the proof. □

Remark 5.5. Assume in addition in Theorem 5.5 that there exists a constant $\kappa > 0$, such that $\min_{K\in\mathcal{T}_h}\kappa_K \ge \kappa$, for all $h > 0$. Then, the constants $\mathtt{c}$ and $\mathtt{C}$ do not depend on κ_K (but on κ).

Remark 5.6. The results of this section extend with the same arguments to the situation where $\Sigma_a \geq 0$ may vanish. Under the assumptions of Theorem 5.3, one has the reliability estimates

$$\left|\left|\left|\zeta - \tilde{\zeta}_h\right|\right|\right|_{\mathcal{T}_h} \leq \left(\sum_{K\in\mathcal{T}_h} \overline{\eta}^2_{r,K}\right)^{1/2} + \left(\sum_{K\in\mathcal{T}_h} \eta^2_{f,K}\right)^{1/2},$$

$$\left|\left|\left|\phi - \tilde{\phi}_h\right|\right|\right|_{p,\mathcal{T}_h} \leq \left(\sum_{K\in\mathcal{T}_h} \left(\overline{\eta}_{r,K} + \eta_{f,K}\right)^2\right)^{1/2},$$

where the residual estimator $\overline{\eta}_{r,K} = \eta_{r,K}\overline{m}_K$ with

$$\eta_{r,K} = \begin{cases} (5.16) & \text{if } \inf_K \Sigma_a > 0, \\ \left\|S_f - \operatorname{div}\mathbf{p}_h - \Sigma_a\tilde{\phi}_h\right\|_{0,K} & \text{otherwise,} \end{cases} \tag{5.33}$$

and

$$\overline{m}_K = \begin{cases} (5.16) & \text{if } \inf_K \Sigma_a > 0, \\ \dfrac{C_{P,d}h_K}{\left(\mathbb{D}^{\min}_K\right)^{1/2}} & \text{otherwise.} \end{cases}$$

Under the assumptions of Theorem 5.5, one has the efficiency[3] estimate

$$\overline{\eta}_{r,K} \leq \begin{cases} \left(\dfrac{\Sigma^{\max}_{a,K}}{\Sigma^{\min}_{a,K}}\right)^{1/2}\left(\mathtt{c}\dfrac{\mathbb{D}^{\max}_K}{\mathbb{D}^{\min}_K} + \mathtt{C}\right)^{1/2}\left|\left|\left|\zeta - \tilde{\zeta}_h\right|\right|\right|_K & \text{if } \inf_K \Sigma_a > 0, \\ \left(\mathtt{c}\dfrac{\mathbb{D}^{\max}_K}{\mathbb{D}^{\min}_K} + \mathtt{C}\dfrac{\Sigma^{\max}_{a,K}h^2_K}{\mathbb{D}^{\min}_K}\right)^{1/2}\left|\left|\left|\zeta - \tilde{\zeta}_h\right|\right|\right|_K & \text{otherwise.} \end{cases}$$

Remark 5.7. Note that the estimators are robust with respect to the interplay of the sizes of $\mathbb{D}$ and Σ_a. We refer to [26, 27] for similar works on this issue.

5.2.3. Estimates in strengthened norm

We define the norm $\|\cdot\|_S$ on $\mathcal{X}$ where, for all $\zeta \in \mathcal{X}$,

$$\begin{aligned} \|\zeta\|^2_S &= d_S(\zeta,\zeta) + \sum_{K\in\mathcal{T}_h} h^2_K\left(\mathbb{D}^{\min}_K\right)^{-1}\|\operatorname{div}\mathbf{p}\|^2_{0,K} \\ &= \left(\mathbb{D}^{-1}\mathbf{p},\mathbf{p}\right)_{0,\Omega} + \left(\Sigma_a\,\phi,\phi\right)_{0,\Omega} + \sum_{K\in\mathcal{T}_h} h^2_K\left(\mathbb{D}^{\min}_K\right)^{-1}\|\operatorname{div}\mathbf{p}\|^2_{0,K}. \end{aligned}$$

Observe that the norm $\|\cdot\|_S$ measures elements of $\mathcal{X}$ in a weighted $\mathbf{H}(\operatorname{div},\mathcal{T}_h)\times L^2(\Omega)$ norm (*cf.* [17], Chapter 8).

For $K \in \mathcal{T}_h$, we introduce $N(K) = \{K' \in \mathcal{T}_h \mid \dim_H(\partial K' \cap \partial K) = d-1\}$, where $\dim_H$ is the Hausdorff dimension, and $\mathcal{X}_K = \{\zeta = (\mathbf{p},\phi) \in \mathcal{X} \mid \operatorname{Supp}(\phi) \subset K, \operatorname{Supp}(\mathbf{p}) \subset N(K)\}$. Then one can define the following $\mathcal{X}_K$-local norm, for all $\zeta \in \mathcal{X}$,

$$|\zeta|_{+,K} = \sup_{\xi\in\mathcal{X}_K,\|\xi\|_S\leq 1} d(\zeta,\xi). \tag{5.34}$$

[3] When $\inf_K \Sigma_a = 0$, there is a h^2_K factor in the upper bound. One still obtains efficiency, since it holds $h_K \leq \operatorname{diam}(\Omega)$ for all h and all $K \in \mathcal{T}_h$.

Lemma 5.2. *Let ζ and ζ_h be respectively the solution to (4.5) and (4.7). Let $\tilde{\zeta}_h = (\mathbf{p}_h, \tilde{\phi}_h) \in \mathbf{Q}_h \times V$ be a reconstruction of ζ_h. We have for all $\xi = (\mathbf{q}, \psi) \in \mathcal{X}$,*

$$d\Big(\zeta - \tilde{\zeta}_h, \xi\Big) = \Big(S_f - \operatorname{div} \mathbf{p}_h - \Sigma_a \tilde{\phi}_h, \psi\Big)_{0,\Omega} - \Big(\mathbb{D}^{-1}\mathbf{p}_h + \mathbf{grad}\, \tilde{\phi}_h, \mathbf{q}\Big)_{0,\Omega}. \tag{5.35}$$

Proof. Let ξ be in $\mathcal{X}$. According to (4.5), we have

$$d\Big(\zeta - \tilde{\zeta}_h, \xi\Big) = \Big(S_f - \operatorname{div} \mathbf{p}_h - \Sigma_a \tilde{\phi}_h, \psi\Big)_{0,\Omega} - (\mathbb{D}^{-1}\mathbf{p}_h, \mathbf{q})_{0,\Omega} + \Big(\tilde{\phi}_h, \operatorname{div} \mathbf{q}\Big)_{0,\Omega}.$$

Owing to the fact that $\tilde{\phi}_h$ is in V, we can integrate by part the last integral:

$$d\Big(\zeta - \tilde{\zeta}_h, \xi\Big) = \Big(S_f - \operatorname{div} \mathbf{p}_h - \Sigma_a \tilde{\phi}_h, \psi\Big)_{0,\Omega} - (\mathbb{D}^{-1}\mathbf{p}_h, \mathbf{q})_{0,\Omega} - \Big(\mathbf{grad}\, \tilde{\phi}_h, \mathbf{q}\Big)_{0,\Omega}.$$

This concludes the proof. □

Theorem 5.6. *Let ζ and ζ_h be respectively the solution to (4.5) and (4.7). Let $\tilde{\zeta}_h = (\mathbf{p}_h, \tilde{\phi}_h) \in \mathbf{Q}_h \times V$ be a reconstruction of $\zeta_h = (\mathbf{p}_h, \phi_h)$. For any $K \in \mathcal{T}_h$, we define the residual estimator $\eta_{r,K}$ as in (5.16), the flux estimator $\eta_{f,K}$ as in (5.17). One has the reliability estimate*

$$\Big|\zeta - \tilde{\zeta}_h\Big|_{+,K} \leq \left(\eta_{r,K}^2 + \sum_{K' \in N(K)} \eta_{f,K'}^2\right)^{1/2}. \tag{5.36}$$

Proof. According to Lemma 5.2, we have

$$d\Big(\zeta - \tilde{\zeta}_h, \xi\Big) = \Big(S_f - \operatorname{div} \mathbf{p}_h - \Sigma_a \tilde{\phi}_h, \psi\Big)_{0,\Omega} - \Big(\mathbb{D}^{-1}\mathbf{p}_h + \mathbf{grad}\, \tilde{\phi}_h, \mathbf{q}\Big)_{0,\Omega}. \tag{5.37}$$

Let $K \in \mathcal{T}_h$ and $\xi = (\mathbf{q}, \psi) \in \mathcal{X}$ be such that $\mathrm{Supp}(\psi) \subset K$, $\mathrm{Supp}(\mathbf{q}) \subset N(K)$. Applying Cauchy–Schwarz inequalities successively in $L^2(K)$, $L^2(K')$ for $K' \in N(K)$, and then in $\mathbb{R}^{1+N(K)}$, we get

$$\begin{aligned} d\Big(\zeta - \tilde{\zeta}_h, \xi\Big) &\leq \eta_{r,K} \left\|\Sigma_a^{1/2}\psi\right\|_{0,K} + \sum_{K' \in N(K)} \eta_{f,K'} \left\|\mathbb{D}^{-1/2}\mathbf{q}\right\|_{0,K'} \\ &\leq \left(\eta_{r,K}^2 + \sum_{K' \in N(K)} \eta_{f,K'}^2\right)^{1/2} \left(\left\|\Sigma_a^{1/2}\psi\right\|_{0,K}^2 + \sum_{K' \in N(K)} \left\|\mathbb{D}^{-1/2}\mathbf{q}\right\|_{0,K'}^2\right)^{1/2}. \end{aligned}$$

We infer (5.36) from the definition of the $|\cdot|_{+,K}$ norm (5.34). □

Theorem 5.7 (Local efficiency of the *a posteriori* error estimators)**.** *Let Assumption 5.2 be fulfilled. For $K \in \mathcal{T}_h$, let $\eta_{r,K}$ and $\eta_{f,K}$ be the residual and flux estimators respectively given by (5.16), and (5.17). In addition, we suppose that ϕ_h is piecewise polynomial on $\mathcal{T}_h$. The following estimates hold true*

$$\eta_{r,K} \leq \mathtt{c} \left(\frac{\Sigma_{a,K}^{\max}}{\Sigma_{a,K}^{\min}}\right)^{1/2} \Big|\zeta - \tilde{\zeta}_h\Big|_{+,K}, \tag{5.38}$$

$$\eta_{f,K} \leq \mathtt{C} \left(\frac{\mathbb{D}_K^{\max}}{\mathbb{D}_K^{\min}}\right)^{1/2} \Big|\zeta - \tilde{\zeta}_h\Big|_{+,K}, \tag{5.39}$$

where $\mathtt{c}$ and $\mathtt{C}$ are constants which depend only on the polynomial degree of S_f, $\mathbb{D}$, Σ_a and $\tilde{\phi}_h$, d, and the shape-regularity parameter κ_K.

Proof. The first part of the proof is similar to that of Theorem 5.5. On a given $K \in \mathcal{T}_h$, let ψ_K be the bubble function, and $\psi_r = (S_f - \mathrm{div}\,\mathbf{p}_h - \Sigma_a \widetilde{\phi}_h)$ on K. As previously, we note that ψ_r is a polynomial in K. In particular, Equations (5.29), (5.30) still hold.

Now, let $\xi_{r,K} = (0, \psi_K \psi_r)$ in K, and 0 elsewhere: as we observed previously, $\xi_{r,K} \in \mathcal{X}$. Then we have, by the definition of the bilinear form d and of ζ

$$d\Big(\zeta - \tilde{\zeta}_h, \xi_{r,K}\Big) = (\psi_r, \psi_K \psi_r)_{0,K}.$$

Since the support of $\xi_{r,K}$ is equal to K, one has actually $\xi_{r,K} \in \mathcal{X}_K$. So, by definition (5.34) of the strengthened $|\cdot|_{+,K}$ norm,

$$\begin{aligned} d\Big(\zeta - \tilde{\zeta}_h, \xi_{r,K}\Big) &\le \Big|\zeta - \tilde{\zeta}_h\Big|_{+,K} \|\xi_{r,K}\|_S \\ &\le \Big|\zeta - \tilde{\zeta}_h\Big|_{+,K} \Big\|\Sigma_a^{1/2} \psi_K \psi_r\Big\|_{0,K}. \end{aligned} \tag{5.40}$$

Combining (5.29), (5.30) and (5.40), one comes to

$$c\|\psi_r\|_{0,K}^2 \le \Big|\zeta - \tilde{\zeta}_h\Big|_{+,K} \|\psi_r\|_{0,K} \big(\Sigma_{a,K}^{\max}\big)^{1/2}.$$

Using the definition of $\eta_{r,K}$ by (5.16) concludes the proof of (5.38):

$$\eta_{r,K} \le \big(\Sigma_{a,K}^{\min}\big)^{-1/2} \|\psi_r\|_{0,K} \le \frac{1}{c} \left(\frac{\Sigma_{a,K}^{\max}}{\Sigma_{a,K}^{\min}}\right)^{1/2} \Big|\zeta - \tilde{\zeta}_h\Big|_{+,K}.$$

We now proceed similarly for the second estimate. Let us denote $\mathbf{q}_f = (\mathbb{D}^{-1}\mathbf{p}_h + \mathbf{grad}\,\tilde{\phi}_h)$ on a given $K \in \mathcal{T}_h$. Note that $\mathbf{q}_f$ is a polynomial in K (thanks to Assumption 5.2 for $\mathbb{D}^{-1}$). Then the equivalence of norms on finite-dimensional spaces, the definition of ψ_K and the inverse inequality (*cf.*, *e.g.* [25], Thm. 3.2.6) give

$$c\|\mathbf{q}_f\|_{0,K}^2 \le (\mathbf{q}_f, \psi_K \mathbf{q}_f)_{0,K}, \tag{5.41}$$

$$\|\psi_K \mathbf{q}_f\|_{0,K} \le \|\mathbf{q}_f\|_{0,K}, \tag{5.42}$$

$$\|\mathrm{div}\,(\psi_K \mathbf{q}_f)\|_{0,K} \le C\, h_K^{-1} \|\psi_K \mathbf{q}_f\|_{0,K}, \tag{5.43}$$

with the constants c and C depending only on the polynomial degree of $\mathbb{D}^{-1}$ and $\tilde{\phi}_h$, d, and κ_K.

Let $\xi_{f,K} = (\psi_K \mathbf{q}_f, 0)$ in K, and 0 elsewhere. We observe that $\psi_K \mathbf{q}_f$ is smooth in K (a closed subset of $\mathbb{R}^d$), and moreover that $(\psi_K \mathbf{q}_f)_{|\partial K} = 0$ thanks to the definition of ψ_K. Hence, $\xi_{f,K} \in \mathcal{X}_K$. According to Lemma 5.2

$$\begin{aligned} -d\Big(\zeta - \tilde{\zeta}_h, \xi_{f,K}\Big) &= \Big(\mathbb{D}^{-1}\mathbf{p}_h + \mathbf{grad}\,\tilde{\phi}_h, \psi_K \mathbf{q}_f\Big)_{0,K} \\ &= (\mathbf{q}_f, \psi_K \mathbf{q}_f)_{0,K}. \end{aligned}$$

By definition (5.34) of the $|\cdot|_{+,K}$ norm, it now follows that

$$\begin{aligned} -d\Big(\zeta - \tilde{\zeta}_h, \xi_{f,K}\Big) &\le \Big|\zeta - \tilde{\zeta}_h\Big|_{+,K} \|\xi_{f,K}\|_S \\ &\le \Big|\zeta - \tilde{\zeta}_h\Big|_{+,K} \left\{ \Big\|\mathbb{D}^{-1/2}(\psi_K \mathbf{q}_f)\Big\|_{0,K}^2 + h_K^2 \big(\mathbb{D}_K^{\min}\big)^{-1} \|\mathrm{div}\,(\psi_K \mathbf{q}_f)\|_{0,K}^2 \right\}^{1/2} \\ &\le \Big|\zeta - \tilde{\zeta}_h\Big|_{+,K} \big(\mathbb{D}_K^{\min}\big)^{-1/2} \Big\{ \|(\psi_K \mathbf{q}_f)\|_{0,K}^2 + h_K^2 \|\mathrm{div}\,(\psi_K \mathbf{q}_f)\|_{0,K}^2 \Big\}^{1/2} \\ &\le \Big|\zeta - \tilde{\zeta}_h\Big|_{+,K} \big(\mathbb{D}_K^{\min}\big)^{-1/2} \big\{1 + C^2\big\}^{1/2} \|(\psi_K \mathbf{q}_f)\|_{0,K}, \end{aligned} \tag{5.44}$$

where we used the inverse inequality (5.43) to reach the last line. Combining (5.41), (5.42) and (5.44), one comes to

$$c\|\mathbf{q}_f\|^2_{0,K} \leq \left|\zeta - \tilde{\zeta}_h\right|_{+,K} \|\mathbf{q}_f\|_{0,K} \left(\mathbb{D}_K^{\min}\right)^{-1/2} \{1 + C^2\}^{1/2}.$$

Considering the definition of $\eta_{f,K}$ by (5.17) concludes the proof. □

Remark 5.8. Assume in addition in Theorem 5.7 that there exists a constant $\kappa > 0$, such that $\min_{K\in\mathcal{T}_h} \kappa_K \geq \kappa$, for all $h > 0$. Then, the constants $\mathtt{c}$ and $\mathtt{C}$ do not depend on κ_K (but on κ).

Remark 5.9. Similarly to Remark 5.6, the results of this section extend with the same arguments to the situation where $\Sigma_a \geq 0$ may vanish if one slightly modifies the definition of the norms by

$$\begin{aligned}
\|\zeta\|^2_{S,\star} &= \left(\mathbb{D}^{-1}\mathbf{p}, \mathbf{p}\right)_{0,\Omega} + (\Sigma_\star\, \phi, \phi)_{0,\Omega} + \sum_{K\in\mathcal{T}_h} h_K^2 \left(\mathbb{D}_K^{\min}\right)^{-1} \|\operatorname{div}\mathbf{p}\|^2_{0,K},\\
|\zeta|_{+,\star,K} &= \sup_{\xi\in\mathcal{X}_K, \|\xi\|_{S,\star}\leq 1} d(\zeta, \xi),
\end{aligned}$$

where $\Sigma_\star$ is defined in (5.5).

Let us define for all $K \in \mathcal{T}_h$,

$$\Sigma^{\max}_{\star,K} = \sup_{\psi\in L^2(K)\setminus\{0\}} \frac{(\Sigma_\star\psi, \psi)_{0,K}}{\|\psi\|^2_{0,K}}, \quad \Sigma^{\min}_{\star,K} = \inf_{\psi\in L^2(K)\setminus\{0\}} \frac{(\Sigma_\star\psi, \psi)_{0,K}}{\|\psi\|^2_{0,K}}.$$

Under the assumptions of Theorem 5.6, one has the reliability estimate

$$\left|\zeta - \tilde{\zeta}_h\right|_{+,K} \leq \left(\eta^2_{r,\star,K} + \sum_{K'\in N(K)} \eta^2_{f,K'}\right)^{1/2},$$

where the residual estimator becomes $\overline{\eta}_{r,\star,K} = \eta_{r,\star,K}\overline{m}_{\star,K}$ with

$$\eta_{r,\star,K} = \left\|\Sigma_\star^{-1/2}\left(S_f - \operatorname{div}\mathbf{p}_h - \Sigma_a\tilde{\phi}_h\right)\right\|_{0,K} \quad \text{and} \quad \overline{m}_{\star,K} = \min\left\{1, \frac{C_{P,d}\, h_K \left(\Sigma^{\max}_{\star,K}\right)^{1/2}}{\left(\mathbb{D}_K^{\min}\right)^{1/2}}\right\}.$$

Under the assumptions of Theorem 5.7, one has the efficiency estimates

$$\begin{aligned}
\eta_{r,\star,K} &\leq \mathtt{c}\left(\frac{\Sigma^{\max}_{\star,K}}{\Sigma^{\min}_{\star,K}}\right)^{1/2} \left|\zeta - \tilde{\zeta}_h\right|_{+,\star,K},\\
\eta_{f,K} &\leq \mathtt{C}\left(\frac{\mathbb{D}_K^{\max}}{\mathbb{D}_K^{\min}}\right)^{1/2} \left|\zeta - \tilde{\zeta}_h\right|_{+,\star,K}.
\end{aligned}$$

6. Numerical results

This section is devoted to the numerical experiments we performed on adaptive mesh refinement strategies (AMR). In fact, the AMR strategy can be classified into several categories: the h-refinement (mesh subdivision), which amounts to refining the mesh where large errors occur [28]; the p-refinement (local high order approximation), which increases the order of the polynomial functions [29], or the r-refinement (moving mesh) that moves the nodes of the mesh to increase the mesh density [30], in the regions of interest where large variations of the solution occur. The above strategies can be mixed, such as hp-refinement [31,32] and hr-refinement [33].

We are interested in the case of heterogeneous coefficients which may induce some singularities in the solution of Problem (3.1), that is a loss of regularity of the solution due to the discontinuities in the data. Therefore, we focus on mesh subdivision strategy in this section.

The performance of adaptive mesh refinement is assessed with respect to various criteria such as the error estimator, the marker strategy and the threshold parameter. We recall that the context of our applications is modelling nuclear reactor cores, in particular geometries composed of rectangular cuboids of $\mathbb{R}^3$. This is the reason why the discretization in this section is applied on Cartesian meshes. We recall that we refer to [19] for the definition of the corresponding finite-dimensional spaces M_h, see footnote[2] page 7.

Mesh subdivision strategies are introduced in Section 6.1. Section 6.2 presents the set of test cases considered throughout the whole Section 6. Section 6.3 focuses on the marker strategies. The sensitivity with respect to the threshold parameter is investigated in Section 6.4. Section 6.5 compares various reconstruction approaches. Section 6.6 examines different error estimators.

6.1. An adaptive mesh refinement strategy

We recall in this section the h-refinement approach.

From the initial mesh $\mathcal{T}_{h_0}$, the AMR strategy generates a sequence of meshes $\mathcal{T}_{h_k}$. This strategy corresponds in general to an iterative loop where at each iteration, we consider the following four modules:

Assuming that the mesh $\mathcal{T}_{h_k}$ is computed, module **Solve** indeed corresponds to solving Problem (4.7). The output of the module **Estimate** is $(\eta_K)_{K\in\mathcal{T}_{h_k}}$ where η_K is an *a posteriori* error estimator. The stopping criterion of the algorithm is given by

$$\max_{K\in\mathcal{T}_h} \eta_K \leq \epsilon_{\mathrm{AMR}},$$

where $\epsilon_{\mathrm{AMR}} > 0$. Module **Mark** returns the set of the marked cells based on the error estimators $(\eta_K)_{K\in\mathcal{T}_{h_k}}$. In other words, this module selects a set of elements to be refined. To be convenient, for $S \subset \mathcal{T}_{h_k}$, let us denote $\eta(S) = (\sum_{K\in S} \eta_K^2)^{1/2}$. For a fixed threshold parameter $0 < \theta \leq 1$, the classical bulk-chasing criterion (Dörfler's marking strategy [34]) is to select the (smallest) set of elements such that

$$\eta(S) \geq \theta\eta(\mathcal{T}_{h_k}). \tag{6.1}$$

Lastly, module **Refine** performs the mesh refinement according to the selected mesh elements.

This strategy is generic and can be applied to any kind of mesh.

In order to preserve the Cartesian structure of the mesh at each iteration, it is essential to refine the mesh according to *lines*, along the directions $(\mathbf{e}_x)_{x=1,d}$, which contain at least one of the selected cells. As a consequence, it is obvious to see that the above *cell marker strategy* is extremely costly since we use the error indicator of some selected cells to refine the other cells located in the same line, for a given direction $(\mathbf{e}_x)_{x=1,d}$. Due to this drawback, it is relevant to consider some other marker cell strategies. Therefore, instead of using the classical bulk-chasing criterion defined on a single cell, we introduce:

- *aggregate error indicators* according to a line containing the cell;
- *the sum of aggregate error indicators*, taken on all lines containing the cell.

We call them respectively the *direction marker* and *cross marker* method.

The *direction marker* method consists in selecting for each direction $\mathbf{e}_x, x = 1, \dots, d$, the smallest set of lines $L_x \subset \mathcal{T}_{h_k}$ parallel to $\mathbf{e}_x$ such that

$$\eta(L_x) \geq \theta_l \eta(\mathcal{T}_{h_k}), \tag{6.2}$$

where $0 < \theta_l \leq 1$ is a fixed threshold parameter. The resulting selected set is $\cup_{x=1,d} L_x$.

The *cross marker* method corresponds to selecting for each $K \in \mathcal{T}_h$, the smallest set of elements $S \subset \mathcal{T}_{h_k}$ such that

$$\sum_{K \in S} \eta(C_K) \geq \theta_c \eta(\mathcal{T}_{h_k}), \tag{6.3}$$

where $0 < \theta_c \leq 1$ is a fixed threshold parameter and C_K is the union over all directions $(\mathbf{e}_x)_{x=1,d}$ of the lines containing K. The resulting selected set is $\cup_{K \in S} C_K$. Interestingly, performing the mesh refinement is straightforward with both the *direction marker* and *cross marker* methods. In addition, they both preserve the Cartesian structure of the mesh.

6.2. Setting of the test cases

This section is devoted to the definition of the test cases considered, namely the Dauge test case, the Checkerboard test case, the Center test case and the Rotation Center test cases. In the following test cases, we perform the numerical simulations on the domain $\Omega = (0,1)^2$. We consider here a simple source term given by $S_f = 1$. Moreover, we assume that the diffusion coefficient $\mathbb{D}$ is a scalar, piecewise constant given by Figure 1. We also set $\Sigma_a = 1$.

In the following, the initial mesh of the AMR strategy is chosen to be uniform in all directions. The mesh size of the initial mesh of the Dauge test case, the Checkerboard test case, the Center test case and the Rotation Center test cases are respectively equal to 1/4, 1/8, 1/6 and 1/8.

6.3. Influence of the marker cell strategy

We now study the influence of the marker cell strategy on the AMR approach for our set of test cases. In this section, the Dauge test case, the Center test case and the Checkerboard test case are performed with RTN_0 finite elements, while the Rotation Center test cases are performed with RTN_1 finite elements. The AMR strategies are based on the error estimator introduced in (5.36) and the average reconstruction (5.1).

The Dauge test case is a singular toy problem (see also in [2,17,35] and references therein for more details). In this test case, the singularity is located at $(0.5, 0.5)$ and we expect refinement in this region. Adaptive mesh refinement is performed with a stopping criterion equal to $\epsilon_{\mathrm{AMR}} = 2 \times 10^{-3}$. Figure 2 shows that mesh refinement is more located near the singularity for the *direction marker* strategy than the other strategies. Moreover, Figure 3 shows that the *direction marker* needs three to four times less mesh elements than the other strategies. All the other test cases yield the same conclusions.

So, from now on, the adaptive mesh refinement is always performed with the *direction marker* method.

6.4. Sensitivity with respect to the threshold parameter

In this section, we evaluate the sensitivity with respect to the threshold parameter θ_l defined in (6.2) on our set of test cases. For unstructured meshes like triangular meshes, the typical value for the threshold parameter θ_l is 0.5. However, the choice of an optimal value for the threshold parameter θ_l remains a difficult question. Therefore, we numerically investigate the optimal value of the threshold parameter.

The stopping criterion of the Checkerboard test case is $\epsilon_{\mathrm{AMR}} = 5 \times 10^{-3}$. For the other test cases, the stopping criterion is set to $\epsilon_{\mathrm{AMR}} = 2 \times 10^{-3}$.

For the sake of brevity, we only show Figure 4 that indicates that the optimal value of θ_l for the Dauge test case is around 0.35. Figure 5 shows the numerical flux on the refined mesh with an optimal value of θ_l for the different test cases.

6.5. Influence of the reconstruction

In this section, we investigate the influence of the reconstruction on the error estimator defined in (5.36). To this aim, we compare the reconstruction approaches defined in Section 5.1 on the Dauge test case.

First, the stopping criterion is fixed at $\epsilon_{\mathrm{AMR}} = 1.5 \times 10^{-3}$. As can be seen in Figure 6, the average reconstruction and RTN post-processing need more elements to reach the stopping criterion than the RTN_0 post-processing.

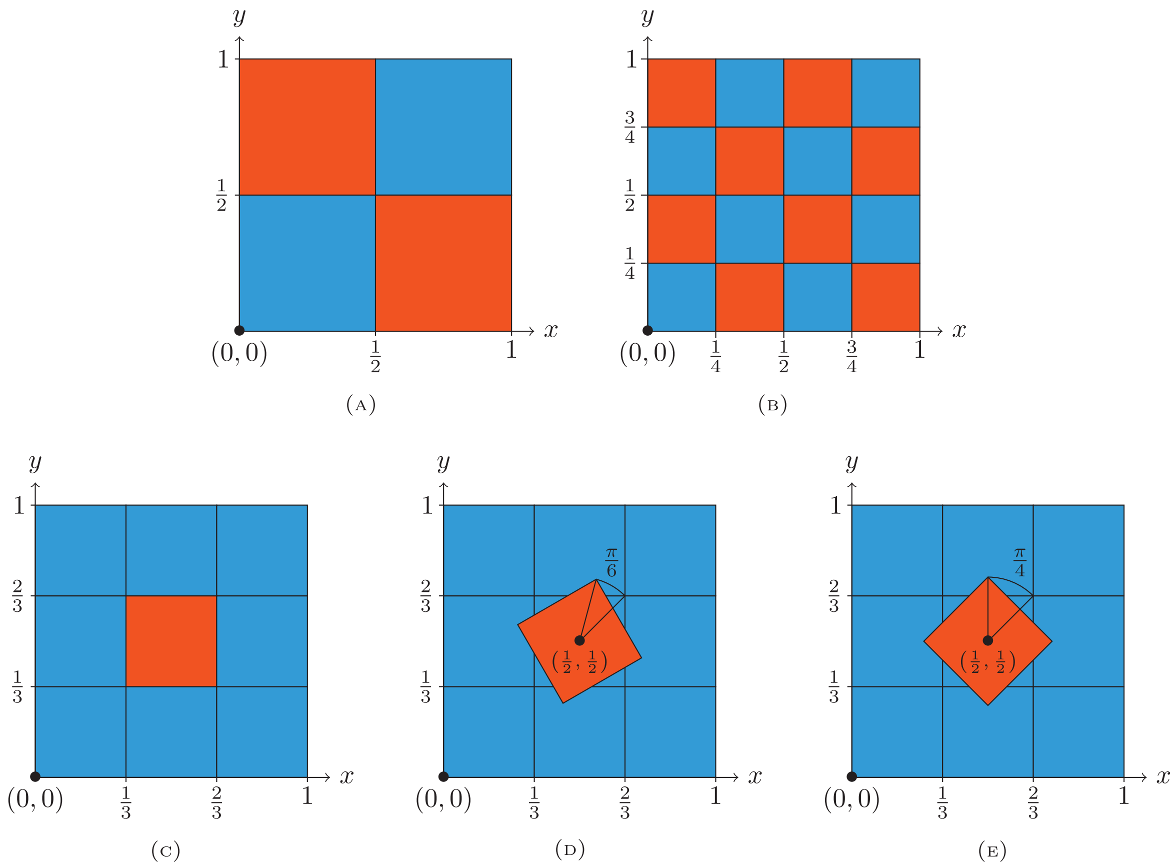


FIGURE 1. The diffusion coefficient $\mathbb{D}$: the region corresponds to $\mathbb{D} = 10$ and the other region stands for $\mathbb{D} = 1$. (A) Dauge. (B) Checkerboard. (C) Center. (D) Rotation Center, $\alpha = \frac{\pi}{6}$. (E) Rotation Center, $\alpha = \frac{\pi}{4}$.

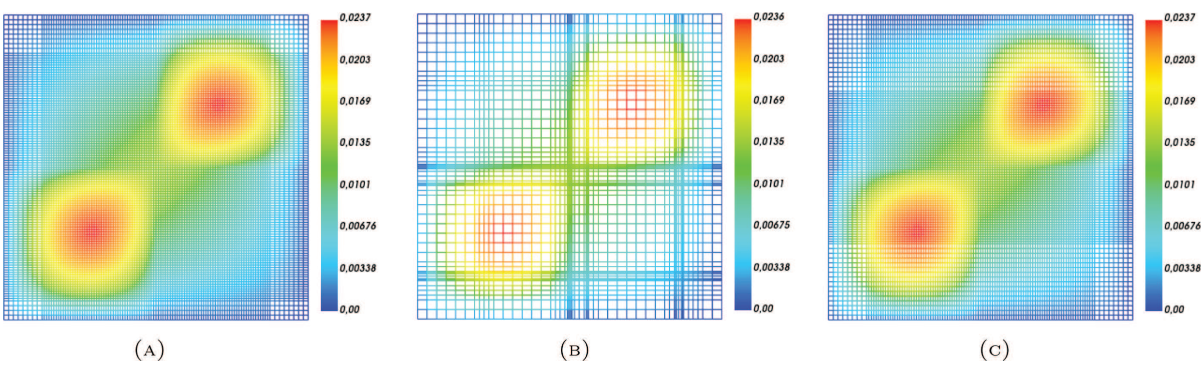
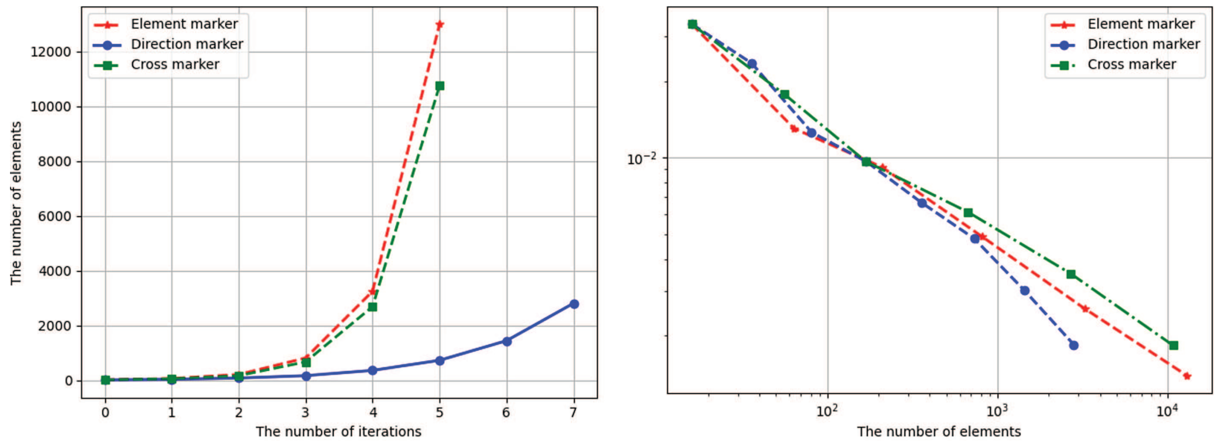
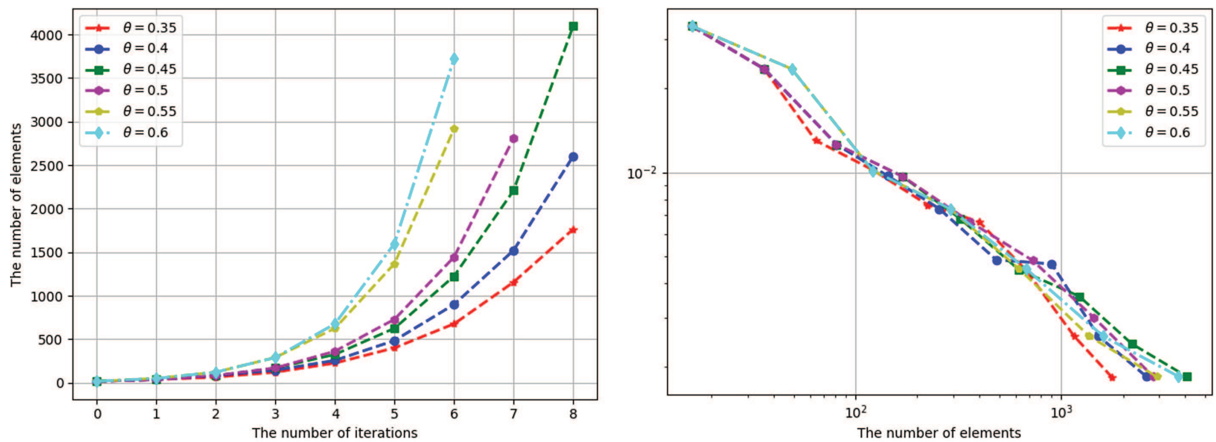


FIGURE 2. Dauge test case: the numerical flux on refined meshes for different marker strategies with RTN_0. (A) Cell marker. (B) Direction marker. (C) Cross marker.



(A) (B)

FIGURE 3. Dauge test case: evolution of the number of elements and the maximum of the total error estimator for different marker cell strategies. (A) Evolution of the number of elements. (B) Evolution of the maximum of total error estimator.



(A) (B)

FIGURE 4. Dauge test case with varying thresholds. (A) Evolution of the number of elements. (B) Evolution of the maximum of total error estimator.

It is related to the fact that the flux estimator is the dominant contribution of the total estimator. Figure 7 shows the numerical flux on the refined mesh for the different reconstructions.

Second, we modify the stopping criterion. Now, the stopping criterion is based on the L_2 error with respect to a reference solution. That is to say, the algorithm is stopped when

$$\frac{\|\phi_{\mathrm{ref}} - \phi_h\|_{0,\Omega}}{\|\phi_{\mathrm{ref}}\|_{0,\Omega}} \leq \epsilon_{\mathrm{rel}}, \tag{6.4}$$

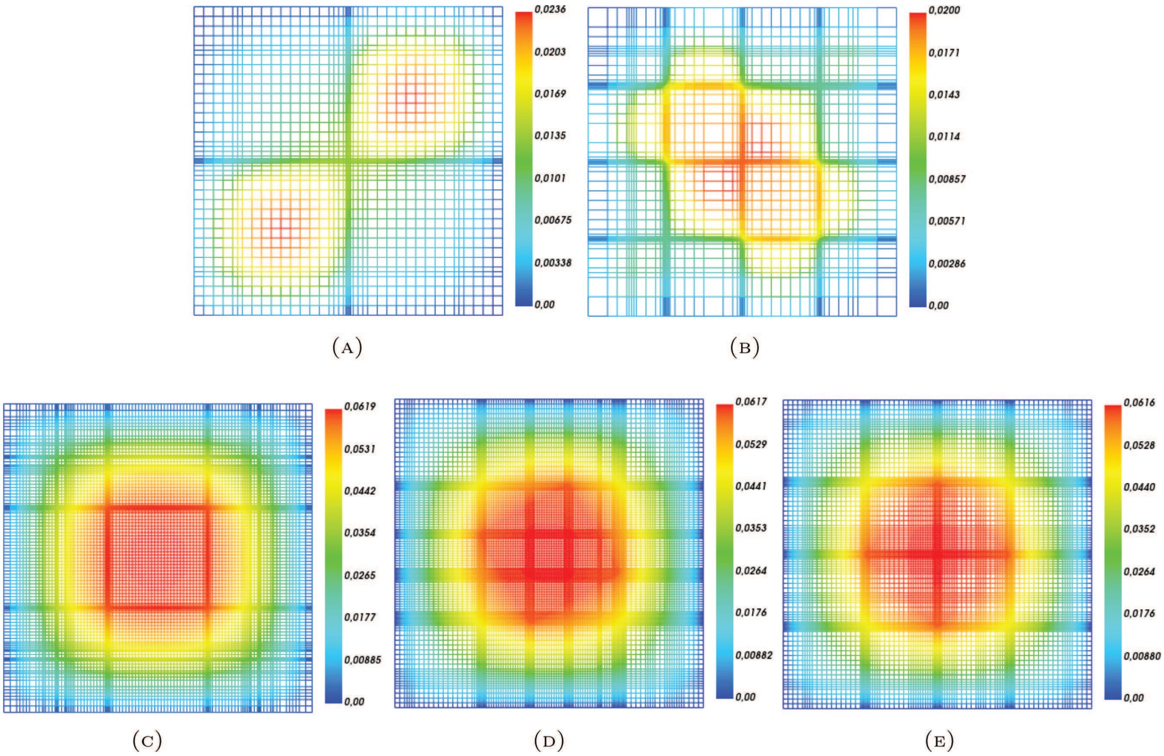


FIGURE 5. The numerical flux ϕ_h of the optimal threshold value for the different test cases. (A) Dauge, $\theta_l = 0.35$. (B) Checkerboard, $\theta_l = 0.35$. (C) Center, $\theta_l = 0.6$. (D) Rotation Center, $\alpha = \pi/6$, $\theta_l = 0.45$. (E) Rotation Center, $\alpha = \pi/4$, $\theta_l = 0.4$.

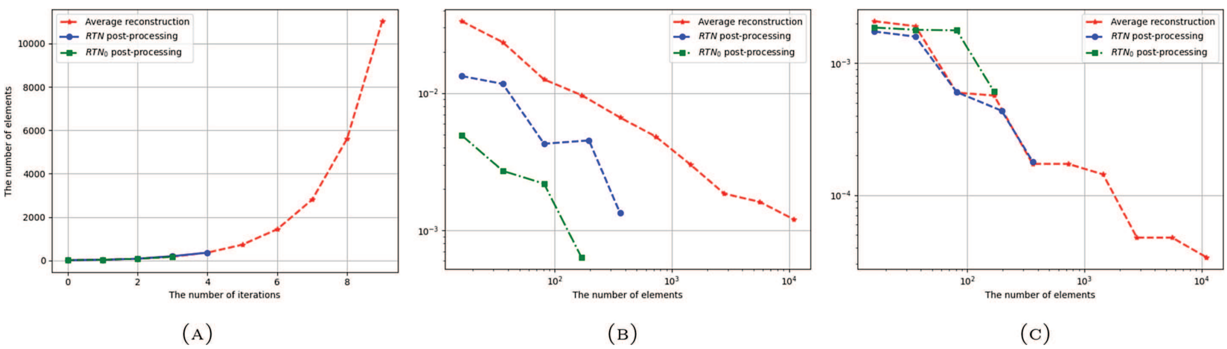
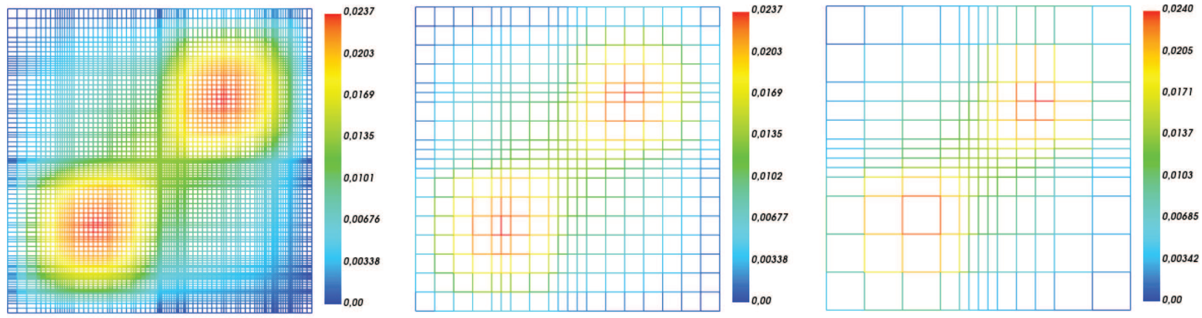
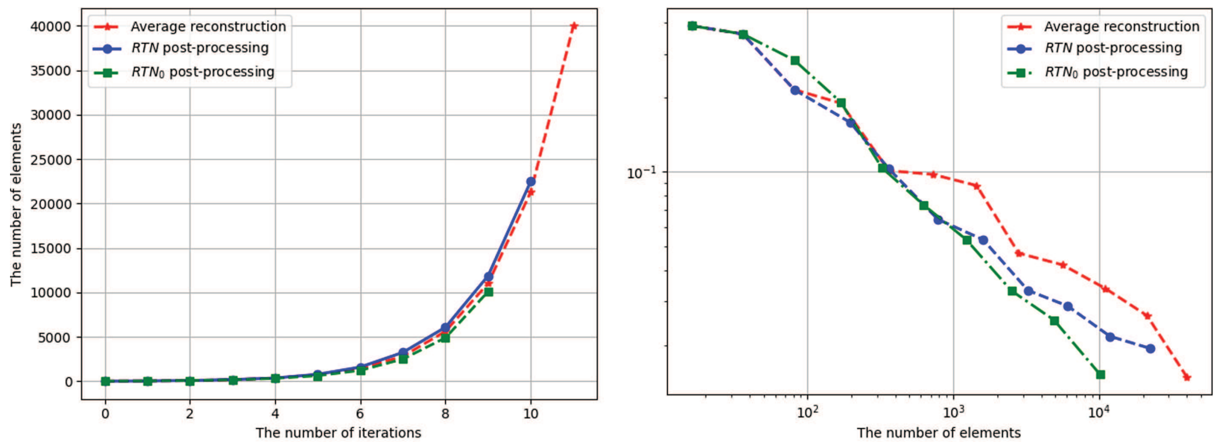


FIGURE 6. Evolution of the number of elements and the maximum of the total estimator by using different reconstruction methods: namely the *average reconstruction* (5.1), the *reconstruction* associated to the RTN post-processing (5.6) and the *reconstruction* associated to the RTN_0 post-processing (5.3). (A) Number of elements. (B) Maximum of the total estimator. (C) Maximum of the residual estimator.



(A) (B) (C)

FIGURE 7. The numerical flux ϕ_h for different reconstruction methods. (A) Average reconstruction (5.1). (B) *Reconstruction* associated to the RTN post-processing (5.6). (C) *Reconstruction* associated to the RTN_0 post-processing (5.3).



(A) (B)

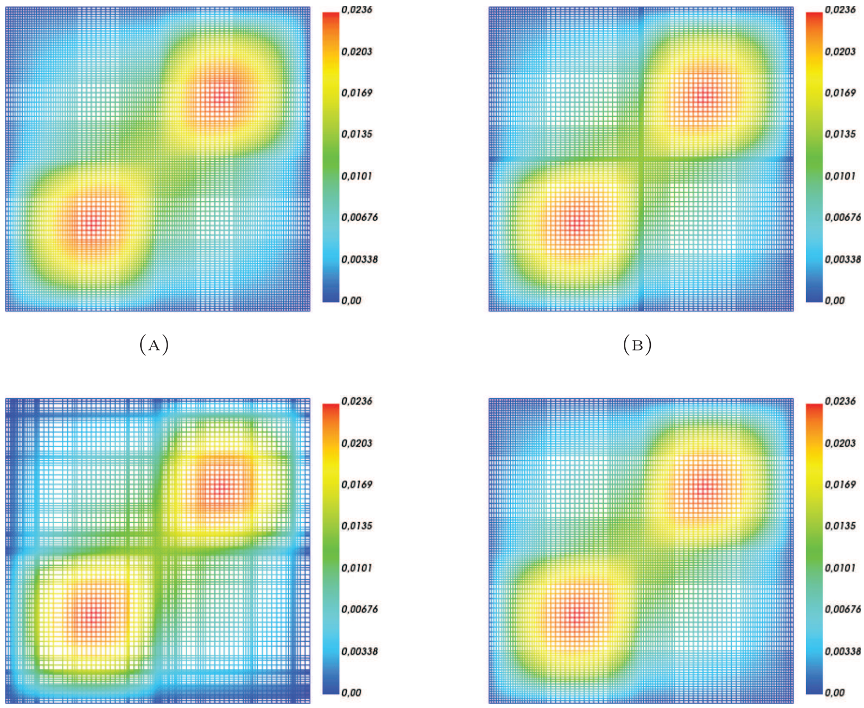
FIGURE 8. Evolution of the number of elements and the relative L^2 error (6.4) for different reconstruction methods : namely the *average reconstruction* (5.1), the *reconstruction* associated to the RTN post-processing (5.6) and the *reconstruction* associated to the RTN_0 post-processing (5.3). (A) Number of elements. (B) Relative L^2 error.

where ϕ_{ref} is a reference solution computed on a fine mesh. We fix $\epsilon_{\text{rel}} = 2 \times 10^{-2}$. Figure 8 shows that the RTN_0 post-processing and RTN post-processing give similar AMR strategies and that the resulting mesh have fewer elements than with the average reconstruction.

6.6. Comparison of the error estimators

In this section, we perform the Dauge test case with a stopping criterion on the relative L^2 error (6.4) with respect to a reference solution at $\epsilon_{\text{rel}} = 2 \times 10^{-2}$. To be convenient, let Estimator 1, Estimator 2, Estimator 3 and Estimator 4 respectively stand for the error estimator defined in ([17], Theorem 8.4), (5.36), (5.18) and (5.11). For the sake of completeness, we recall here the different estimators for all $K \in \mathcal{T}_h$,

$$\eta_K^1 = \left(\widehat{\eta}_{r,K}^2 + \eta_{f,K}^2 + 9\eta_{nc,K}^2\right)^{1/2} \quad \text{where } \eta_{nc,K} = \left\| \Sigma_a^{1/2}\left(\phi_h - \tilde{\phi}_h\right)\right\|_{0,K},$$



(A) (B) (C) (D)

FIGURE 9. The numerical flux ϕ_h for different error estimators. (A) Estimator 1. (B) Estimator 2. (C) Estimator 3. (D) Estimator 4.

$$\begin{aligned}
\eta_K^2 &= \left(\eta_{r,K}^2 + \sum_{K' \in N(K)} \eta_{f,K'}^2\right)^{1/2}, \\
\eta_K^3 &= \overline{\eta}_{r,K} + \eta_{f,K}, \\
\eta_K^4 &= \tilde{\eta}_{nc,K} + \tilde{\eta}_{r,K}.
\end{aligned}$$

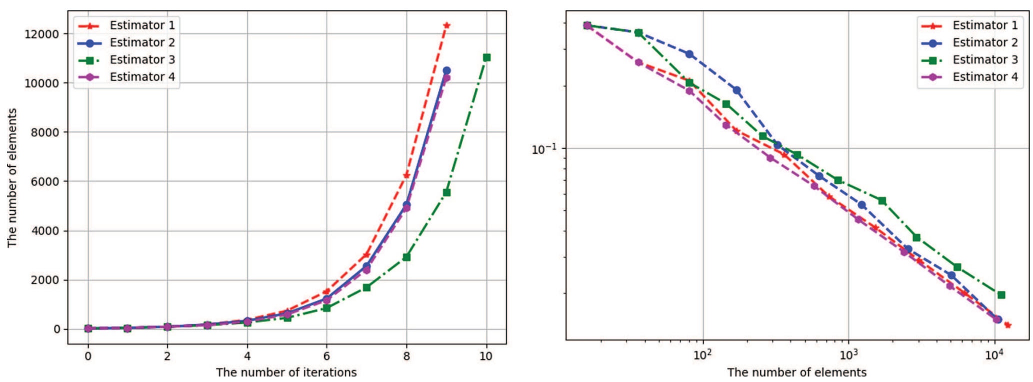
We apply the *reconstruction* associated to the RTN_0 post-processing with bubble correction (5.3)–(5.7) to Error estimators 1, 2 and 3 (we refer to Thm. 5.1 for Estimator 4). As can be seen in Figure 9, we obtain similar meshes for the various AMR strategies using Estimators 1, 2 and 4. On the other hand, there is more refinement near the discontinuities for Estimator 3. Also, the relative L^2 error on the neutron flux are similar for the different AMR strategies according to Figure 10.

7. Conclusion

In this manuscript, we derive *a posteriori* estimates associated to different norms for the numerical solution of the neutron diffusion equation in mixed form.

We discuss the approach presented in [17], Chapter 8. Although reliability can be proven, it remains difficult to achieve local efficiency of the estimators. We address this issue by proposing *a posteriori* estimates that are both reliable and locally efficient. We also propose two norms to measure the errors.

Regarding the numerical aspects, we focus on Cartesian meshes, since such structures are relevant in nuclear core applications, and outline a robust marker strategy for this specific constraint, the *direction marker* strategy.



(A) (B)

FIGURE 10. Evolution of the number of elements and the relative L^2 error (6.4) for different error estimators. (A) Number of elements. (B) Relative L^2 error.

We observe numerically that the AMR strategy is sensitive to the choice of the threshold parameter. We compare various *a posteriori* estimators under different criteria. We show that the choice of the reconstruction has a strong influence on the AMR strategy. The post-processing approaches are shown to be more efficient than the average reconstruction. In the case of the lowest-order Raviart–Thomas–Nédélec finite element, we observe that the RTN_0 post-processing gives a more accurate reconstruction compared to the RTN post-processing. Also, we compare the different estimators with the same choice of reconstruction. And we note that, if the stopping criterion is based on the L^2 error with respect to a reference solution, the various refinement strategies yield similar results.

In a companion paper, we will consider more general models or settings. First, the extension of our *a posteriori* estimators to a Domain Decomposition Method, the so-called DD+L^2 jumps method. Then, we will study a more general model, widely used for nuclear core simulations, the multigroup diffusion problem, for which we will also provide *a posteriori* estimators. In both cases, these will be proven to be reliable and locally efficient.

Acknowledgements. The authors thank the referees for their many comments on the original version of this work.

References

[1] J.J. Duderstadt and L.J. Hamilton, Nuclear Reactor Analysis. John Wiley & Sons, Inc., New York (1976).
[2] P. Ciarlet, Jr., E. Jamelot and F.D. Kpadonou, Domain decomposition methods for the diffusion equation with low-regularity solution. *Comput. Math. App.* **74** (2017) 2369–2384.
[3] P. Ciarlet, Jr., L. Giret, E. Jamelot and F.D. Kpadonou, Numerical analysis of the mixed finite element method for the neutron diffusion eigenproblem with heterogeneous coefficients. *ESAIM: Math. Modell. Numer. Anal.* **52** (2018) 2003–2035.
[4] C. Carstensen, A posteriori error estimate for the mixed finite element method. *Math. Comp.* **66** (1997) 465–476.
[5] M.G. Larson and A. Målqvist, A posteriori error estimates for mixed finite element approximations of elliptic problems. *Numer. Math.* **108** (2008) 487–500.
[6] C. Lovadina and R. Stenberg, Energy norm a posteriori error estimates for mixed finite element methods. *Math. Comp.* **75** (2006) 1659–1674.
[7] M. Vohralík, Unified primal formulation-based a priori and a posteriori error analysis of mixed finite element methods. *Math. Comp.* **79** (2010) 2001–2032.
[8] B. Wohlmuth and R. Hoppe, A comparison of a posteriori error estimators for mixed finite element discretizations by Raviart-Thomas elements. *Math. Comp.* **68** (1999) 1347–1378.
[9] M.F. Wheeler and I. Yotov, A posteriori error estimates for the mortar mixed finite element method. *SIAM J. Numer. Anal.* **43** (2005) 1021–1042.
[10] M. Vohralík, A posteriori error estimates for lowest-order mixed finite element discretizations of convection-diffusion-reaction equations. *SIAM J. Numer. Anal.* **45** (2007) 1570–1599.

[11] E. Jamelot, A.-M. Baudron and J.-J. Lautard, Domain decomposition for the SP_N solver MINOS. *Transp. Theory Stat. Phys.* **41** (2012) 495–512.
[12] E. Jamelot and P. Ciarlet, Jr., Fast non-overlapping Schwarz domain decomposition methods for solving the neutron diffusion equation. *J. Comput. Phys.* **241** (2013) 445–463.
[13] A. Ern and J.-L. Guermond, Theory and Practice of Finite Elements. Springer-Verlag, Berlin (2004).
[14] P.-A. Raviart and J.-M. Thomas, A mixed finite element method for second order elliptic problems. In: Mathematical Aspects of Finite Element Methods. Vol. 606 of *Lecture Notes in Mathematics*. Springer (1977) 292–315.
[15] J.-C. Nédélec, Mixed finite elements in $\mathbb{R}^3$. *Numer. Math.* **35** (1980) 315–341.
[16] D. Boffi, F. Brezzi and M. Fortin, Mixed and Hybrid Finite Element Methods and Applications. Springer-Verlag, Berlin (2013).
[17] L. Giret, *Non-conforming domain decomposition for the multigroup neutron SP_N equations*. Ph.D. thesis, Université Paris Saclay (2018).
[18] P. Oswald, On a BPX-preconditioner for P1 elements. *Computing* **51** (1993) 125–133.
[19] T. Arbogast and Z. Chen, On the implementation of mixed methods as nonconforming methods for second-order elliptic problems. *Math. Comp.* **64** (1995) 943–972.
[20] F. Févotte, Une méthode de post-traitement des éléments finis de Raviart-Thomas appliquée à la neutronique. CMAP Seminar, Palaiseau, May 21 (2019).
[21] A. Ern and M. Vohralík, A posteriori error estimation based on potential and flux reconstruction for the heat equation. *SIAM J. Numer. Anal.* **48** (2010) 198–223.
[22] L.E. Payne and H.F. Weinberger, An optimal Poincaré inequality for convex domains. *Arch. Ration. Mech. Anal.* **5** (1960) 286–292.
[23] M. Bebendorf, A note on the Poincaré inequality for convex domains. *Zeitschrift für Analysis und ihre Anwendungen* **22** (2003) 751–756.
[24] W. Prager and J.L. Synge, Approximation in elasticity based on the concept of functions spaces. *Q. Appl. Math.* **5** (1947) 241–269.
[25] P.G. Ciarlet, The Finite Element Method for Elliptic Problems. Vol. 40 of *Classics in Applied Mathmetics*. SIAM (2002).
[26] I. Cheddadi, R. Fučík, M.I. Prieto and M. Vohralík, Guaranteed and robust a posteriori error estimates for singularly perturbed reaction–diffusion problems. *ESAIM: Math. Modell. Numer. Anal.* **43** (2009) 867–888.
[27] R. Verfürth, Robust a posteriori error estimators for a singularly perturbed reaction–diffusion equation. *Numer. Math.* **78** (1998) 479–493.
[28] R. Verfürth, A posteriori error estimation and adaptive mesh-refinement techniques. *J. Comput. Appl. Math.* **50** (1994) 67–83.
[29] I. Babuška, B.A. Szabo and I.N. Katz, The *p*-version of the finite element method. *SIAM J. Numer. Anal.* **18** (1981) 515–545.
[30] W. Cao, W. Huang and R.D. Russell, An *r*-adaptive finite element method based upon moving mesh PDEs. *J. Comput. Phys.* **149** (1999) 221–244.
[31] I. Babuška and M. Suri, The *p* and *hp* versions of the finite element method, basic principles and properties. *SIAM Rev.* **36** (1994) 578–632.
[32] P. Daniel, A. Ern, I. Smears and M. Vohralík, An adaptive *hp*-refinement strategy with computable guaranteed bound on the error reduction factor. *Comput. Math. Applic.* **76** (2018) 967–983.
[33] J. Lang, W. Cao, W. Huang and R.D. Russell, A two-dimensional moving finite element method with local refinement based on a posteriori error estimates. *Appl. Numer. Math.* **46** (2003) 75–94.
[34] W. Dörfler, A convergent adaptive algorithm for Poisson's equation. *SIAM J. Numer. Anal.* **33** (1996) 1106–1124.
[35] M. Dauge, Benchmark computations for Maxwell equations for the approximation of highly singular solutions. Available at: https://perso.univ-rennes1.fr/monique.dauge/core/index.html (2004).