# Reliability and efficiency of an anisotropic Zienkiewicz-Zhu error estimator


Stefano Micheletti Simona Perotto

MOX– Modellistica e Calcolo Scientifico
Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy
stefano.micheletti@mate.polimi.it, simona.perotto@mate.polimi.it



**Keywords**
Reliability and efficiency of an error estimator, recovery-based error estimators, anisotropic mesh adaption, finite elements.



**Abstract**

In this paper we study the efficiency and the reliability of an anisotropic a posteriori error estimator in the case of the Poisson problem supplied with mixed boundary conditions. The error estimator may be classified as a residual-based one, but its novelty is twofold: firstly, it employs anisotropic estimates of the interpolation error for linear triangular finite elements and, secondly, it makes use of the Zienkiewicz-Zhu recovery procedure to approximate the gradient of the exact solution. Finally, we describe the adaptive procedure used to obtain a numerical solution satisfying a given accuracy, and we include some numerical test cases to assess the robustness of the proposed numerical algorithm.


## 1 Introduction and motivations

In [27, 31] an error estimator computationally cheap but, at the same time, able to detect the directional features of the solution of the problem at hand is introduced. These good properties are obtained by suitably combining the Zienkiewicz-Zhu (ZZ) gradient recovery procedure [40, 41, 42, 43] with the anisotropic error estimates of [11, 12]. This is carried out first by developing a residual-based error estimator. Then the error in the interpolation terms is bounded via suitable anisotropic error estimates. Finally, the derivatives of the exact solution entering these anisotropic terms are replaced by recovered quantities, in the spirit of the ZZ procedure.

Let us show in more detail how this error estimator is obtained in a quite general setting. Suppose that the weak form of the problem at hand is:
find $u \in V$ such that, for any $v \in V$,

$$B(u, v) = L(v), \tag{1}$$

where $V$ is a Hilbert space, $B(\cdot,\cdot) : V \times V \to \mathbb{R}$ is a coercive symmetric bilinear form, $L : V \to \mathbb{R}$ is an element of the dual space $V'$ of $V$. Then the approximated problem is:
find $u_h \in V_h$ such that, for any $v_h \in V_h$,

$$B(u_h, v_h) = L(v_h),$$

where $V_h \subset V$ is a suitable finite dimensional subspace of $V$. Then it follows that, for any $v \in V$,

$$B(u - u_h, v) = L(v) - B(u_h, v) = \mathcal{R}(v),$$

$\mathcal{R} \in V'$ being the (weak) residual associated with (1). The well-known Galerkin orthogonality property ($\mathcal{R}(v_h) = 0$, for any $v_h \in V_h$), yields

$$B(u - u_h, v) = \mathcal{R}(v - v_h).$$

By localizing the residual term over the elements $K$ and the edges of the triangulation $\mathcal{T}_h$, and using the Cauchy-Schwarz inequality as well as suitable interpolation error estimates, we have

$$|B(u - u_h, v)| \leq C \sum_{K \in \mathcal{T}_h} \alpha_K \, \rho_K(u_h) \, w_K(v), \tag{2}$$

where $\alpha_K$ are area-dependent coefficients, while $\rho_K(u_h)$ and $w_K(v)$ are the local (interior and edge) residual terms and the anisotropic weights associated with the function $v$, respectively, with $C$ a suitable constant. By identifying in (2) $v$ with the discretization error $e_h = u - u_h$, we get

$$B(e_h, e_h) \leq C \sum_{K \in \mathcal{T}_h} \alpha_K \, \rho_K(u_h) \, w_K(e_h). \tag{3}$$

We may characterize such a result as an *implicit* estimate for the energy norm of $e_h$, since $e_h$ appears at both the left and right-hand sides of (3), the energy norm being defined by the bilinear form itself. The idea now is to replace the above term $w_K(e_h)$, usually depending on the first and/or the second derivatives of $e_h$, with the computable quantity $w_K(e_h^*)$, obtained by employing, for instance, recovered derivatives of $u$ instead of the exact ones, following the ZZ approach. Thus the final estimator for the energy norm of $e_h$ is defined by

$$\eta = \Big( \sum_{K \in \mathcal{T}_h} \eta_K^2 \Big)^{1/2}, \tag{4}$$

with $\eta_K = \big(\alpha_K \rho_K(u_h) w_K(e_h^*)\big)^{1/2}$.

Since the pioneering work [40] dealing with the linear elastic problem, and some further papers [42, 43], it has been attempted to theoretically understand the amazingly good properties of the ZZ error estimator, obtained by approximating the true gradient, for instance, with the recovered one. One of the first work in which some averaging technique is studied is [20], though the idea is nearly as old as the finite element method itself (see, e.g., [39]). In the literature emphasis is often given to superconvergence results, that is, the phenomenon observed when using, for example, continuous piecewise linear finite elements: the convergence rate of the averaged gradient to the exact gradient in the $L^2$-norm can locally be higher, even by one order, than that of the original piecewise constant discrete gradient, under some smoothness assumption on the solution and on the domain, and under some regularity constraints on the mesh. For instance, in [20] *uniform triangulations* are necessary; in [21] a *regular family of uniform triangulations* of a polygonal domain is considered; in [10] a *quasi-parallelism* assumption is made; in [24] generalizations of previous results assuming *fully-structured* partitions or *strongly-regular* meshes to *globally mildly structured meshes* are derived. Theoretical properties of different types of ZZ-like error estimators are considered also in e.g. [3, 5, 23, 37, 38].

Aim of this paper is to study the efficiency and the reliability of an anisotropic a posteriori error estimator of the type (4), in the case of the Poisson problem provided with mixed boundary conditions.
Some numerical tests on anisotropic error estimators of type (4) have already been presented in [27, 31] while generalization to other elliptic problems and to parabolic problems are discussed in [32].

The outline of the paper is as follows: after introducing the anisotropic setting in Section 2, we derive in Section 3 the anisotropic error estimator of type (4) for a model elliptic boundary value problem. In Section 4 we provide the theoretical tools for studying the reliability and the efficiency of the error estimator, carried out in Section 5 and 6, respectively. Finally, in Section 7 we discuss how the anisotropic error estimator can be used to generate an adapted mesh and we numerically validate the proposed theory on some test cases.

## 2 Functional and anisotropic framework

Let us introduce the functional spaces used in the sequel. Let $\Omega$ be a polygonal domain of $\mathbb{R}^2$ with Lipschitz continuous boundary $\partial\Omega$. First, let $L^2(\Omega)$ be the space of the Lebesgue square-integrable functions with norm $\|\cdot\|_{L^2(\Omega)}$ and scalar product $(\cdot,\cdot)$.
Then let $W^{k,p}(\Omega)$ be the classical Sobolev spaces of functions for which the $p$-th power of their distributional derivatives of order up to $k \geq 0$ is Lebesgue-measurable and $1 \leq p < \infty$ [25]. In particular, for $p = 2$, the space $W^{k,2}(\Omega)$ is denoted with $H^k(\Omega)$, with norm and seminorm $\|\cdot\|_{H^k(\Omega)}$ and $|\cdot|_{H^k(\Omega)}$, respectively. When these norms or seminorms are referred to some subset $S$ of $\Omega$, they are written as $\|\cdot\|_{L^2(S)}$, $\|\cdot\|_{H^k(S)}$ and $|\cdot|_{H^k(S)}$.
Moreover, in the case $p = 2$ and $k = 1$ we let $H^1_\Gamma(\Omega)$ be the subspace of functions of $H^1(\Omega)$ satisfying homogeneous Dirichlet boundary conditions on a subset $\Gamma$ of the boundary $\partial\Omega$ of $\Omega$, with $\Gamma \neq \emptyset$. Finally, we recall that $L^\infty(\Omega)$ is the space of bounded functions a.e. in $\Omega$.
The remaining part of this section is devoted to the introduction of the anisotropic setting used to derive the anisotropic a posteriori error estimator in Section 3.1. The details of the anisotropic analysis we are referring to are covered essentially in [11, 12].

For any $0 < h \leq 1$, let $\{\mathcal{T}_h\}_h$ be a family of conforming triangulations of $\overline{\Omega}$ into triangles $K$ of diameter $h_K \leq h$. Since we are working with strongly anisotropic meshes, the standard regularity assumption on the mesh does not hold [7]. Let us introduce the standard invertible affine map $T_K : \widehat{K} \to K$ from the reference triangle $\widehat{K}$ to the general element $K$ of the triangulation $\mathcal{T}_h$ (see Fig. 1). Although the results in [11, 12] are independent of $\widehat{K}$, in the sequel, we identify $\widehat{K}$ with the unitary equilateral triangle $(-1/2, 0), (1/2, 0), (0, \sqrt{3}/2)$. This turns out to be a practical and rather standard choice for an anisotropic analysis [2, 9, 18, 22, 35].

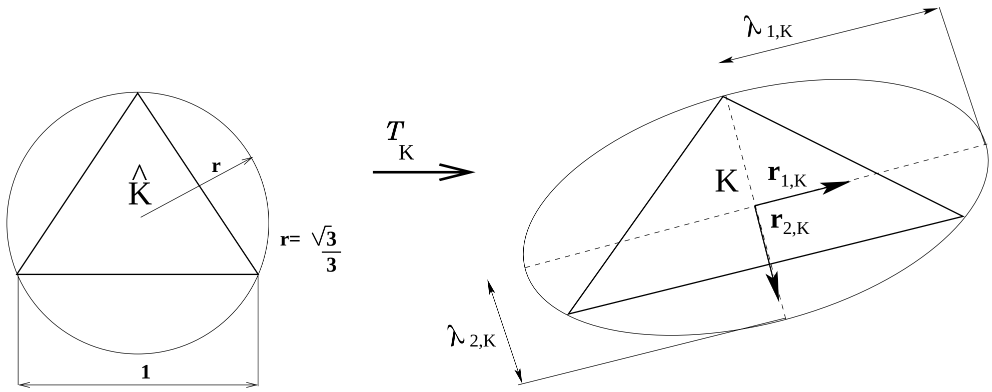


Figure 1: The affine map $T_K$.

For any $K \in \mathcal{T}_h$, let $M_K \in \mathbb{R}^{2\times 2}$ and $\vec{t}_K \in \mathbb{R}^2$ be the matrix and the vector defining the map $T_K$, that is, for any $\vec{\widehat{x}} = (\widehat{x}_1, \widehat{x}_2)^T \in \widehat{K}$,

$$\vec{x} = (x_1, x_2)^T = T_K(\vec{\widehat{x}}) = M_K \vec{\widehat{x}} + \vec{t}_K \in K\,. \tag{5}$$

The anisotropic information about the size and the orientation of the mesh element $K$ are derived by the spectral properties of the map $T_K$. In more detail, let us consider the polar decomposition $M_K = B_K\, Z_K$ of the matrix $M_K$ in (5), with $B_K$ and $Z_K \in \mathbb{R}^{2\times 2}$ symmetric positive definite and orthogonal matrices, respectively (see, e.g., [17]). Then let us factorize the matrix $B_K$ in terms of its eigenvalues $\lambda_{i,K}$ and eigenvectors $\vec{r}_{i,K}$ to obtain $M_K = R_K^T\, \Lambda_K\, R_K\, Z_K$, with $R_K^T = [\vec{r}_{1,K}\ \vec{r}_{2,K}]$ and $\Lambda_K = \mathrm{diag}(\lambda_{1,K},\, \lambda_{2,K})$. In the sequel the non-restrictive assumption $\lambda_{1,K} \geq \lambda_{2,K}$ is made.
The deformation of any $K \in \mathcal{T}_h$ with respect to $\widehat{K}$ can thus be measured in terms of the quantities $\lambda_{i,K}$ by defining the so-called stretching factor $s_K = \lambda_{1,K}/\lambda_{2,K}$ ($\geq 1$), $s_{\widehat{K}}$ being equal to one. Notice that the matrix $B_K$ and all the quantities related to it are independent of the local numbering of the nodes of $K$ only when $\widehat{K}$ is the (isotropic) equilateral triangle.

In view of the a posteriori error analysis below, after introducing the finite element space $W_h \subset H^1(\Omega)$ consisting of piecewise continuous polynomials of degree one, let $I_h^1 : L^2(\Omega) \to W_h$ be the standard Clément linear interpolant [8], and let $I_K^1$ be its restriction to $K$, for any $K \in \mathcal{T}_h$. Throughout two requirements are made on the patch $\Delta_K$ involved in the definition of the operator $I_h^1$, $\Delta_K$ being the union of all the elements sharing a vertex with $K$. We assume the cardinality of any patch $\Delta_K$ as well as the diameter of the reference patch $\Delta_{\widehat{K}} = T_K^{-1}(\Delta_K)$ to be uniformly bounded, independently of the geometry of the mesh, i.e., for any $K \in \mathcal{T}_h$,

$$\mathrm{card}(\Delta_K) < N \quad \text{and} \quad \mathrm{diam}(\Delta_{\widehat{K}}) \leq C_\Delta \simeq O(1), \tag{6}$$

with $C_\Delta \geq h_{\widehat{K}}$ [26]. In particular, the latter hypothesis rules out some too distorted reference patches (see Fig. 2 where examples of acceptable and non-acceptable patches are provided).

For the Clèment operator $I_h^1$ the following anisotropic interpolation error estimates can be proved [11, 12].

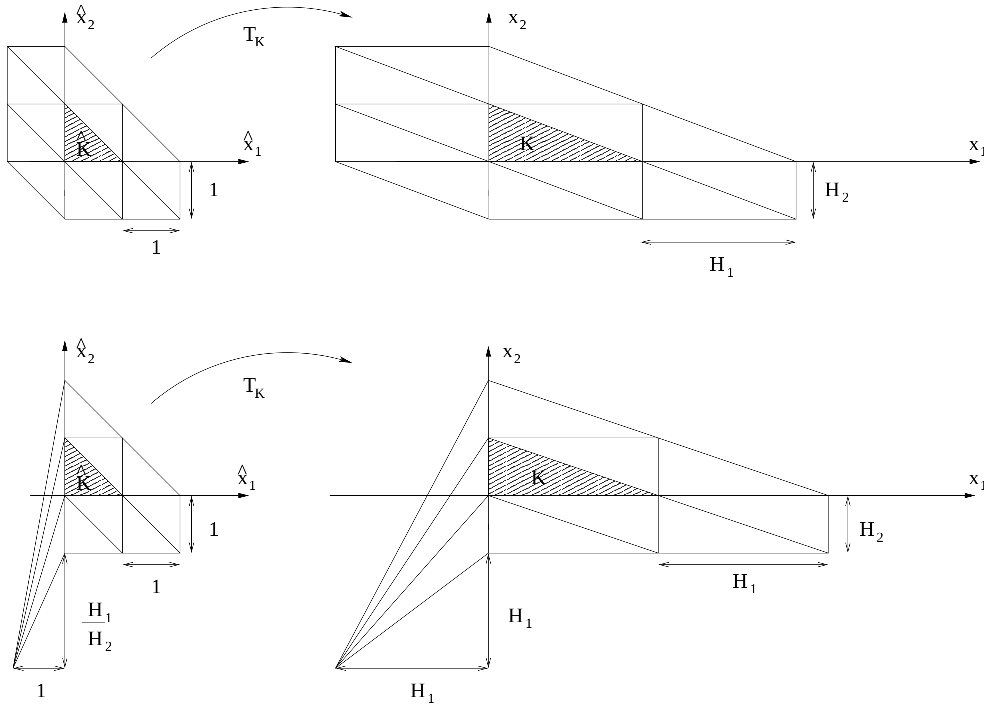


Figure 2: Examples of an acceptable (top) and of a non-acceptable (bottom) patch.

**Lemma 2.1** *Let $v \in H^1(\Omega)$. Then there exist two constants $\mathcal{C}_1 = \mathcal{C}_1(N, C_\Delta)$ and $\mathcal{C}_2 = \mathcal{C}_2(N, C_\Delta)$ such that, for any $K \in \mathcal{T}_h$,*

$$\begin{aligned} \|v - I_K^1(v)\|_{L^2(K)} &\le \mathcal{C}_1 \Big[ \sum_{i=1}^{2} \lambda_{i,K}^2 \left( \vec{r}_{i,K}^{\,T}\, G_K(v)\, \vec{r}_{i,K} \right) \Big]^{1/2}, \\ \|v - I_K^1(v)\|_{L^2(\partial K)} &\le \mathcal{C}_2\, h_K^{1/2} \left[ s_K \left( \vec{r}_{1,K}^{\,T}\, G_K(v)\, \vec{r}_{1,K} \right) + \frac{1}{s_K} \left( \vec{r}_{2,K}^{\,T}\, G_K(v)\, \vec{r}_{2,K} \right) \right]^{1/2}, \end{aligned} \tag{7}$$

*$G_K(v)$ being the symmetric positive semi-definite matrix given by*

$$G_K(v) = \sum_{T \in \Delta_K} \begin{bmatrix} \displaystyle\int_T \left( \frac{\partial v}{\partial x_1} \right)^2 d\vec{x} & \displaystyle\int_T \frac{\partial v}{\partial x_1} \frac{\partial v}{\partial x_2}\, d\vec{x} \\ \displaystyle\int_T \frac{\partial v}{\partial x_1} \frac{\partial v}{\partial x_2}\, d\vec{x} & \displaystyle\int_T \left( \frac{\partial v}{\partial x_2} \right)^2 d\vec{x} \end{bmatrix}. \tag{8}$$

**Remark 2.1** *Estimates* (7) *hold also for a more general Clèment like operator such as, for instance, the Scott-Zhang interpolant* [34]*. In such a case it suffices to suitably modify the definition of the patch $\Delta_K$ in* (8)*.*

## 3 The model problem

Let us consider the model elliptic boundary value problem: find $u : \Omega \to \mathbb{R}$ such that

$$\begin{cases} -\displaystyle\sum_{i,\,j=1}^{2} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) + \gamma u = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_D, \\ \dfrac{\partial u}{\partial \vec{n}_L} = g & \text{on } \Gamma_N, \end{cases} \tag{9}$$

where $\Gamma_D$ and $\Gamma_N$, with $\Gamma_D \neq \emptyset$, denote two disjoint boundary segments such that $\overline{\Gamma}_D \cup \overline{\Gamma}_N = \partial\Omega$; $f \in L^2(\Omega)$, $g \in L^2(\Gamma_N)$, $\gamma = \gamma(\vec{x}) \geq 0$ a.e. in $\Omega$ and $a_{ij} = a_{ij}(\vec{x}) = a_{ji}(\vec{x}) \in L^\infty(\Omega)$ are given functions, and

$$\frac{\partial u}{\partial \vec{n}_L} = \sum_{i,\,j=1}^{2} a_{ij}\frac{\partial u}{\partial x_j} n_i$$

is the conormal derivative of $u$, $\vec{n} = (n_1, n_2)^T$ being the unit outward normal vector to the boundary $\partial\Omega$ of the domain $\Omega$. Moreover, we assume that the differential operator defined in (9) is elliptic, i.e. that there exists a constant $\delta > 0$ such that

$$\sum_{i,\,j=1}^{2} a_{ij}(\vec{x})\, \xi_i\, \xi_j \geq \delta\, \| \vec{\xi}\, \|_2^2$$

for any $\vec{\xi} = (\xi_1, \xi_2)^T \in \mathbb{R}^2$ and for a.e. $\vec{x} \in \Omega$, $\| \cdot \|_2$ denoting the standard Euclidean norm. The weak form of (9) reads: find $u \in V \equiv H^1_{\Gamma_D}(\Omega)$ such that

$$B(u, v) = L(v) \quad \text{for any} \quad v \in V, \tag{10}$$

where

$$B(u, v) = \int_\Omega \Big( \sum_{i,\,j=1}^{2} a_{ij}\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i} + \gamma\, u\, v \Big)\, d\vec{x} \quad \text{and} \quad L(v) = \int_\Omega f\, v\, d\vec{x} + \int_{\Gamma_N} g\, v\, ds.$$

The hypotheses made above on the data of problem (9) guarantee the existence and the uniqueness of the solution $u$ of the weak formulation (10). Let us endow the space $V$ with the energy norm $|||\cdot|||$ defined by

$$|||v||| = [B(v, v)]^{1/2} \quad \text{for any } v \in V. \tag{11}$$

In what follows, without any explicit specification, we will refer such a norm to the whole domain $\Omega$. Otherwise the considered subset of $\Omega$ will be specified by a corresponding subscript. Let us introduce the subspace $V_h \subset V$ consisting of piecewise continuous polynomials of maximum degree one [7, 33]. Then the discrete form of (10) reads: find $u_h \in V_h$ such that

$$B(u_h, v_h) = L(v_h) \quad \text{for any} \quad v_h \in V_h. \tag{12}$$

Existence and uniqueness of $u_h$ are again guaranteed by the hypotheses made above on the data of problem (9). Recalling that $e_h = u - u_h$ is the discretization error associated with the finite element solution $u_h$, this quantity satisfies the so-called Galerkin orthogonality property given by

$$B(e_h, v_h) = 0 \quad \text{for any} \quad v_h \in V_h. \tag{13}$$

Moving from [27, 31], we are now in a position to build the desired anisotropic counterpart of the standard ZZ error estimator [40, 41, 42, 43].

### 3.1 An anisotropic recovery-based a posteriori error estimator

First, let us introduce some quantities used below. For any $K \in \mathcal{T}_h$, let

$$r_K(u_h) = \Big( f + \sum_{i,\,j=1}^{2} \frac{\partial}{\partial x_i} \Big( a_{ij} \frac{\partial u_h}{\partial x_j} \Big) - \gamma\, u_h \Big) \Big|_K \tag{14}$$

and

$$R_K(u_h) = \begin{cases} 0 & \text{for any } \ E \in \mathcal{E}(K) \cap \mathcal{E}_{h,D}, \\ 2\Big( g - \dfrac{\partial u_h}{\partial \vec{n}_{L,K}} \Big) \Big|_E & \text{for any } \ E \in \mathcal{E}(K) \cap \mathcal{E}_{h,N}, \\ -\Big[ \dfrac{\partial u_h}{\partial \vec{n}_{L,K}} \Big]_E & \text{for any } \ E \in \mathcal{E}(K) \cap \mathcal{E}_{h,\Omega} \end{cases} \tag{15}$$

be the element interior and boundary residuals, respectively associated with the finite element approximation $u_h$. We have distinguished the edges $E$ constituting the skeleton $\mathcal{E}_h$ of the triangulation $\mathcal{T}_h$ as $\mathcal{E}_{h,\Omega}$, $\mathcal{E}_{h,D}$ and $\mathcal{E}_{h,N}$ according to whether they are internal, Dirichlet or Neumann edges, respectively, while with $\mathcal{E}(K)$ we let the set of the edges of the generic triangle $K$. Moreover,

$$\Big[ \frac{\partial u_h}{\partial \vec{n}_{L,K}} \Big]_E = \frac{\partial u_h}{\partial \vec{n}_{L,K}} + \frac{\partial u_h}{\partial \vec{n}_{L,K'}} \qquad \text{for any } E \in \mathcal{E}(K) \cap \mathcal{E}_{h,\Omega},$$

where $K'$ is the triangle sharing the edge $E$ with $K$ and $\partial u_h / \partial \vec{n}_{L,K}$ and $\partial u_h / \partial \vec{n}_{L,K'}$ denote the conormal derivatives of $u_h$ associated with the elements $K$ and $K'$, respectively.
First, let us prove an *implicit* estimate for the energy norm of the discretization error, where implicit is understood in the sense mentioned in Section 1.

**Proposition 3.1** *Let $u$ be the solution of* (10) *and $u_h$ be the corresponding finite element approximation, solution of* (12)*. Then there exists a constant $C = C(N, C_\Delta)$ such that*

$$|||e_h||| \le C \Big( \sum_{K \in \mathcal{T}_h} \alpha_K\, \rho_K(u_h)\, w_K(e_h) \Big)^{1/2}, \tag{16}$$

*where*

$$\begin{aligned} &\alpha_K = \lambda_{1,K}^{1/2}\, \lambda_{2,K}^{1/2}, \\ &\rho_K(u_h) = \|r_K(u_h)\|_{L^2(K)} + \frac{1}{2\,\lambda_{2,K}^{1/2}} \|R_K(u_h)\|_{L^2(\partial K)}, \\ &w_K(e_h) = \Big[ s_K \left( \vec{r}_{1,K}^{\,T}\, G_K(e_h)\, \vec{r}_{1,K} \right) + \frac{1}{s_K} \left( \vec{r}_{2,K}^{\,T}\, G_K(e_h)\, \vec{r}_{2,K} \right) \Big]^{1/2}, \end{aligned} \tag{17}$$

*$G_K$ is the matrix defined in* (8)*, and $r_K(u_h)$ and $R_K(u_h)$ are given by* (14) *and* (15)*, respectively.*

**Proof.** Let us suitably rewrite the bilinear form $B(e_h, v)$: for any $v \in V$,

$$\begin{aligned} B(e_h, v) &= \sum_{K \in \mathcal{T}_h} \int_K \Big( \sum_{i,\,j=1}^{2} a_{ij} \frac{\partial e_h}{\partial x_j} \frac{\partial v}{\partial x_i} + \gamma\, e_h\, v \Big)\, d\vec{x} \\ &= \sum_{K \in \mathcal{T}_h} \Big\{ \int_K f\, v\, d\vec{x} + \int_{\partial K \cap \Gamma_N} g\, v\, ds \Big\} - \sum_{K \in \mathcal{T}_h} \int_K \Big( \sum_{i,\,j=1}^{2} a_{ij} \frac{\partial u_h}{\partial x_j} \frac{\partial v}{\partial x_i} + \gamma\, u_h\, v \Big)\, d\vec{x}. \end{aligned} \tag{18}$$

An integration by parts of the integral in the last sum yields

$$
\begin{aligned}
-\sum_{K\in\mathcal{T}_h} \int_K \Big( \sum_{i,\,j=1}^{2} a_{ij} \frac{\partial u_h}{\partial x_j} \frac{\partial v}{\partial x_i} + \gamma\, u_h\, v \Big)\, d\vec{x} = \sum_{K\in\mathcal{T}_h} \Big\{ \int_K \Big( \sum_{i,\,j=1}^{2} \frac{\partial}{\partial x_i} \Big( a_{ij} \frac{\partial u_h}{\partial x_j} \Big) \\
- \gamma\, u_h \Big) v\, d\vec{x} \;-\; \int_{\partial K \cap \Gamma_N} \frac{\partial u_h}{\partial \vec{n}_{L,K}}\, v\, ds \;-\; \int_{\partial K \cap \mathcal{E}_{h,\Omega}} \frac{\partial u_h}{\partial \vec{n}_{L,K}}\, v\, ds \Big\}.
\end{aligned}
$$

Thus, going back to (18) and thanks to (14) and (15), we get

$$
\begin{aligned}
B(e_h, v) &= \sum_{K\in\mathcal{T}_h} \Big\{ \int_K \Big( f + \sum_{i,\,j=1}^{2} \frac{\partial}{\partial x_i} \Big( a_{ij} \frac{\partial u_h}{\partial x_j} \Big) - \gamma\, u_h \Big) v\, d\vec{x} \\
&+ \int_{\partial K \cap \Gamma_N} \Big( g - \frac{\partial u_h}{\partial \vec{n}_{L,K}} \Big) v\, ds \;-\; \int_{\partial K \cap \mathcal{E}_{h,\Omega}} \frac{\partial u_h}{\partial \vec{n}_{L,K}} v\, ds \Big\} \\
&= \sum_{K\in\mathcal{T}_h} \Big\{ \int_K r_K(u_h)\, v\, d\vec{x} + \frac{1}{2} \int_{\partial K} R_K(u_h)\, v\, ds \Big\}.
\end{aligned}
$$

Now the Galerkin orthogonality property (13) (with $v_h = I_h^1(v)$) together with the Cauchy-Schwarz inequality and Lemma 2.1 provide the estimate

$$
\begin{aligned}
|B(e_h, v)| &= \big| B\big(e_h, v - I_h^1(v)\big) \big| \le \sum_{K\in\mathcal{T}_h} \Big\{ \|r_K(u_h)\|_{L^2(K)}\, \|v - I_K^1(v)\|_{L^2(K)} \\
&+ \frac{1}{2} \|R_K(u_h)\|_{L^2(\partial K)}\, \|v - I_K^1(v)\,\|_{L^2(\partial K)} \Big\} \\
&\le C \sum_{K\in\mathcal{T}_h} \Big\{ \|r_K(u_h)\|_{L^2(K)} \Big[ \sum_{i=1}^{2} \lambda_{i,K}^2 \left( \vec{r}_{i,K}^{\,T}\, G_K(v)\, \vec{r}_{i,K} \right) \Big]^{1/2} \\
&+ \frac{h_K^{1/2}}{2} \|R_K(u_h)\|_{L^2(\partial K)} \Big[ s_K \left( \vec{r}_{1,K}^{\,T}\, G_K(v)\, \vec{r}_{1,K} \right) + \frac{1}{s_K} \left( \vec{r}_{2,K}^{\,T}\, G_K(v)\, \vec{r}_{2,K} \right) \Big]^{1/2} \Big\} \\
&\le C \sum_{K\in\mathcal{T}_h} \Big\{ \Big[ \lambda_{1,K}^{1/2}\, \lambda_{2,K}^{1/2} \|r_K(u_h)\|_{L^2(K)} + \frac{\lambda_{1,K}^{1/2}}{2} \|R_K(u_h)\|_{L^2(\partial K)} \Big] \\
&\qquad \Big[ s_K \left( \vec{r}_{1,K}^{\,T}\, G_K(v)\, \vec{r}_{1,K} \right) + \frac{1}{s_K} \left( \vec{r}_{2,K}^{\,T}\, G_K(v)\, \vec{r}_{2,K} \right) \Big]^{1/2} \Big\},
\end{aligned}
$$

i.e. result (16) after choosing $v = e_h$. Notice that in the last inequality the geometrical relation

$$
h_E \le h_K \le h_{\widehat{K}} \lambda_{1,K}, \qquad \text{for any } K \in \mathcal{T}_h, \tag{19}
$$

has been exploited also. $\square$

To get information from estimate (16), we replace the matrix $G_K(e_h)$ in the definition of the weights $w_K(e_h)$ with a computable quantity, $e_h$ depending on the unknown solution $u$. With this aim, we exploit the ZZ recovery technique and replace $G_K(e_h)$ with the new matrix $G_K(e_h^*)$ defined by

$$
(G_K(e_h^*))_{ij} = \sum_{T\in\Delta_K} \int_T \Big( G_i u_h - \frac{\partial u_h}{\partial x_i} \Big) \Big( G_j u_h - \frac{\partial u_h}{\partial x_j} \Big)\, d\vec{x} \quad \text{with } i,\, j = 1,\, 2, \tag{20}
$$

$G^{ZZ} u_h = (G_1 u_h, G_2 u_h)^T \in (W_h)^2$ denoting the ZZ recovered gradient.

**Remark 3.1** *Throughout we employ the following definition of $G^{ZZ}u_h$:*

$$G^{ZZ}u_h(\vec{x}_i) = \frac{1}{|\Delta_i|} \sum_{T \in \Delta_i} |T| \, \nabla u_h\big|_T,$$

*where $\Delta_i$ is the patch of elements sharing the generic node $\vec{x}_i$, and $|T|$, $|\Delta_i|$ are the measures of $T$ and $\Delta_i$, respectively. This corresponds to an approximate $L^2$-projection, where the scalar product are evaluated using the trapezoidal quadrature formula.*

Matrix (20) allows us to provide the definition below.

**Definition 3.1** *An anisotropic a posteriori error estimator for the energy norm of the discretization error $e_h$ associated with the finite element approximation $u_h$ of problem* (10)*, is given by the quantity*

$$\eta = \Big( \sum_{K \in \mathcal{T}_h} \eta_K^2 \Big)^{1/2}, \tag{21}$$

*where $\eta_K = \big(\alpha_K \, \rho_K(u_h) \, w_K(e_h^*)\big)^{1/2}$ is the* element error indicator*. According to* $(17)_3$*,*

$$w_K(e_h^*) = \left[ s_K \left(\vec{r}_{1,K}^{\,T} \, G_K(e_h^*) \, \vec{r}_{1,K}\right) \, + \, \frac{1}{s_K} \left(\vec{r}_{2,K}^{\,T} \, G_K(e_h^*) \, \vec{r}_{2,K}\right) \right]^{1/2}, \tag{22}$$

*while $\alpha_K$ and $\rho_K(u_h)$ are defined as in* $(17)_1$ *and* $(17)_2$*, respectively.*

We point out that the error estimator (21) is of residual type (see, e.g., [1, 36]). However, though computationally cheap, it allows us to estimate only the energy norm of the discretization error $e_h$. If linear functionals of $e_h$ have to be controlled, dual-based error estimators, involving the solution of a suitable adjoint problem, should be considered [4, 16, 30]. Anisotropic error estimates for the control of linear functionals of the discretization error are considered, for instance, in [12, 13, 14, 28].

# 4 Foreword to the analysis

Given a generic estimator $\eta$ of the discretization error $e_h$ in the energy norm, checking the robustness of $\eta$ means verifying its efficiency and reliability i.e. the existence of two strictly positive constants $\overline{C}$, $\underline{C}$, independent of the mesh size, such that

$$|||e_h||| \leq \overline{C} \, \eta \, + \, H.O.T._1 \quad \text{(reliability)} \tag{23}$$

and

$$\eta_K \leq \underline{C} \, |||e_h|||_{\Delta_K} + H.O.T._2 \quad \text{(efficiency)}, \tag{24}$$

where $\eta_K$ is the element error indicator associated with $\eta$ and $H.O.T._i$, with $i = 1, 2$, are higher order terms related to the data oscillations (see [29]).

Essentially, relations (23) and (24) state the upper (global) and lower (local) boundedness of the energy norm of $e_h$ in terms of the global and of the element error indicators $\eta$ and $\eta_K$, respectively.

The reliability and the efficiency of the error estimator $\eta$ in (21) are analyzed in Sections 5 and 6, respectively after making some simplifying choices on the starting problem (9): the diffusive matrix $A = \{a_{ij}\}$ is assumed constant, while the reaction term $\gamma u$ is neglected. Thus, the element internal residual reduces to $r_K(u_h) = f|_K$ because of the

choice of the finite element space.
Then we replace the data $f$ and $g$, usually not exactly integrable, by suitable functions $f_K$ and $g_E$, piecewise constant with respect to $\mathcal{T}_h$ and $\mathcal{E}_{h,N}$, respectively [29, 36]. The element error indicator $\eta_K$ in (21) can thus be explicitly rewritten as

$$\begin{aligned}\eta_{K,T} &= \Big[\lambda_{1,K}^{1/2}\lambda_{2,K}^{1/2}\,\|f_K\|_{L^2(K)} + \frac{\lambda_{1,K}^{1/2}}{2}\sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,\Omega}}\Big\|\Big[\frac{\partial u_h}{\partial\vec{n}_{L,K}}\Big]_E\Big\|_{L^2(E)} \\ &+ \lambda_{1,K}^{1/2}\sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,N}}\Big\|\Big(g_E - \frac{\partial u_h}{\partial\vec{n}_{L,K}}\Big)\Big|_E\Big\|_{L^2(E)}\Big]^{1/2}\big(w_K(e_h^*)\big)^{1/2},\end{aligned} \tag{25}$$

$\eta_{K,T}$ being now an exactly computable quantity.
Finally, the reliability and the efficiency of the error indicator

$$\eta_T = \Big(\sum_{K\in\mathcal{T}_h}\eta_{K,T}^2\Big)^{1/2} \tag{26}$$

will be studied under the additional

**Assumption 4.1** *For any $K\in\mathcal{T}_h$,*

$$\Big\|\frac{\partial u}{\partial x_i} - G_i u_h\Big\|_{L^2(\Delta_K)} \le \nu_K\,\Big\|\frac{\partial e_h}{\partial x_i}\Big\|_{L^2(\Delta_K)} \quad \textit{for} \quad i=1,\,2, \tag{27}$$

*with $\nu_K$'s constants such that $0\le\nu_K<1$.*

Assumption 4.1 is not so unusual in the literature. It is the main idea of the ZZ recovery technique, i.e. that the reconstruction $G^{ZZ}u_h$ of the approximation $\nabla u_h$ for the gradient $\nabla u$ turns out to be better than $\nabla u_h$ itself. However, we remark that assumption (27) is stronger than what is usually assumed, that is $\|\nabla u - G^{ZZ}u\|_{L^2(\Omega)} \le \nu\|\nabla e_h\|_{L^2(\Omega)}$, with $0\le\nu<1$ [6, 36].

## 4.1 Some useful results

Let us provide some results used in the sequel to assess the reliability and the efficiency of the error estimator (26).

**Lemma 4.1** *For any function $v\in H^1(\Delta_K)$ and for any $\alpha,\beta>0$, it holds*

$$\min(\alpha,\beta) \le \frac{\alpha\left(\vec{r}_{1,K}^{\,T}\,G_K(v)\,\vec{r}_{1,K}\right) + \beta\left(\vec{r}_{2,K}^{\,T}\,G_K(v)\,\vec{r}_{2,K}\right)}{|v|^2_{H^1(\Delta_K)}} \le \max(\alpha,\beta),$$

*$G_K$ being the matrix defined in (8).*

**Proof.** Without loss of generality, let us assume that $\alpha\ge\beta$. Vice versa it suffices to exchange in the following the roles played by $\vec{r}_{1,K}$ and $\vec{r}_{2,K}$. As $\vec{r}_{1,K}$ and $\vec{r}_{2,K}$ are orthonormal (eigen)vectors, let $\vec{r}_{1,K}=[c\ \ s]^T$ and $\vec{r}_{2,K}=[-s\ \ c]^T$, with $c=\cos\theta, s=\sin\theta$ and $\theta\in[0,\pi[$.

Let $W(v)$ be the matrix with components $\big(W(v)\big)_{i,j} = (\partial v/\partial x_i)\,(\partial v/\partial x_j)$, for $i,j = 1,2$. Then we have

$$
\begin{aligned}
& \alpha\,(\vec{r}_{1,K}^{\,T}\,W(v)\,\vec{r}_{1,K}) + \beta\,(\vec{r}_{2,K}^{\,T}\,W(v)\,\vec{r}_{2,K}) \\
= \;& \alpha\Big[\Big(\frac{\partial v}{\partial x_1}\Big)^2 c^2 + 2\,\frac{\partial v}{\partial x_1}\,\frac{\partial v}{\partial x_2}\,s\,c + \Big(\frac{\partial v}{\partial x_2}\Big)^2 s^2\Big] \\
+ \;& \beta\Big[\Big(\frac{\partial v}{\partial x_1}\Big)^2 s^2 - 2\,\frac{\partial v}{\partial x_1}\,\frac{\partial v}{\partial x_2}\,s\,c + \Big(\frac{\partial v}{\partial x_2}\Big)^2 c^2\Big] \\
= \;& \Big(\frac{\partial v}{\partial x_1}\Big)^2(\alpha\,c^2 + \beta\,s^2) + 2\,\frac{\partial v}{\partial x_1}\,\frac{\partial v}{\partial x_2}\,(\alpha-\beta)\,s\,c + \Big(\frac{\partial v}{\partial x_2}\Big)^2(\alpha\,s^2 + \beta\,c^2) \\
= \;& \begin{bmatrix} \dfrac{\partial v}{\partial x_1} & \dfrac{\partial v}{\partial x_2} \end{bmatrix}
\begin{bmatrix} \alpha\,c^2 + \beta\,s^2 & (\alpha-\beta)s\,c \\ (\alpha-\beta)s\,c & \alpha\,s^2 + \beta\,c^2 \end{bmatrix}
\begin{bmatrix} \dfrac{\partial v}{\partial x_1} \\[2mm] \dfrac{\partial v}{\partial x_2} \end{bmatrix}.
\end{aligned}
$$

Thus we are led to bound the last term of the chain of equalities above to obtain

$$
\begin{aligned}
& \beta\Big[\Big(\frac{\partial v}{\partial x_1}\Big)^2 + \Big(\frac{\partial v}{\partial x_2}\Big)^2\Big] \\
& \le \begin{bmatrix} \dfrac{\partial v}{\partial x_1} & \dfrac{\partial v}{\partial x_2} \end{bmatrix}
\begin{bmatrix} \alpha\,c^2 + \beta\,s^2 & (\alpha-\beta)s\,c \\ (\alpha-\beta)s\,c & \alpha\,s^2 + \beta\,c^2 \end{bmatrix}
\begin{bmatrix} \dfrac{\partial v}{\partial x_1} \\[2mm] \dfrac{\partial v}{\partial x_2} \end{bmatrix} \\
& \le \alpha\Big[\Big(\frac{\partial v}{\partial x_1}\Big)^2 + \Big(\frac{\partial v}{\partial x_2}\Big)^2\Big]
\end{aligned}
$$

since $\alpha$ and $\beta$ are easily shown to be the maximum and minimum eigenvalue of the matrix above. After integrating over $\Delta_K$ the thesis follows. $\square$

Analogously to Lemma 4.1, we have

**Lemma 4.2** *For any $\alpha, \beta > 0$, we have*

$$
\min(\alpha,\beta) \le \frac{\alpha\,\big(\vec{r}_{1,K}^{\,T}\,G_K(e_h^*)\,\vec{r}_{1,K}\big) + \beta\,\big(\vec{r}_{2,K}^{\,T}\,G_K(e_h^*)\,\vec{r}_{2,K}\big)}{\|G^{ZZ}u_h - \nabla u_h\|^2_{L^2(\Delta_K)}} \le \max(\alpha,\beta), \tag{28}
$$

*$G_K(e_h^*)$ being defined as in* (20).

**Proof.** It is enough to repeat the proof of Lemma 4.1 simply by replacing the matrix $W(v)$ with the matrix of components $\Big(G_i u_h - \dfrac{\partial u_h}{\partial x_i}\Big)\Big(G_j u_h - \dfrac{\partial u_h}{\partial x_j}\Big)$. $\square$

Notice that both the upper and lower bounds in Lemmas 4.1 and 4.2 are sharp.

Let us derive now a relation between $\nabla e_h$ and the corresponding "recovered" quantity $(G^{ZZ}u_h - \nabla u_h)$.

**Lemma 4.3** *Under the Assumption* 4.1 *and for any $K \in \mathcal{T}_h$, we have that*

$$
\frac{1}{(1+\nu_K)}\,\|G^{ZZ}u_h - \nabla u_h\|_{L^2(\Delta_K)} \le \|\nabla e_h\|_{L^2(\Delta_K)} \le \frac{1}{(1-\nu_K)}\,\|G^{ZZ}u_h - \nabla u_h\|_{L^2(\Delta_K)}. \tag{29}
$$

**Proof.** From (27) and thanks to the triangle inequality, we deduce that, for $i = 1, 2$,

$$
\Big\|\frac{\partial e_h}{\partial x_i}\Big\|_{L^2(\Delta_K)} \le \nu_K\,\Big\|\frac{\partial e_h}{\partial x_i}\Big\|_{L^2(\Delta_K)} + \Big\|G_i u_h - \frac{\partial u_h}{\partial x_i}\Big\|_{L^2(\Delta_K)}
$$

that is

$$\Big\|\frac{\partial e_h}{\partial x_i}\Big\|_{L^2(\Delta_K)} \le \frac{1}{1-\nu_K}\;\Big\|G_i u_h - \frac{\partial u_h}{\partial x_i}\Big\|_{L^2(\Delta_K)}. \tag{30}$$

The upper bound in (29) immediately follows from (30). Likewise, it can be inferred that, for $i = 1,\, 2$,

$$\Big\|G_i u_h - \frac{\partial u_h}{\partial x_i}\Big\|_{L^2(\Delta_K)} \le (1+\nu_k)\;\Big\|\frac{\partial e_h}{\partial x_i}\Big\|_{L^2(\Delta_K)}$$

i.e.

$$\Big\|\frac{\partial e_h}{\partial x_i}\Big\|_{L^2(\Delta_K)} \ge \frac{1}{1+\nu_k}\;\Big\|G_i u_h - \frac{\partial u_h}{\partial x_i}\Big\|_{L^2(\Delta_K)}.$$

This completes the proof of (29). □

The next Lemma relates the $L^2(K)$-norm of the gradient of any function $v \in H^1(K)$ to the energy norm of $v$, on a generic triangle $K$.

**Lemma 4.4** *Let $\gamma_{max}$ and $\gamma_{min}$ denote the maximum and minimum eigenvalue of the diffusive matrix $A$, respectively. Then for any $K \in \mathcal{T}_h$ and for any $v \in H^1(K)$, the following equivalence can be proved*

$$\gamma_{min}^{1/2}\, \|\nabla v\|_{L^2(K)} \le |||v|||_K \le \gamma_{max}^{1/2}\, \|\nabla v\|_{L^2(K)}. \tag{31}$$

**Proof.** The simplifying hypotheses made at the beginning of this section reduce the definition (11) of the energy norm of $v$ on $K$ to

$$|||v|||_K^2 = B(v,v)\big|_K = \int_K \Big\{ a_{11}\Big(\frac{\partial v}{\partial x_1}\Big)^2 + 2\,a_{12}\frac{\partial v}{\partial x_1}\,\frac{\partial v}{\partial x_2} + a_{22}\Big(\frac{\partial v}{\partial x_2}\Big)^2 \Big\}\, d\vec{x}. \tag{32}$$

The integrand of (32) can be identified with the numerator of the Rayleigh quotient $\big((\nabla v)^T A\, \nabla v\big) \,/\, \big((\nabla v)^T\, \nabla v\big)$. Now, as

$$\gamma_{\min}\, (\nabla v)^T \nabla v \le (\nabla v)^T A\, \nabla v \le \gamma_{\max}\, (\nabla v)^T \nabla v,$$

we can integrate such relations on the triangle $K$ to obtain

$$\gamma_{\min}\, \|\nabla v\|_{L^2(K)}^2 \le |||v|||_K^2 \le \gamma_{\max}\, \|\nabla v\|_{L^2(K)}^2,$$

i.e. (31). □

**Remark 4.1** *Result* (31) *can be reformulated on any patch of elements $\Delta_K$ simply by extending the integration step from $K$ to $\Delta_K$ and under the assumption $v \in H^1(\Delta_K)$.*

The last result of this section turns out to be an essential tool in the proof of both the efficiency and the reliability of the error estimator (26).

**Lemma 4.5** *Under the Assumption* 4.1 *it can be proved that, for any $K \in \mathcal{T}_h$,*

$$\frac{1-\nu_K}{s_K^{1/2}\,\gamma_{max}^{1/2}}\, |||e_h|||_{\Delta_K} \le w_K(e_h^*) \le \frac{s_K^{1/2}\,(1+\nu_K)}{\gamma_{min}^{1/2}}\, |||e_h|||_{\Delta_K}, \tag{33}$$

*$\gamma_{max}$ and $\gamma_{min}$ being defined as in Lemma* 4.4.

**Proof.** First, let us exploit relation (28) by choosing $\alpha = s_K$ and $\beta = 1/s_K$. This yields

$$\frac{1}{s_K^{1/2}} \|G^{ZZ}u_h - \nabla u_h\|_{L^2(\Delta_K)} \leq w_K(e_h^*) \leq s_K^{1/2} \|G^{ZZ}u_h - \nabla u_h\|_{L^2(\Delta_K)}.$$

Now, for any $K \in \mathcal{T}_h$, from (29), we have

$$\frac{(1-\nu_K)}{s_K^{1/2}} \|\nabla e_h\|_{L^2(\Delta_K)} \leq w_K(e_h^*) \leq s_K^{1/2}(1+\nu_K) \|\nabla e_h\|_{L^2(\Delta_K)}$$

which immediately provides (33) thanks to Lemma 4.4 extended to the whole patch $\Delta_K$. □

## 4.2 Anisotropic bubble functions

Typically, the efficiency of an error estimator is studied by using the properties of bubble functions (see, e.g., [1, 36]).
In particular, we base the efficiency analysis of Section 6 on a new type of bubble functions which we define *anisotropic bubble functions*. This turns out to be the main novelty of our analysis.
Like in the case of the standard bubble functions, we distinguish between triangle and edge bubble functions, denoted in the sequel with $b_K^A$ and $b_E^A$, respectively. We define both types of bubbles through the solution of suitable eigenvalue problems.

For any $K \in \mathcal{T}_h$, the triangle anisotropic bubble function $b_K^A$ is defined by the relation $b_K^A = b_{\widehat{K}} \circ T_K^{-1}$, where $b_{\widehat{K}}$ solves the problem

$$\begin{cases} -\Delta b_{\widehat{K}} = \widehat{\lambda}\, b_{\widehat{K}} & \text{in } \widehat{K} \\ b_{\widehat{K}} = 0 & \text{on } \partial\widehat{K}, \end{cases} \tag{34}$$

$\Delta$ denoting the standard Laplacian operator. Thus $b_K^A$ is determined by computing the eigenfunction $b_{\widehat{K}}$ in $\widehat{K}$ associated with the Laplacian operator provided with homogeneous Dirichlet boundary conditions, and then mapping it to triangle $K$ via the transformation $T_K$. It is also understood that the eigenfunction $b_{\widehat{K}}$ corresponds to the least (positive) eigenvalue $\widehat{\lambda}$ of (34) and it is normalized such that $\max_{\vec{x}\in K} b_K^A(\vec{x}) = 1$ (see Fig. 3, left).

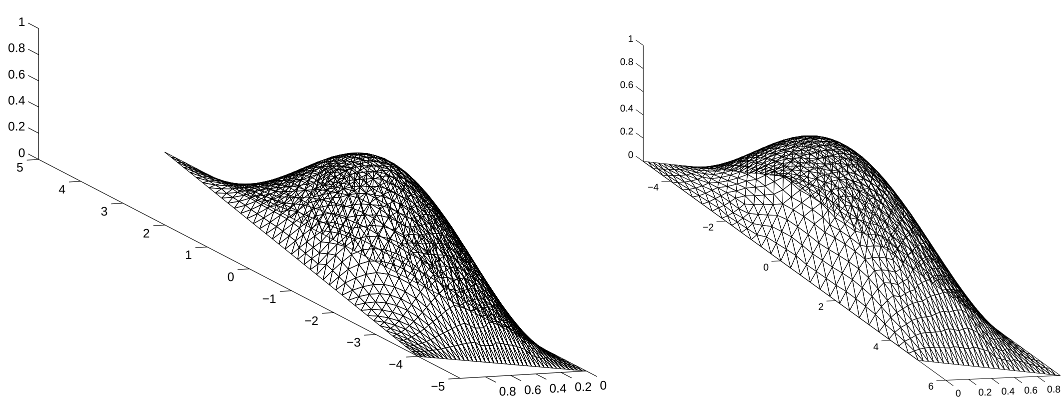

Figure 3: The anisotropic triangle (left) and edge (right) bubble functions $b_K^A$ and $b_E^A$ both corresponding to $s_K = 10$.

Likewise for any pair of triangles $K, K'$ sharing the edge $E \in \mathcal{E}_{h,\Omega}$, the edge bubble function $b_E^A$, supported in $Q_E = K \cup K'$, is defined by the relations $b_E^A\big|_K = b_{\widehat{E}} \circ T_K^{-1}$, $b_E^A\big|_{K'} = b_{\widehat{E}} \circ \widetilde{T}_{K'}^{-1}$, where $b_{\widehat{E}}$ solves the problem

$$\begin{cases} -\Delta b_{\widehat{E}} = \widetilde{\lambda}\, b_{\widehat{E}} & \text{in } \widehat{Q}_{\widehat{E}}, \\ b_{\widehat{E}} = 0 & \text{on } \partial \widehat{Q}_{\widehat{E}}, \end{cases} \tag{35}$$

$\widehat{Q}_{\widehat{E}}$ is the quadrilateral obtained by joining $\widehat{K}$ with the equilateral triangle $\widehat{K}'$, $(1/2, 0), (1, \sqrt{3}/2), (0, \sqrt{3}/2)$, along the side $\widehat{E}$ with extremes $(0, \sqrt{3}/2)$, $(1/2, 0)$, and $\widetilde{T}_{K'}$ is the map from triangle $\widehat{K}'$ to $K'$.
Notice that the choice made for $\widehat{E}$ is not restrictive, any other edge of $\widehat{K}$ will do, because of its symmetry. Moreover, since $\widehat{K}'$ coincides with $\widehat{K}$ up to a rigid motion, $\widetilde{T}_{K'}$ is characterized by the same eigenvalues $\lambda_{1,\,K'}$, $\lambda_{2,\,K'}$ as the ones obtained by mapping $\widehat{K}$ directly to $K'$ via the mapping $T_{K'}$. Finally, $\widetilde{T}_{K'}$ and $T_K$ coincide on $\widehat{E}$.
As above, it is understood that the eigenfunction $b_{\widehat{E}}$ corresponds to the least (positive) eigenvalue $\widetilde{\lambda}$ of (35) and it is normalized such that $\max\limits_{\vec{x} \in Q_E} b_E^A(\vec{x}) = 1$ (see Fig. 3, right).

Results corresponding to those stated for the standard bubble functions (see, e.g., [36]) can be proved also for $b_K^A$ and $b_E^A$. Let us summarize these properties in the following

**Lemma 4.6** *For any $K \in \mathcal{T}_h$, $E \in \mathcal{E}_{h,\Omega}$ and $Q_E = K \cup K'$, with $K$ and $K'$ triangles sharing the edge $E$, the following properties hold:*

$$supp\,(b_K^A) \subset K, \quad 0 \le b_K^A(\vec{x}) \le 1 \quad \textit{for any } \vec{x} \in K, \quad \max_{\vec{x} \in K} b_K^A(\vec{x}) = 1, \tag{36}$$

$$supp\,(b_E^A) \subset Q_E, \quad 0 \le b_E^A(\vec{x}) \le 1 \quad \textit{for any } \vec{x} \in Q_E, \quad \max_{\vec{x} \in E} b_E^A(\vec{x}) = 1, \tag{37}$$

$$\int_K b_K^A \, d\vec{x} = C_{\widehat{K}} \, |K|, \tag{38}$$

$$\int_E b_E^A \, ds = C_{\widehat{K}}^* \, h_E, \tag{39}$$

$$\int_T b_E^A \, d\vec{x} \le \widehat{C} \, s_T \, h_E^2 \qquad \textit{with } \; T \in \{K, K'\}, \tag{40}$$

$$\|\nabla b_K^A\|_{L^2(K)} = \frac{\widehat{\lambda}^{1/2} \, (1 + s_K^2)^{1/2}}{2^{1/2} \, \lambda_{1,K}} \, \|b_K^A\|_{L^2(K)} \,, \tag{41}$$

$$\|\nabla b_E^A\|_{L^2(T)} \le \frac{\widetilde{\lambda}^{1/2} \, s_T}{\lambda_{1,T}} \, \|b_E^A\|_{L^2(T)} \qquad \textit{with } \; T \in \{K, K'\}, \tag{42}$$

*where $C_{\widehat{K}}$, $C_{\widehat{K}}^*$, $\widehat{C}$ are constants depending only on the reference triangle $\widehat{K}$ and $s_T$ is the stretching factor of the element $T$.*

**Proof.** Let us start with the properties of the triangle bubble function $b_K^A$. Relations (36) follow immediately by the definition of $b_K^A$ and by the choice made for its normalization.

Let us now prove (38). Employing the relation $|K| = \lambda_{1,K}\lambda_{2,K}|\widehat{K}|$, we have

$$\int_K b_K^A \, d\vec{x} = \lambda_{1,K}\lambda_{2,K} \int_{\widehat{K}} b_{\widehat{K}} \, d\vec{\widehat{x}} = \frac{|K|}{|\widehat{K}|} \int_{\widehat{K}} b_{\widehat{K}} \, d\vec{\widehat{x}}.$$

Thus,

$$\int_K b_K^A \, d\vec{x} = C_{\widehat{K}} |K|,$$

where $C_{\widehat{K}} = \int_{\widehat{K}} b_{\widehat{K}} \, d\vec{\widehat{x}}/|\widehat{K}|$ depends on the reference triangle only.

To prove (41), the weak form of (34) immediately yields

$$\|\nabla b_{\widehat{K}}\|^2_{L^2(\widehat{K})} = \widehat{\lambda}\, \|b_{\widehat{K}}\|^2_{L^2(\widehat{K})} \, .$$

When passing to triangle $K$ this relation becomes

$$s_K \left(\vec{r}_{1,K}^{\,T} G_K(b_K^A)\vec{r}_{1,K}\right) + \frac{1}{s_K}\left(\vec{r}_{2,K}^{\,T} G_K(b_K^A)\vec{r}_{2,K}\right) = \frac{\widehat{\lambda}}{\lambda_{1,K}\lambda_{2,K}} \|b_K^A\|^2_{L^2(K)}, \tag{43}$$

where the summation implied by the definition of the matrix $G_K$ runs only on $K$ since $\mathrm{supp}(b_K^A) \subset K$. We first show that

$$s_K \left(\vec{r}_{1,K}^{\,T} G_K(b_K^A)\vec{r}_{1,K}\right) = \frac{1}{s_K}\left(\vec{r}_{2,K}^{\,T} G_K(b_K^A)\vec{r}_{2,K}\right). \tag{44}$$

The argument combines a slight modification of the proof of Lemma 2.2 in [11] with a symmetry argument. Let us introduce a unit vector $\vec{\chi}$. We express the directional derivative along direction $\vec{\chi}$ of a generic function $\widehat{v} \in H^1(\widehat{K})$, in terms of the derivatives of its image $v$ defined in $K$. We have,

$$\widehat{\nabla}\widehat{v} \cdot \vec{\chi} = \vec{\chi}^{\,T} M_K^T \nabla v = \vec{\chi}^{\,T} Z_K^T B_K \nabla v = \vec{\chi}^{\,T} Z_K^T R_K^T \Lambda_K R_K \nabla v \, .$$

It then follows

$$|\widehat{\nabla}\widehat{v} \cdot \vec{\chi}|^2 = |(R_K Z_K \vec{\chi})^T \Lambda_K R_K \nabla v|^2 \, .$$

Choosing $\vec{\chi} = \vec{\chi}_1$ such that $R_K Z_K \vec{\chi}_1 = (1,0)^T$, we obtain

$$\begin{aligned} \int_{\widehat{K}} |\widehat{\nabla}\widehat{v} \cdot \vec{\chi}_1|^2 \, d\vec{\widehat{x}} &= \int_{\widehat{K}} \lambda_{1,K}^2 (\vec{r}_{1,K}^{\,T} \nabla v)^2 \, d\vec{\widehat{x}} \\ &= \int_K \frac{\lambda_{1,K}}{\lambda_{2,K}} (\vec{r}_{1,K}^{\,T} \nabla v)^2 \, d\vec{x} = s_K \left(\vec{r}_{1,K}^{\,T} G_K(v)\vec{r}_{1,K}\right), \end{aligned} \tag{45}$$

while picking $\vec{\chi} = \vec{\chi}_2$ such that $R_K Z_K \vec{\chi}_2 = (0,1)^T$, we have

$$\begin{aligned} \int_{\widehat{K}} |\widehat{\nabla}\widehat{v} \cdot \vec{\chi}_2|^2 \, d\vec{\widehat{x}} &= \int_{\widehat{K}} \lambda_{2,K}^2 (\vec{r}_{2,K}^{\,T} \nabla v)^2 \, d\vec{\widehat{x}} \\ &= \int_K \frac{\lambda_{2,K}}{\lambda_{1,K}} (\vec{r}_{2,K}^{\,T} \nabla v)^2 \, d\vec{x} = \frac{1}{s_K} \left(\vec{r}_{2,K}^{\,T} G_K(v)\vec{r}_{2,K}\right). \end{aligned} \tag{46}$$

Notice that, by construction, $\vec{\chi}_1 \cdot \vec{\chi}_2 = 0$. Identifying in (45) and (46) $v$ with $b_K^A$, we infer that terms at the left and right-hand sides in (44) can be rewritten, with respect to the orthogonal directions $\vec{\chi}_1$ and $\vec{\chi}_2$, as

$$s_K \left(\vec{r}_{1,K} G_K(b_K^A)\vec{r}_{1,K}\right) = \int_{\widehat{K}} |\widehat{\nabla} b_{\widehat{K}} \cdot \vec{\chi}_1|^2 \, d\vec{\widehat{x}} \, ,$$
$$\frac{1}{s_K}\left(\vec{r}_{2,K}^{\,T} G_K(b_K^A)\vec{r}_{2,K}\right) = \int_{\widehat{K}} |\widehat{\nabla} b_{\widehat{K}} \cdot \vec{\chi}_2|^2 \, d\vec{\widehat{x}} \, ,$$

respectively. Finally, relation (44) follows on noticing that $b_{\widehat{K}}$ is invariant by rotations, thus implying that the two integral above are equal. Property (44) together with the relation

$$\vec{r}_{1,K}^{\,T} G_K(b_K^A)\vec{r}_{1,K} + \vec{r}_{2,K}^{\,T} G_K(b_K^A)\vec{r}_{2,K} = \|\nabla b_K^A\|^2_{L^2(K)}$$

and (43), allow us to conclude

$$\|\nabla b_K^A\|^2_{L^2(K)} = \frac{\widehat{\lambda}(1+s_K^2)}{2s_K\lambda_{1,K}\lambda_{2,K}}\,\|b_K^A\|^2_{L^2(K)} = \frac{\widehat{\lambda}(1+s_K^2)}{2\lambda_{1,K}^2}\,\|b_K^A\|^2_{L^2(K)}\,,$$

that is (41).

Let us now deal with the properties associated with the edge bubble function $b_E^A$. Relations (37) follow immediately from the definition and normalization of $b_E^A$, and observing that, also by a symmetry argument, the maximum of $b_{\widehat{E}}$ is assumed at the midpoint of $\widehat{E}$, which implies that the maximum of $b_E^A$ is taken at the midpoint of $E$ as well.

Equality (39) follows from

$$\int_E b_E^A\,ds = \frac{h_E}{h_{\widehat{E}}}\int_{\widehat{E}} b_{\widehat{E}}\,ds = C^*_{\widehat{K}} h_E\,,$$

the constant $C^*_{\widehat{K}}$ being defined as $C^*_{\widehat{K}} = h_{\widehat{E}}^{-1}\int_{\widehat{E}} b_{\widehat{E}}\,ds$.

Let us now prove (40) by choosing $T = K$. We have

$$\int_K b_E^A\,d\vec{x} = \lambda_{1,K}\lambda_{2,K}\int_{\widehat{K}} b_{\widehat{E}}\,d\vec{\widehat{x}} = C_{\widehat{E}}\frac{\lambda_{1,K}\lambda_{2,K}}{h_E^2}h_E^2 \le \widehat{C}s_K h_E^2\,,$$

where $C_{\widehat{E}} = \int_{\widehat{K}} b_{\widehat{E}}\,d\vec{\widehat{x}}$ and having also used the relation

$$\frac{1}{h_E} \le \frac{1}{\rho_K} \le \frac{1}{\rho_{\widehat{K}}\lambda_{2,K}}, \tag{47}$$

$\rho_K, \rho_{\widehat{K}}$ being the diameters of the balls inscribed in $K$ and $\widehat{K}$, respectively. Thus, $\widehat{C} = C_{\widehat{E}}/\rho^2_{\widehat{K}}$. Likewise an analogous relation holds when $T = K'$.

Finally, by the invariance of the domain $\widehat{Q}_{\widehat{E}}$ and of the operator $\Delta$ in (35) with respect to a rotation of an angle $\pi$, it follows that $\partial b_{\widehat{E}}/\partial\vec{n}_{\widehat{E}} = 0$, where $\vec{n}_{\widehat{E}}$ is the unit normal across $\widehat{E}$. This property allows us to rewrite (35) in each of the two triangles $\widehat{K}, \widehat{K}'$

$$\begin{cases} -\Delta b_{\widehat{E}} = \widetilde{\lambda} b_{\widehat{E}} & \text{in } \widehat{K} \\ \dfrac{\partial b_{\widehat{E}}}{\partial \vec{n}_{\widehat{E}}} = 0 & \text{on } \widehat{E} \\ b_{\widehat{E}} = 0 & \text{on } \partial\widehat{K}\setminus\widehat{E} \end{cases} \tag{48}$$

and similarly in $\widehat{K}'$. We are now in the same position as in the case of $b_{\widehat{K}}$ as the weak form of (48) gives

$$\|\nabla b_{\widehat{E}}\|^2_{L^2(\widehat{K})} = \widetilde{\lambda}\|b_{\widehat{E}}\|^2_{L^2(\widehat{K})}.$$

Thus a property analogous to (43) holds also for $b_E^A|_K$ and $b_E^A|_{K'}$, where it is understood that the summation implied by the definition of the matrices $G_K$ and $G_{K'}$ runs only on $K$ and $K'$, respectively. However, it is no longer possible to prove the analogue to (44) using the same argument, as $b_{\widehat{E}}|_{\widehat{K}}$ is not invariant by rotations in $\widehat{K}$. On the other hand, applying Lemma 4.1 on $K$ with $v = b_E^A|_K$, $\alpha = s_K$ and $\beta = 1/s_K$ (and similarly on $K'$), we obtain

$$s_K\left(\vec{r}_{1,K}^{\,T} G_K(b_E^A\big|_K)\vec{r}_{1,K}\right) + \frac{1}{s_K}\left(\vec{r}_{2,K}^{\,T} G_K(b_E^A\big|_K)\vec{r}_{2,K}\right) \ge \frac{1}{s_K}\,\|\nabla b_E^A\|^2_{L^2(K)},$$

and then, from the analogue to (43) for $b_E^A$,

$$||\nabla b_E^A||^2_{L^2(K)} \le \frac{\widetilde{\lambda}}{\lambda^2_{2,K}} ||b_E^A||^2_{L^2(K)} = \frac{\widetilde{\lambda}\, s_K^2}{\lambda^2_{1,K}}\, ||b_E^A||^2_{L^2(K)},$$

which proves (42). Notice that (42) is less sharp than (41), the latter being and equality and because

$$\Big(\frac{1+s_K^2}{2}\Big)^{1/2} \le s_K,$$

for any $s_K \ge 1$, with equality holding only when $s_K = 1$. $\square$

**Remark 4.2** *The definition provided for the edge bubble function $b_E^A$ via the problem* (35) *is not suited when $E \in \mathcal{E}_{h,N}$. In such a case the boundary problem* (35) *has to be replaced by a new one similar to* (48).

We are now in a position to study the reliability and the efficiency of the error estimator (26).

# 5 Reliability of $\eta_T$

Let us first prove the following Lemma, which, basically, establishes an equivalence between the quantities $w_K(e_h)$ and $w_K(e_h^*)$ provided that the stretching factor $s_K$ is bounded from above by a quantity depending on $\nu_K$ and such that the smaller $\nu_K$, the larger $s_K$ can be. On the contrary, should this bound not hold, then the constants in the equivalence relations would depend on $s_K$.

**Lemma 5.1** *Under the assumption that*

$$\nu_K(2+\nu_K)\Big(1+\frac{s_K^2+1/s_K^2}{2}\Big) \le C_*,$$

*for some positive constant $C_* < 1/2$, where $\nu_K$ is defined via* (27)*, then it holds that*

$$C_1\, w_K^2(e_h^*) \le w_K^2(e_h) \le C_2\, w_K^2(e_h^*), \tag{49}$$

*where $C_1 = C_1(C_*) > 0, C_2 = C_2(C_*) > C_1$ and $\lim_{\nu_K\to 0} C_1 = \lim_{\nu_K\to 0} C_2 = 1$.*

**Proof.** Let us rewrite the expression for $w_K(e_h)$ and $w_K(e_h^*)$ given by (17) and (22), respectively. Let $\vec{r}_{1,K} = [c \;\; s]^T$ and $\vec{r}_{2,K} = [-s \;\; c]^T$, with $c = \cos\theta$ and $s = \sin\theta$, for some $\theta \in [0,\pi[$. We have $w_K^2(e_h) = \vec{r}_{1,K}^{\,T}\, M\, \vec{r}_{1,K}$, $w_K^2(e_h^*) = \vec{r}_{1,K}^{\,T}\, \widetilde{M}\, \vec{r}_{1,K}$ where $M = \{m_{ij}\}$ and $\widetilde{M} = \{\widetilde{m}_{ij}\} \in \mathbb{R}^{2\times 2}$ are the symmetric positive definite matrices given by

$$m_{ij} = \begin{cases} s_K \Big\|\dfrac{\partial e_h}{\partial x_1}\Big\|^2_{L^2(\Delta_K)} + \dfrac{1}{s_K}\Big\|\dfrac{\partial e_h}{\partial x_2}\Big\|^2_{L^2(\Delta_K)}, & i=j=1,\\[2ex] \Big(s_K - \dfrac{1}{s_K}\Big)\displaystyle\int_{\Delta_K} \dfrac{\partial e_h}{\partial x_1}\dfrac{\partial e_h}{\partial x_2}\, d\vec{x}, & i\ne j,\\[2ex] s_K \Big\|\dfrac{\partial e_h}{\partial x_2}\Big\|^2_{L^2(\Delta_K)} + \dfrac{1}{s_K}\Big\|\dfrac{\partial e_h}{\partial x_1}\Big\|^2_{L^2(\Delta_K)}, & i=j=2,\end{cases}$$

and

$$\widetilde{m}_{ij} = \begin{cases} s_K \left\| G_1 u_h - \dfrac{\partial u_h}{\partial x_1} \right\|^2_{L^2(\Delta_K)} + \dfrac{1}{s_K} \left\| G_2 u_h - \dfrac{\partial u_h}{\partial x_2} \right\|^2_{L^2(\Delta_K)}, & i = j = 1, \\ \Big(s_K - \dfrac{1}{s_K}\Big) \displaystyle\int_{\Delta_K} \Big(G_1 u_h - \dfrac{\partial u_h}{\partial x_1}\Big)\Big(G_2 u_h - \dfrac{\partial u_h}{\partial x_2}\Big)\, d\vec{x}, & i \neq j, \\ s_K \left\| G_2 u_h - \dfrac{\partial u_h}{\partial x_2} \right\|^2_{L^2(\Delta_K)} + \dfrac{1}{s_K} \left\| G_1 u_h - \dfrac{\partial u_h}{\partial x_1} \right\|^2_{L^2(\Delta_K)}, & i = j = 2, \end{cases}$$

respectively. To prove (49) it suffices to show under what conditions

$$\frac{w_K^2(e_h^*)}{w_K^2(e_h)} = \frac{\vec{r}_{1,K}^{\,T}\, \widetilde{M}\, \vec{r}_{1,K}}{\vec{r}_{1,K}^{\,T}\, M\, \vec{r}_{1,K}}$$

is bounded from below and from above, independently of $s_K$. This is equivalent to requiring that the eigenvalues $\mu$ of the generalized eigenvalue problem

$$\widetilde{M}\vec{x} = \mu M \vec{x},$$

or alternatively, that the eigenvalues of the symmetric part of the positive definite matrix $M^{-1}\widetilde{M}$ do not depend on $s_K$.

For ease of notation, throughout we let $e_{ij} = \displaystyle\int_{\Delta_K} \frac{\partial e_h}{\partial x_i}\frac{\partial e_h}{\partial x_j}\, d\vec{x}$ and $E_{ij} = \displaystyle\int_{\Delta_K} \big(G_i u_h - \frac{\partial u_h}{\partial x_i}\big)\big(G_j u_h - \frac{\partial u_h}{\partial x_j}\big)\, d\vec{x}$, for $i, j = 1, 2$, and for generality we let $s_K$ and $1/s_K$ be replaced by any two positive constants $\alpha, \beta$, respectively, with $\alpha \geq \beta$.

Let us rewrite the elements of $\widetilde{M}$ as $\widetilde{m}_{ij} = m_{ij} + \epsilon_{ij}$, for $i, j = 1, 2$, with

$$\epsilon_{ij} = \begin{cases} \quad \alpha\Big[\big\|G_1 u_h - \dfrac{\partial u}{\partial x_1}\big\|^2_{L^2(\Delta_K)} + 2\displaystyle\int_{\Delta_K} \dfrac{\partial e_h}{\partial x_1}\Big(G_1 u_h - \dfrac{\partial u}{\partial x_1}\Big)\, d\vec{x}\Big] \\ + \quad \beta\Big[\big\|G_2 u_h - \dfrac{\partial u}{\partial x_2}\big\|^2_{L^2(\Delta_K)} + 2\displaystyle\int_{\Delta_K} \dfrac{\partial e_h}{\partial x_2}\Big(G_2 u_h - \dfrac{\partial u}{\partial x_2}\Big)\, d\vec{x}\Big], \quad i = j = 1, \\ \quad (\alpha - \beta)\Big[\displaystyle\int_{\Delta_K} \dfrac{\partial e_h}{\partial x_1}\Big(G_2 u_h - \dfrac{\partial u}{\partial x_2}\Big)\, d\vec{x} + \displaystyle\int_{\Delta_K} \dfrac{\partial e_h}{\partial x_2}\Big(G_1 u_h - \dfrac{\partial u}{\partial x_1}\Big)\, d\vec{x} \\ + \quad \displaystyle\int_{\Delta_K} \Big(G_1 u_h - \dfrac{\partial u}{\partial x_1}\Big)\Big(G_2 u_h - \dfrac{\partial u}{\partial x_2}\Big)\, d\vec{x}\Big], \quad i \neq j, \\ \quad \alpha\Big[\big\|G_2 u_h - \dfrac{\partial u}{\partial x_2}\big\|^2_{L^2(\Delta_K)} + 2\displaystyle\int_{\Delta_K} \dfrac{\partial e_h}{\partial x_2}\Big(G_2 u_h - \dfrac{\partial u}{\partial x_2}\Big)\, d\vec{x}\Big] \\ + \quad \beta\Big[\big\|G_1 u_h - \dfrac{\partial u}{\partial x_1}\big\|^2_{L^2(\Delta_K)} + 2\displaystyle\int_{\Delta_K} \dfrac{\partial e_h}{\partial x_1}\Big(G_1 u_h - \dfrac{\partial u}{\partial x_1}\Big)\, d\vec{x}\Big], \quad i = j = 2. \end{cases}$$

The "perturbations" $\epsilon_{ij}$ can be bounded using (27) and the Cauchy-Schwarz inequality as

$$\begin{array}{lcll} -2\nu_K\Big(\alpha e_{11} + \beta e_{22}\Big) & \leq & \epsilon_{11} & \leq \nu_K(\nu_K + 2)\Big(\alpha e_{11} + \beta e_{22}\Big), \\ -\nu_K(\nu_K + 2)(\alpha - \beta) e_{11}^{1/2} e_{22}^{1/2} & \leq & \epsilon_{12} & \leq \nu_K(\nu_K + 2)(\alpha - \beta) e_{11}^{1/2} e_{22}^{1/2}, \\ -2\nu_K\Big(\alpha e_{22} + \beta e_{11}\Big) & \leq & \epsilon_{22} & \leq \nu_K(\nu_K + 2)\Big(\alpha e_{22} + \beta e_{11}\Big). \end{array} \tag{50}$$

A simple computation shows that

$$M^{-1}\widetilde{M} = I + \frac{1}{\det(M)} \begin{bmatrix} m_{22}\epsilon_{11} - m_{12}\epsilon_{12} & m_{22}\epsilon_{12} - m_{12}\epsilon_{22} \\ m_{11}\epsilon_{12} - m_{12}\epsilon_{11} & m_{11}\epsilon_{22} - m_{12}\epsilon_{12} \end{bmatrix},$$

where $I$ is the identity matrix and $\det(M)$ is the determinant of $M$. The symmetric part of $M^{-1}\widetilde{M}$ is thus given by $(M^{-1}\widetilde{M})_{\rm sym} = I + \dfrac{1}{\det(M)}\,\mathcal{C}$, with

$$\mathcal{C} = \{c_{ij}\} = \begin{bmatrix} m_{22}\epsilon_{11} - m_{12}\epsilon_{12} & \dfrac{(m_{11}+m_{22})\epsilon_{12} - m_{12}(\epsilon_{11}+\epsilon_{22})}{2} \\ \dfrac{(m_{11}+m_{22})\epsilon_{12} - m_{12}(\epsilon_{11}+\epsilon_{22})}{2} & m_{11}\epsilon_{22} - m_{12}\epsilon_{12} \end{bmatrix}.$$

We next bound the entries of $\mathcal{C}$ by exploiting inequalities (50). Tedious but straightforward computations yield

$$\begin{aligned} -\nu_K^2(\alpha-\beta)^2 e_{11}e_{22} - 2\nu_K\Big(2(\alpha^2+\beta^2)e_{11}e_{22} + \alpha\beta(e_{11}-e_{22})^2\Big) \le c_{11}, c_{22} \\ \le \nu_k(\nu_K+2)\Big(2(\alpha^2+\beta^2)e_{11}e_{22} + \alpha\beta(e_{11}-e_{22})^2\Big), \\ |c_{12}| \le \nu_K(\nu_K+2)(\alpha^2-\beta^2)e_{11}^{1/2}e_{22}^{1/2}(e_{11}+e_{22}). \end{aligned} \tag{51}$$

Moreover, using the Cauchy-Schwarz inequality

$$\begin{aligned} \det(M) &= m_{11}m_{22} - m_{12}^2 = (\alpha e_{11} + \beta e_{22})(\alpha e_{22} + \beta e_{11}) - (\alpha-\beta)^2 e_{12}^2 \\ &\ge (\alpha e_{11} + \beta e_{22})(\alpha e_{22} + \beta e_{11}) - (\alpha-\beta)^2 e_{11}e_{22} = \alpha\beta(e_{11}+e_{22})^2. \end{aligned}$$

It can be checked that the largest bound of the absolute value for $c_{11}, c_{22}$ in (51) is the upper bound, so that we obtain

$$\begin{aligned} \frac{|c_{11}|}{m_{11}m_{22} - m_{12}^2} &\le \nu_K(\nu_K+2)\frac{\big(2(\alpha^2+\beta^2)e_{11}e_{22} + \alpha\beta(e_{11}-e_{22})^2\big)}{\alpha\beta(e_{11}+e_{22})^2} \\ &= \nu_K(\nu_K+2)\Bigg(\underbrace{\left(\frac{e_{11}-e_{22}}{e_{11}+e_{22}}\right)^2}_{a_1} + 2\frac{\alpha^2+\beta^2}{\alpha\beta}\underbrace{\left(\frac{e_{11}^{1/2}e_{22}^{1/2}}{e_{11}+e_{22}}\right)^2}_{a_2}\Bigg) \\ &\le \nu_K(\nu_K+2)\Big(1 + \frac{\alpha^2+\beta^2}{2\alpha\beta}\Big), \end{aligned} \tag{52}$$

as $a_1 \le 1$ and $a_2 \le 1/4$. Analogously, we have

$$\begin{aligned} \frac{|c_{12}|}{m_{11}m_{22} - m_{12}^2} &\le \nu_K(\nu_K+2)\,\frac{\alpha^2-\beta^2}{\alpha\beta}\,\frac{e_{11}^{1/2}e_{22}^{1/2}(e_{11}+e_{22})}{(e_{11}+e_{22})^2} \\ &\le \nu_K(\nu_K+2)\,\frac{\alpha^2-\beta^2}{\alpha\beta}\underbrace{\left(\frac{e_{11}^{1/2}e_{22}^{1/2}}{e_{11}+e_{22}}\right)}_{\sqrt{a_2}} \le \nu_K(\nu_K+2)\,\frac{\alpha^2-\beta^2}{2\alpha\beta}, \end{aligned} \tag{53}$$

as $\sqrt{a_2} \le 1/2$. The thesis follows using Gershgorin theorem to bound the eigenvalues of $(M^{-1}\widetilde{M})_{\rm sym}$, on noting that the term $1 + (\alpha^2+\beta^2)/(2\alpha\beta)$ in (52) is always greater than $(\alpha^2-\beta^2)/(2\alpha\beta)$ in (53), and recalling that, for the case at hand, $\alpha = s_K$, and $\beta = 1/s_K$. Thus, both the radius and the center of the two circles containing the eigenvalues of $(M^{-1}\widetilde{M})_{\rm sym} - I$ are bounded by the same quantity, i.e. the right-hand side of (52), so that the constraint $C_* < 1/2$ guarantees that the lower bound for the estimate of the eigenvalues of the matrix is positive. We note that, while the true eigenvalues of the matrix $M^{-1}\widetilde{M}$ are positive, this may not be the case for their estimates, unless the requirement $C_* < 1/2$ is made. Moreover, as $\nu_K \to 0$, both $C_1$ and $C_2$ tend to one, since $M^{-1}\widetilde{M} \to I$, due to (50). □

The reliability of the error estimator $\eta_T$ in (26) is stated by the following

**Proposition 5.1** *The error estimator (26) is reliable, i.e.,*

$$
\begin{aligned}
|||e_h||| \quad &\leq \quad \overline{C}\Big\{ \sum_{K\in\mathcal{T}_h} \eta_{K,T}^2 + \sum_{K\in\mathcal{T}_h} \lambda_{1,K}^2 \|f - f_K\|^2_{L^2(K)} \\
&+ \quad \sum_{K\in\mathcal{T}_h} \frac{\lambda_{1,K}^2}{\lambda_{2,K}} \sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,N}} \|g - g_E\|^2_{L^2(E)} \Big\}^{1/2}
\end{aligned}
\tag{54}
$$

*where* $\overline{C} = \overline{C}(N, C_\Delta, C_2, \gamma_{min})$.

**Proof.** First, let us prove the intermediate result

$$|||e_h||| \leq C_R\,\eta, \tag{55}$$

where $\eta$ is the global error indicator (21). This follows immediately from Proposition 3.1 and Lemma 5.1, with $C_R = C_R(N, C_\Delta, C_2)$.

From (55) and recalling the definitions of $\eta_K$ and $\eta_{K,T}$, we obtain

$$
\begin{aligned}
&|||e_h|||^2 \leq C_R \sum_{K\in\mathcal{T}_h} \Big\{ \Big[ \lambda_{1,K}^{1/2}\lambda_{2,K}^{1/2}\, \|f - f_K\|_{L^2(K)} + \lambda_{1,K}^{1/2}\lambda_{2,K}^{1/2}\, \|f_K\|_{L^2(K)} \\
&+ \frac{\lambda_{1,K}^{1/2}}{2} \sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,\Omega}} \Big\| \Big[\frac{\partial u_h}{\partial \vec{n}_{L,K}}\Big]_E \Big\|_{L^2(E)} + \lambda_{1,K}^{1/2} \sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,N}} \|g - g_E\|_{L^2(E)} \\
&+ \lambda_{1,K}^{1/2} \sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,N}} \Big\| g_E - \frac{\partial u_h}{\partial \vec{n}_{L,K}} \Big\|_{L^2(E)} \Big] w_K(e_h^*) \Big\} = C_R \sum_{K\in\mathcal{T}_h} \eta_{K,T}^2 \\
&+ C_R \sum_{K\in\mathcal{T}_h} \Big\{ \Big[ \lambda_{1,K}^{1/2}\lambda_{2,K}^{1/2}\, \|f - f_K\|_{L^2(K)} + \lambda_{1,K}^{1/2} \sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,N}} \|g - g_E\|_{L^2(E)} \Big] w_K(e_h^*) \Big\}.
\end{aligned}
\tag{56}
$$

Our goal is to bound the right-hand side of (56) in terms of $\eta_{K,T}$ and of the data perturbations $(f - f_K)$ and $(g - g_E)$ only. Let us exploit Lemma 4.5 and Young inequality to get

$$
\begin{aligned}
|||e_h|||^2 \quad &\leq \quad C_R \sum_{K\in\mathcal{T}_h} \eta_{K,T}^2 + C_R \sum_{K\in\mathcal{T}_h} \Big\{ \Big[ \lambda_{1,K}\, \|f - f_K\|_{L^2(K)} \\
&+ \quad \frac{\lambda_{1,K}}{\lambda_{2,K}^{1/2}} \sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,N}} \|g - g_E\|_{L^2(E)} \Big] \frac{(1+\nu_K)}{\gamma_{min}^{1/2}}\, |||e_h|||_{\Delta_K} \Big\} \\
&\leq \quad C_R \sum_{K\in\mathcal{T}_h} \eta_{K,T}^2 + C_R \sum_{K\in\mathcal{T}_h} \Big\{ \frac{2\,\varepsilon\,(1+\nu_K)^2}{\gamma_{min}}\, |||e_h|||^2_{\Delta_K} + \frac{1}{4\varepsilon} \Big[ \lambda_{1,K}^2\, \|f - f_K\|^2_{L^2(K)} \\
&+ \quad \frac{\lambda_{1,K}^2}{\lambda_{2,K}} \Big( \sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,N}} \|g - g_E\|_{L^2(E)} \Big)^2 \Big] \Big\} \\
&\leq \quad C_R \Big\{ \sum_{K\in\mathcal{T}_h} \eta_{K,T}^2 + \frac{2\,\varepsilon\,(1+\nu)^2}{\gamma_{min}} \sum_{K\in\mathcal{T}_h} |||e_h|||^2_{\Delta_K} + \frac{1}{4\,\varepsilon} \sum_{K\in\mathcal{T}_h} \lambda_{1,K}^2 \|f - f_K\|^2_{L^2(K)} \\
&+ \quad \frac{1}{2\varepsilon} \sum_{K\in\mathcal{T}_h} \frac{\lambda_{1,K}^2}{\lambda_{2,K}} \sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,N}} \|g - g_E\|^2_{L^2(E)} \Big\},
\end{aligned}
$$

where we have used the fact that, for any $K \in \mathcal{T}_h$, the number of Neumann edges is at most equal to 2, while $\nu = \max_{K\in\mathcal{T}_h} \nu_K$, $\nu_K$ being defined via (27), and $\varepsilon$ is the parameter associated with the

Young inequality, to be suitably chosen. As from (6) it follows that $\sum_{K\in\mathcal{T}_h} |||e_h|||^2_{\Delta_K} \leq N\, |||e_h|||^2$, we get

$$\begin{aligned}
&\Big(1-\frac{2C_R\varepsilon N(1+\nu)^2}{\gamma_{min}}\Big)\, |||e_h|||^2 \leq C_R \Big\{ \sum_{K\in\mathcal{T}_h} \eta^2_{K,T} \\
&+ \frac{1}{4\varepsilon}\sum_{K\in\mathcal{T}_h} \lambda^2_{1,K} \|f-f_K\|^2_{L^2(K)} + \frac{1}{2\,\varepsilon}\sum_{K\in\mathcal{T}_h} \frac{\lambda^2_{1,K}}{\lambda_{2,K}} \sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,N}} \|g-g_E\|^2_{L^2(E)} \Big\}.
\end{aligned} \tag{57}$$

To make inequality (57) meaningful, we have to assure that

$$1-\frac{2C_R\varepsilon N(1+\nu)^2}{\gamma_{min}} > 0 \qquad \text{i.e.} \qquad \varepsilon < \frac{\gamma_{min}}{2C_R N(1+\nu)^2}.$$

For instance, the choice $\varepsilon = \gamma_{min}/(4C_R N(1+\nu)^2)$ guarantees that

$$\begin{aligned}
|||e_h||| \quad &\leq \quad C_R \Big\{ \sum_{K\in\mathcal{T}_h} \eta^2_{K,T} + \frac{C_R N(1+\nu)^2}{\gamma_{min}} \sum_{K\in\mathcal{T}_h} \lambda^2_{1,K} \|f-f_K\|^2_{L^2(K)} \\
&+ \quad \frac{2C_R N(1+\nu)^2}{\gamma_{min}} \sum_{K\in\mathcal{T}_h} \frac{\lambda^2_{1,K}}{\lambda_{2,K}} \sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,N}} \|g-g_E\|^2_{L^2(E)} \Big\}^{1/2},
\end{aligned}$$

namely the reliability result (54), where $\overline{C} = C_R \max\Big(1, \dfrac{8N}{\gamma_{min}} C_R\Big)$, with $C_R$ defined through (55). □

**Remark 5.1** *Notice that the contribution of the data oscillation increases as $N$ gets larger, i.e., when the maximum number of elements of a patch increases. Moreover, the reliability result does not depend explicitly on $\nu$ since this quantity appears in the form $1+\nu$ which is uniformly bounded between $1$ and $2$.*

# 6 Efficiency of $\eta_T$

To prove the efficiency of the error estimator defined in (26), let us study separately the three terms at the right-hand side of (25):

$$\begin{aligned}
\eta^2_{K,T} \quad = \quad & \underbrace{\lambda^{1/2}_{1,K}\, \lambda^{1/2}_{2,K} \|f_K\|_{L^2(K)}\, w_K(e^*_h)}_{(\,\mathrm{I}\,)} \\
+ \quad & \underbrace{\frac{\lambda^{1/2}_{1,K}}{2} \Big( \sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,\Omega}} \Big\| \Big[ \frac{\partial u_h}{\partial \vec{n}_{L,K}} \Big]_E \Big\|_{L^2(E)} \Big)\, w_K(e^*_h)}_{(\,\mathrm{II}\,)} \\
+ \quad & \underbrace{\lambda^{1/2}_{1,K} \Big( \sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,N}} \Big\| \Big( g_E - \frac{\partial u_h}{\partial \vec{n}_{L,K}} \Big)\Big|_E \Big\|_{L^2(E)} \Big)\, w_K(e^*_h)}_{(\,\mathrm{III}\,)}.
\end{aligned} \tag{58}$$

In the Lemmas below we bound the three quantities ( I ), ( II ) and ( III ) in terms of the data perturbations $(f-f_K)$ and $(g-g_E)$, and of the energy norm $|||e_h|||_{\Delta_K}$ of the discretization error on the patch $\Delta_K$.

First, let us provide two results used in the sequel. The first one is relation (18) rewritten on the whole domain $\Omega$ and by recalling the simplifying assumptions made in Section 4 on the initial problem (9). We have

$$\int_\Omega \Big( \sum_{i,j=1}^{2} a_{ij} \, \frac{\partial e_h}{\partial x_j} \, \frac{\partial v}{\partial x_i} \Big) \, d\vec{x} = \int_\Omega f v \, d\vec{x} + \int_{\Gamma_N} g v \, ds - \int_\Omega \Big( \sum_{i,j=1}^{2} a_{ij} \, \frac{\partial u_h}{\partial x_j} \, \frac{\partial v}{\partial x_i} \Big) \, d\vec{x}. \tag{59}$$

The second result is obtained by suitably integrating by parts the right-hand side of (59):

$$\begin{aligned} & \int_\Omega f v \, d\vec{x} + \int_{\Gamma_N} g v \, ds - \int_\Omega \Big( \sum_{i,j=1}^{2} a_{ij} \, \frac{\partial u_h}{\partial x_j} \, \frac{\partial v}{\partial x_i} \Big) \, d\vec{x} \\ = \; & \int_\Omega f v \, d\vec{x} + \int_{\Gamma_N} g v \, ds + \sum_{K \in \mathcal{T}_h} \Big\{ \int_K \Big( \sum_{i,j=1}^{2} a_{ij} \frac{\partial^2 u_h}{\partial x_i \, \partial x_j} \; v \Big) \, d\vec{x} - \int_{\partial K} \frac{\partial u_h}{\partial \vec{n}_{L,K}} \, v \, ds \Big\} \\ = \; & \sum_{K \in \mathcal{T}_h} \int_K f v \, d\vec{x} + \sum_{E \in \mathcal{E}_{h,N}} \int_E \Big( g - \frac{\partial u_h}{\partial \vec{n}_{L,K}} \Big) v \, ds - \sum_{E \in \mathcal{E}_{h,\Omega}} \int_E \Big[ \frac{\partial u_h}{\partial \vec{n}_{L,K}} \Big]_E v \, ds, \end{aligned} \tag{60}$$

the diffusive matrix $A$ having been assumed constant.
Let us begin to analyze the term $(\,\mathrm{I}\,)$ in (58).

**Lemma 6.1** *Under the Assumption* 4.1*, the bound*

$$(\,\mathrm{I}\,) \leq \frac{1+\nu_K}{C_{\widehat{K}}^{1/2} \, \gamma_{min}} \Big[ \Big( \frac{C_A \, \widehat{\lambda}^{1/2} \, (1+s_K^2)^{1/2}}{2^{1/2}} + \frac{\gamma_{min}}{4} \Big) \, |||e_h|||_{\Delta_K}^2 + \lambda_{1,K}^2 \, \| f - f_K \|_{L^2(K)}^2 \Big] \tag{61}$$

*holds, where $\gamma_{min}$ is the minimum eigenvalue of the constant diffusive matrix $A$, $\widehat{\lambda}$ is the eigenvalue of the problem* (34)*, $C_A = 4 \max_{i,j=1,\,2} |a_{ij}|$, and $\nu_K$ and $C_{\widehat{K}}$ are defined by relations* (27) *and* (38)*, respectively.*

**Proof.** First, let us introduce the auxiliary function $\sigma_K$ defined by

$$\sigma_K = \mathrm{sign}(f_K) \, b_K^A \, |K|^{-1/2},$$

sign denoting the sign-function and where $b_K^A$ is the triangle anisotropic bubble function introduced in Section 4.2. Lemma 4.6 immediately yields

$$\int_K f_K \, \sigma_K \, d\vec{x} = |f_K| \, |K|^{-1/2} \int_K b_K^A \, d\vec{x} = C_{\widehat{K}} \, \| f_K \|_{L^2(K)},$$

that is

$$\| f_K \|_{L^2(K)} = \frac{1}{C_{\widehat{K}}} \int_K f_K \, \sigma_K \, d\vec{x}. \tag{62}$$

Now, since $\mathrm{supp}(\sigma_K) \subset K$, and from equations (60), (59) and Lemma 4.6 we get

$$
\begin{aligned}
&\int_K f_K\, \sigma_K\, d\vec{x} = \int_K f\, \sigma_K\, d\vec{x} + \int_K (f_K - f)\, \sigma_K\, d\vec{x} \\
= \quad & \int_\Omega f\, \sigma_K\, d\vec{x} + \int_{\Gamma_N} g\, \sigma_K\, ds - \int_\Omega \Big( \sum_{i,j=1}^{2} a_{ij} \frac{\partial u_h}{\partial x_j} \frac{\partial \sigma_K}{\partial x_i} \Big)\, d\vec{x} + \int_K (f_K - f)\, \sigma_K\, d\vec{x} \\
= \quad & \int_K \Big( \sum_{i,j=1}^{2} a_{ij} \frac{\partial e_h}{\partial x_j} \frac{\partial \sigma_K}{\partial x_i} \Big)\, d\vec{x} + \int_K (f_K - f)\, \sigma_K\, d\vec{x} \\
\le \quad & C_A\, \|\nabla e_h\|_{L^2(K)} \|\nabla \sigma_K\|_{L^2(K)} + \|f - f_K\|_{L^2(K)} \|\sigma_K\|_{L^2(K)} \\
\le \quad & |K|^{-1/2} \Big( \int_K b_K^A\, d\vec{x} \Big)^{1/2} \Big( C_A\, \frac{\widehat{\lambda}^{1/2}\, (1 + s_K^2)^{1/2}}{2^{1/2}\, \lambda_{1,K}} \|\nabla e_h\|_{L^2(K)} + \|f - f_K\|_{L^2(K)} \Big) \\
= \quad & C_{\widehat{K}}^{1/2} \Big[ C_A\, \frac{\widehat{\lambda}^{1/2}\, (1 + s_K^2)^{1/2}}{2^{1/2}\, \lambda_{1,K}}\, \|\nabla e_h\|_{L^2(K)} + \|f - f_K\|_{L^2(K)} \Big].
\end{aligned}
\tag{63}
$$

Going back to the term $(\,\mathrm{I}\,)$ in (58) we have

$$
\begin{aligned}
&(\,\mathrm{I}\,) \le \lambda_{1,K}^{1/2} \lambda_{2,K}^{1/2}\, \frac{1}{C_{\widehat{K}}^{1/2}} \Big[ C_A\, \frac{\widehat{\lambda}^{1/2}\, (1 + s_K^2)^{1/2}}{2^{1/2}\, \lambda_{1,K}}\, \|\nabla e_h\|_{L^2(K)} + \|f - f_K\|_{L^2(K)} \Big]\, w_K(e_h^*) \\
&\le \lambda_{1,K}\, \frac{1}{C_{\widehat{K}}^{1/2}} \Big[ \frac{C_A}{\gamma_{min}^{1/2}}\, \frac{\widehat{\lambda}^{1/2}\, (1 + s_K^2)^{1/2}}{2^{1/2}\, \lambda_{1,K}}\, |||e_h|||_K + \|f - f_K\|_{L^2(K)} \Big]\, \frac{1 + \nu_K}{\gamma_{min}^{1/2}}\, |||e_h|||_{\Delta_K} \\
&\le \frac{1 + \nu_K}{(C_{\widehat{K}}\, \gamma_{min})^{1/2}} \left( \frac{C_A}{\gamma_{min}^{1/2}}\, \frac{\widehat{\lambda}^{1/2}\, (1 + s_K^2)^{1/2}}{2^{1/2}}\, |||e_h|||_{\Delta_K}^2 + \lambda_{1,K}\, \|f - f_K\|_{L^2(K)} |||e_h|||_{\Delta_K} \right),
\end{aligned}
\tag{64}
$$

where relations (62), (63), (31) and (33) have been exploited. Let us further rewrite the product $\lambda_{1,K}\, \|f - f_K\|_{L^2(K)}\, |||e_h|||_{\Delta_K}$ via the Young inequality:

$$
\lambda_{1,K}\, \|f - f_K\|_{L^2(K)}\, |||e_h|||_{\Delta_K} \le \frac{\lambda_{1,K}^2}{\gamma_{min}^{1/2}}\, \|f - f_K\|_{L^2(K)}^2 + \frac{\gamma_{min}^{1/2}}{4}\, |||e_h|||_{\Delta_K}^2,
$$

which, inserted into (64), provides the estimate (61). $\square$

The term $(\,\mathrm{II}\,)$ in (58) can be bounded thanks to a similar procedure.

**Lemma 6.2** *Under the Assumption* 4.1 *it can be proved that*

$$
\begin{aligned}
(\,\mathrm{II}\,) \le \frac{3\, \overline{C}^*(1 + \nu_K)}{2\, C_{\widehat{K}}^*\, \gamma_{min}} \Big[ & s_K \left( \frac{\gamma_{min}}{4} h_{\widehat{K}}^{1/2} + \frac{C_A}{\rho_{\widehat{K}}^{1/2}} \right) |||e_h|||_{\Delta_K}^2 \\
& + \lambda_{1,K}^2\, N\, h_{\widehat{K}}^{1/2} \sum_{T \in \Delta_K} \|f - f_T\|_{L^2(T)}^2 \Big],
\end{aligned}
\tag{65}
$$

*where*

$$
\overline{C}^* = \widehat{C}^{1/2}\, \max(s_K^{1/2}, s_{K'}^{1/2})\, \max\Big\{ 1 + \frac{1}{C_{\widehat{K}}^{1/2}},\, \max(s_K, s_{K'})\, h_{\widehat{K}} \left( \widetilde{\lambda}^{1/2} + \Big( \frac{2\, \widehat{\lambda}}{C_{\widehat{K}}} \Big)^{1/2} \right) \Big\},
$$

$K'$ *is the triangle sharing edge* $E$ *with* $K$, $\nu_K$, $C_A$, $\gamma_{min}$, $C_{\widehat{K}}$ *and* $\widehat{\lambda}$ *are defined as in Lemma* 6.1, $\widetilde{\lambda}$ *is the eigenvalue of the problem* (35), *the constants* $N$, $C_{\widehat{K}}^*$ *and* $\widehat{C}$ *are* *defined by relations* (6), (39) *and* (40), *and* $h_{\widehat{K}}, \rho_{\widehat{K}}$ *are the diameter of* $\widehat{K}$ *and of the ball inscribed in it, respectively.*

**Proof.** Let us introduce the auxiliary function $\sigma_E$ associated with a generic internal edge $E \in \mathcal{E}(K) \cap \mathcal{E}_{h,\Omega}$,

$$\sigma_E = \mathrm{sign}\Big( \Big[ \frac{\partial u_h}{\partial \vec{n}_{L,K}} \Big]_E \Big)\, h_E^{-1/2}\, b_E^A ,$$

$b_E^A$ denoting the edge anisotropic bubble function defined in Section 4.2. As, thanks to Lemma 4.6,

$$\int_E \Big[ \frac{\partial u_h}{\partial \vec{n}_{L,K}} \Big]_E \sigma_E \, ds = \Big| \Big[ \frac{\partial u_h}{\partial \vec{n}_{L,K}} \Big]_E \Big|\, h_E^{-1/2} \int_E b_E^A \, ds = C^*_{\widehat{K}} \, \Big\| \Big[ \frac{\partial u_h}{\partial \vec{n}_{L,K}} \Big]_E \Big\|_{L^2(E)},$$

we immediately deduce that

$$\Big\| \Big[ \frac{\partial u_h}{\partial \vec{n}_{L,K}} \Big]_E \Big\|_{L^2(E)} = \frac{1}{C^*_{\widehat{K}}} \int_E \Big[ \frac{\partial u_h}{\partial \vec{n}_{L,K}} \Big]_E \sigma_E \, ds. \tag{66}$$

As $\mathrm{supp}(\sigma_E) \subset Q_E$, and thanks to results (60), (59) and Lemma 4.6 we infer

$$\begin{aligned}
& \int_E \Big[ \frac{\partial u_h}{\partial \vec{n}_{L,K}} \Big]_E \sigma_E \, ds \\
= \;& \sum_{T \in Q_E} \int_T f \, \sigma_E \, d\vec{x} - \int_\Omega f \, \sigma_E \, d\vec{x} - \int_{\Gamma_N} g \, \sigma_E \, ds + \int_\Omega \Big( \sum_{i,j=1}^{2} a_{ij} \frac{\partial u_h}{\partial x_j} \frac{\partial \sigma_E}{\partial x_i} \Big) d\vec{x} \\
= \;& \sum_{T \in Q_E} \int_T f \, \sigma_E \, d\vec{x} - \int_{Q_E} \Big( \sum_{i,j=1}^{2} a_{ij} \frac{\partial e_h}{\partial x_j} \frac{\partial \sigma_E}{\partial x_i} \Big) d\vec{x} \\
\le \;& \|f\|_{L^2(Q_E)} \|\sigma_E\|_{L^2(Q_E)} + C_A \|\nabla e_h\|_{L^2(Q_E)} \|\nabla \sigma_E\|_{L^2(Q_E)} \\
\le \;& h_E^{-1/2} \Big( \int_{Q_E} b_E^A \, d\vec{x} \Big)^{1/2} \Big( \|f\|_{L^2(Q_E)} \\
& \qquad\qquad + C_A \, h_{\widehat{K}} \, \widetilde{\lambda}^{1/2} \, h_E^{-1} \, \max(s_K, s_{K'}) \, \|\nabla e_h\|_{L^2(Q_E)} \Big) \\
\le \;& \widehat{C}^{1/2} \max(s_K^{1/2}, s_{K'}^{1/2}) \Big( h_E^{1/2} \|f\|_{L^2(Q_E)} \\
& \qquad\qquad + C_A \, h_{\widehat{K}} \, h_E^{-1/2} \, \widetilde{\lambda}^{1/2} \max(s_K, s_{K'}) \|\nabla e_h\|_{L^2(Q_E)} \Big) \\
\le \;& \widehat{C}^{1/2} \max(s_K^{1/2}, s_{K'}^{1/2}) \Big( h_E^{1/2} \sum_{T \in Q_E} \|f - f_T\|_{L^2(T)} \\
& \quad + h_E^{1/2} \sum_{T \in Q_E} \|f_T\|_{L^2(T)} + C_A \, h_{\widehat{K}} \, h_E^{-1/2} \, \widetilde{\lambda}^{1/2} \max(s_K, s_{K'}) \|\nabla e_h\|_{L^2(Q_E)} \Big),
\end{aligned}$$

where relation (19) has been exploited. Now, by suitably using equalities (62) and (63) to estimate the norms $\|f_T\|_{L^2(T)}$, we get

$$\int_E \Big[ \frac{\partial u_h}{\partial \vec{n}_{L,K}} \Big]_E \sigma_E \, ds \le \overline{C}^* \Big[ h_E^{1/2} \sum_{T \in Q_E} \|f - f_T\|_{L^2(T)} + \frac{C_A}{(\rho_{\widehat{K}} \lambda_{2,K})^{1/2}} \|\nabla e_h\|_{L^2(Q_E)} \Big],$$

where relation (19), (47), together with

$$\sum_{T \in Q_E} \|\nabla e_h\|_{L^2(T)} \le \sqrt{2}\, \|\nabla e_h\|_{L^2(Q_E)},$$

and

$$\frac{h_E^{1/2}\,(1+s_T^2)^{1/2}}{2^{1/2}\,\lambda_{1,T}} \le h_E^{1/2}\,s_T\,\frac{h_{\widehat{K}}}{h_E} = h_E^{-1/2}\,h_{\widehat{K}}\,s_T\,,$$

have been used. Thanks to the relations (31), (33) and (66), we get

$$\begin{aligned}
(\,\mathrm{II}\,) \quad &\le \quad \frac{\lambda_{1,K}^{1/2}\,\overline{C}^*}{2\,C^*_{\widehat{K}}}\Big(\sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,\Omega}}\Big[h_E^{1/2}\sum_{T\in Q_E}\|f-f_T\|_{L^2(T)} \\
&\qquad\qquad + \frac{C_A}{(\gamma_{min}\,\rho_{\widehat{K}}\,\lambda_{2,K})^{1/2}}\,|||e_h|||_{Q_E}\Big]\Big)w_K(e_h^*) \\
&\le \quad \frac{3\,\overline{C}^*}{2\,C^*_{\widehat{K}}}\Big[\lambda_{1,K}\,h_{\widehat{K}}^{1/2}\sum_{T\in\Delta_K}\|f-f_T\|_{L^2(T)} + C_A\Big(\frac{s_K}{\rho_{\widehat{K}}\,\gamma_{min}}\Big)^{1/2}|||e_h|||_{\Delta_K}\Big] \\
&\qquad \frac{s_K^{1/2}\,(1+\nu_K)}{\gamma_{min}^{1/2}}\,|||e_h|||_{\Delta_K} \\
&= \quad \frac{3\,\overline{C}^*\,(1+\nu_K)}{2\,C^*_{\widehat{K}}\,\gamma_{min}^{1/2}}\Big[\frac{\lambda_{1,K}^{3/2}}{\lambda_{2,K}^{1/2}}\,h_{\widehat{K}}^{1/2}\,|||e_h|||_{\Delta_K}\sum_{T\in\Delta_K}\|f-f_T\|_{L^2(T)} \\
&\qquad\qquad + \frac{s_K}{(\rho_{\widehat{K}}\,\gamma_{min})^{1/2}}\,C_A\,|||e_h|||^2_{\Delta_K}\Big]\,,
\end{aligned}$$

where relation

$$\bigcup_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,\Omega}} Q_E \subseteq \Delta_K,$$

along with (19), and the fact that $\mathcal{E}(K)$ consists of three edges have been used. Finally, we exploit the Young inequality to split the term
$|||e_h|||_{\Delta_K}\sum_{T\in\Delta_K}\|f-f_T\|_{L^2(T)}$:

$$\begin{aligned}
\frac{\lambda_{1,K}^{3/2}}{\lambda_{2,K}^{1/2}}\,|||e_h|||_{\Delta_K}\sum_{T\in\Delta_K}\|f-f_T\|_{L^2(T)} &\le \frac{\gamma_{min}^{1/2}}{4}\,s_K\,|||e_h|||^2_{\Delta_K} \\
&\quad + \frac{\lambda_{1,K}^2}{\gamma_{min}^{1/2}}\,N\sum_{T\in\Delta_K}\|f-f_T\|^2_{L^2(T)}.
\end{aligned}$$

This yields result (65). □

Finally, let us consider the term ( III ) in (58).

**Lemma 6.3** *Under the Assumption* 4.1 *we have*

$$\begin{aligned}
(\,\mathrm{III}\,) \quad \le \quad &\frac{2\,\overline{C}^{**}}{C^*_{\widehat{K}}}\,\frac{(1+\nu_K)}{\gamma_{min}}\Big[s_K\Big(\frac{C_A}{\rho_{\widehat{K}}^{1/2}} + \frac{\gamma_{min}}{4}\,h_{\widehat{K}}^{1/2} + \frac{\gamma_{min}}{4}\Big)|||e_h|||^2_{\Delta_K} \\
&+ h_{\widehat{K}}^{1/2}\lambda_{1,K}^2\,\|f-f_K\|^2_{L^2(K)} + 2\,\lambda_{1,K}\sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,N}}\|g_E-g\|^2_{L^2(E)}\Big],
\end{aligned} \tag{67}$$

*where*

$$\begin{aligned}
\overline{C}^{**} = \max\Big\{&\widehat{C}^{1/2}\,s_K^{1/2}\Big(1+\frac{1}{C_{\widehat{K}}^{1/2}}\Big),\,(C^*_{\widehat{K}})^{1/2}\,,\widehat{C}^{1/2}\,s_K^{3/2}h_{\widehat{K}}\,\widetilde{\lambda}^{1/2} \\
&+\Big(\frac{\widehat{C}}{C_{\widehat{K}}}\Big)^{1/2}\,\widehat{\lambda}^{1/2}\,h_{\widehat{K}}s_K^{3/2}\Big\},
\end{aligned}$$

$C_A$, $\gamma_{min}$, $\widehat{\lambda}$, $\widetilde{\lambda}$, $\widehat{C}$, $C_{\widehat{K}}$, $h_{\widehat{K}}$, $\rho_{\widehat{K}}$, $\nu_K$ *and* $C^*_{\widehat{K}}$ *being defined as in Lemma* 6.2.

**Proof.** Let us introduce the auxiliary function

$$\sigma_E = \mathrm{sign}\Big(\Big(g_E - \frac{\partial u_h}{\partial \vec{n}_{L,K}}\Big)\Big|_E\Big)\, h_E^{-1/2}\, b_E^A$$

associated with a generic edge $E \in \mathcal{E}(K) \cap \mathcal{E}_{h,N}$. From Lemma 4.6, we have

$$\begin{aligned}\int_E \Big(g_E - \frac{\partial u_h}{\partial \vec{n}_{L,K}}\Big)\, \sigma_E \, ds &= \Big|\Big(g_E - \frac{\partial u_h}{\partial \vec{n}_{L,K}}\Big)\Big|_E\Big|\, h_E^{-1/2} \int_E b_E^A \, ds \\ &= C^*_{\widehat{K}} \Big\|\Big(g_E - \frac{\partial u_h}{\partial \vec{n}_{L,K}}\Big)\Big\|_{L^2(E)},\end{aligned}$$

that is,

$$\Big\|\Big(g_E - \frac{\partial u_h}{\partial \vec{n}_{L,K}}\Big)\Big\|_{L^2(E)} = \frac{1}{C^*_{\widehat{K}}} \int_E \Big(g_E - \frac{\partial u_h}{\partial \vec{n}_{L,K}}\Big)\, \sigma_E \, ds. \tag{68}$$

Results (60), (59), together with Lemma 4.6 yield the inequalities

$$\begin{aligned}
&\int_E \Big(g_E - \frac{\partial u_h}{\partial \vec{n}_{L,K}}\Big)\, \sigma_E \, ds = \int_E \Big(g - \frac{\partial u_h}{\partial \vec{n}_{L,K}}\Big)\, \sigma_E \, ds + \int_E (g_E - g)\, \sigma_E \, ds \\
&= \int_\Omega f\, \sigma_E \, d\vec{x} + \int_{\Gamma_N} g\, \sigma_E \, ds - \int_\Omega \Big(\sum_{i,j=1}^2 a_{ij}\, \frac{\partial u_h}{\partial x_j}\, \frac{\partial \sigma_E}{\partial x_i}\Big)\, d\vec{x} - \int_K f\, \sigma_E \, d\vec{x} \\
&+ \int_E (g_E - g)\, \sigma_E \, ds = \int_K \Big(\sum_{i,j=1}^2 a_{ij}\, \frac{\partial e_h}{\partial x_j}\, \frac{\partial \sigma_E}{\partial x_i}\Big)\, d\vec{x} - \int_K f\, \sigma_E \, d\vec{x} + \int_E (g_E - g)\, \sigma_E \, ds \\
&\le C_A\, \|\nabla e_h\|_{L^2(K)} \|\nabla \sigma_E\|_{L^2(K)} + \|f\|_{L^2(K)} \|\sigma_E\|_{L^2(K)} + \|g_E - g\|_{L^2(E)} \|\sigma_E\|_{L^2(E)} \\
&\le C_A\, \|\nabla e_h\|_{L^2(K)}\, h_E^{-3/2}\, \widetilde{\lambda}^{1/2}\, h_{\widehat{K}}\, s_K\, \Big(\int_K b_E^A \, d\vec{x}\Big)^{1/2} \\
&+ \|f\|_{L^2(K)}\, h_E^{-1/2} \Big(\int_K b_E^A \, d\vec{x}\Big)^{1/2} + \|g_E - g\|_{L^2(E)}\, h_E^{-1/2} \Big(\int_E b_E^A \, ds\Big)^{1/2} \\
&\le \widehat{C}^{1/2}\, s_K^{1/2} \Big[C_A\, \widetilde{\lambda}^{1/2}\, h_{\widehat{K}}\, s_K\, h_E^{-1/2}\, \|\nabla e_h\|_{L^2(K)} + h_E^{1/2}\, \|f\|_{L^2(K)}\Big] \\
&+ (C^*_{\widehat{K}})^{1/2}\, \|g_E - g\|_{L^2(E)} \le \widehat{C}^{1/2}\, s_K^{1/2} \Big[C_A\, \widetilde{\lambda}^{1/2}\, h_{\widehat{K}}\, s_K\, h_E^{-1/2}\, \|\nabla e_h\|_{L^2(K)} \\
&+ h_E^{1/2}\, \|f - f_K\|_{L^2(K)} + h_E^{1/2}\, \|f_K\|_{L^2(K)}\Big] + (C^*_{\widehat{K}})^{1/2}\, \|g_E - g\|_{L^2(E)},
\end{aligned}$$

where the inclusion $\mathrm{supp}(\sigma_E) \subset K$ has been exploited too. Now, by applying results (62) and (63) to the term $\|f_K\|_{L^2(K)}$, we derive that

$$\begin{aligned}\int_E \Big(g_E - \frac{\partial u_h}{\partial \vec{n}_{L,K}}\Big)\, \sigma_E \, ds &\le \overline{C}^{**} \Big[C_A h_E^{-1/2}\, \|\nabla e_h\|_{L^2(K)} \\ &+ h_E^{1/2}\, \|f - f_K\|_{L^2(K)} + \|g_E - g\|_{L^2(E)}\Big].\end{aligned}$$

Now we are in a position to bound the term (III) in (58). Moving from (68) and thanks to the geometrical relations (47) and (19), and results (31) and (33), we have

$$
\begin{aligned}
(\mathrm{III}) \quad &\le \quad \frac{\lambda_{1,K}^{1/2}\,\overline{C}^{**}}{C^*_{\widehat{K}}}\Big(\sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,N}}\Big[\frac{h_E^{-1/2}}{\gamma_{min}^{1/2}}C_A\,|||e_h|||_K + h_E^{1/2}\,\|f-f_K\|_{L^2(K)} \\
&+ \quad \|g_E-g\|_{L^2(E)}\Big]\Big)\,w_K(e_h^*) \\
&\le \quad \frac{2\,\lambda_{1,K}^{1/2}\,\overline{C}^{**}}{C^*_{\widehat{K}}}\Big[\frac{C_A}{(\rho_{\widehat{K}}\,\gamma_{min}\,\lambda_{2,K})^{1/2}}\,|||e_h|||_{\Delta_K} + \lambda_{1,K}^{1/2}\,h_{\widehat{K}}^{1/2}\|f-f_K\|_{L^2(K)} \\
&+ \quad \sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,N}}\|g_E-g\|_{L^2(E)}\Big]\,w_K(e_h^*) \\
&\le \quad \frac{2\,\overline{C}^{**}}{C^*_{\widehat{K}}}\Big[\frac{s_K^{1/2}}{(\rho_{\widehat{K}}\,\gamma_{min})^{1/2}}\,C_A\,|||e_h|||_{\Delta_K} + \lambda_{1,K}\,h_{\widehat{K}}^{1/2}\|f-f_K\|_{L^2(K)} \\
&+ \quad \lambda_{1,K}^{1/2}\sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,N}}\|g_E-g\|_{L^2(E)}\Big]\,s_K^{1/2}\,\frac{(1+\nu_K)}{\gamma_{min}^{1/2}}\,|||e_h|||_{\Delta_K} \\
&= \quad \frac{2\,\overline{C}^{**}}{C^*_{\widehat{K}}}\,\frac{(1+\nu_K)}{\gamma_{min}^{1/2}}\Big[\frac{s_K}{(\rho_{\widehat{K}}\,\gamma_{min})^{1/2}}\,C_A\,|||e_h|||^2_{\Delta_K} \\
&\qquad\qquad + \frac{\lambda_{1,K}^{3/2}}{\lambda_{2,K}^{1/2}}\,h_{\widehat{K}}^{1/2}\,|||e_h|||_{\Delta_K}\,\|f-f_K\|_{L^2(K)} \\
&\qquad\qquad + \frac{\lambda_{1,K}}{\lambda_{2,K}^{1/2}}\,|||e_h|||_{\Delta_K}\sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,N}}\|g_E-g\|_{L^2(E)}\Big].
\end{aligned}
$$

Finally, employing Young's inequality on the last two terms, we obtain

$$
\frac{\lambda_{1,K}^{3/2}}{\lambda_{2,K}^{1/2}}\,|||e_h|||_{\Delta_K}\|f-f_K\|_{L^2(K)} \le \frac{\gamma_{min}^{1/2}}{4}\,s_K\,|||e_h|||^2_{\Delta_K} + \frac{\lambda_{1,K}^2}{\gamma_{min}^{1/2}}\|f-f_K\|^2_{L^2(K)},
$$

and

$$
\begin{aligned}
&\frac{\lambda_{1,K}}{\lambda_{2,K}^{1/2}}\,|||e_h|||_{\Delta_K}\sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,N}}\|g_E-g\|_{L^2(E)} \\
&\le \frac{\gamma_{min}^{1/2}}{4}\,s_K\,|||e_h|||^2_{\Delta_K} + \frac{2}{\gamma_{min}^{1/2}}\,\lambda_{1,K}\sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,N}}\|g_E-g\|^2_{L^2(E)},
\end{aligned}
$$

respectively. These inequalities provide result (67). □

Lemmas 6.1, 6.2 and 6.3 yield the desired efficiency estimate (24) for the local error indicator $\eta_{K,T}$:

**Proposition 6.1** *Under the Assumption* 4.1*, it can be proved that*

$$
\begin{aligned}
\eta_{K,T} \le \underline{C}\Big[ \qquad & |||e_h|||^2_{\Delta_K} + \lambda_{1,K}^2\sum_{T\in\Delta_K}\|f-f_T\|^2_{L^2(T)} \\
+ \quad & \lambda_{1,K}\sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_{h,N}}\|g_E-g\|^2_{L^2(E)}\Big]^{1/2},
\end{aligned}
$$

*with* $\underline{C} = \underline{C}(\widehat{K},\, N,\, C_\Delta,\, \nu_K,\, \max_{T\in\Delta_K} s_T,\, C_A,\, \gamma_{min})$.

**Remark 6.1 (A cautionary tale)** *It is well known that recovery-based estimators, though possessing several attractive features, such as, their ease of implementation, generality and ability to produce quite accurate estimators, also have some drawbacks. For example, a kind of dangerous behavior is highlighted in* [1]*, Section* 4.7*, and it is referred to as* a cautionary tale*. The authors construct an example in which the recovery-based estimator produces an estimated error equal to zero, while the actual error can be arbitrarily large. A similar phenomenon is also addressed in* [36]*, Remark* 1.7*, where it is shown how to construct a problem having $u_h = 0$ as discrete solution, while $||u - u_h||_{H^1(\Omega)} \neq 0$. We point out that this situation is independent of the estimator being employed, e.g. recovery-based or residual-based estimator. As long as the estimator uses only "finite" or "lumped" information extracted from the numerical solution and/or from the data of the problem, it will always be possible to devise cases when the estimators fail to be reliable. In particular, in* [36]*, the author's conclusion is that this kind of situations will always occur as long as it is not possible to evaluate exactly $||f||_{L^2(K)}$, and that this problem is cured by further refinements of the mesh. In other words, this phenomenon is related to the so-called* data oscillation*, i.e. $||f - f_K||_{L^2(K)}$. This term may dominate entirely the error estimate but it is usually not included in the definition of the local error estimator, due to its uncomputable nature. In these cases, it is obvious that the error estimator is not reliable. Conditions guaranteeing that the data oscillation is small should be satisfied, then no phenomena such as the cautionary tale might occur.*

# 7 Numerical algorithm and validation

In this section we first describe the numerical algorithm used to compute a numerical solution satisfying a given tolerance, from an error estimator as (26). Then we validate this algorithm on some numerical test cases.

## 7.1 Generation of the metric

The anisotropic information provided by the estimator (26) can be employed in two different ways,

1. one just computes, on a given mesh, the quantity (26), thus obtaining an estimate for the energy norm of the error;

2. one uses (26) in a *predictive fashion*, i.e. to construct a mesh satisfying an optimality condition. Typical choices for this are

   a) given a constraint on the maximum number of elements, find the mesh providing the most accurate numerical solution;
   b) given a constraint on the accuracy of the numerical solution, find the mesh with the least number of elements.

In what follows, we describe our approach, which fits into the 2.b) case.

Our numerical procedure is based on the definition of mesh metric (see [15]). In particular, it is a standard way to endow the domain $\Omega$ with a *metric*, induced by a symmetric positive definite tensor field $\widetilde{M_\Omega} : \Omega \to \mathbb{R}^2$ such that

$$\widetilde{M_\Omega} = \widetilde{R}^T \widetilde{\Lambda}^{-2} \widetilde{R}.$$

The tensors $\widetilde{\Lambda} = \mathrm{diag}(\widetilde{\lambda}_1, \widetilde{\lambda}_2)$ and $\widetilde{R}^{\mathrm{T}} = [\vec{\widetilde{r}}_1 \ \vec{\widetilde{r}}_2]$ are positive diagonal and orthogonal, respectively while the quantities $\vec{\widetilde{r}}_i$ and $\widetilde{\lambda}_i$ provide the stretching directions and spacing of the grid to be generated (see Fig. 1). The metric $\widetilde{M}_\Omega$, however, is not known explicitly (e.g. as a function of $\vec{x} = (x_1, x_2)$), but is defined implicitly by the error estimator (26) and the requirement b) above.

Suppose first that $\widetilde{M}_\Omega$ is given; then we show how the problem of constructing a mesh associated with the metric, in some sense to be defined shortly, can be posed in terms of a "matching condition". With this aim, we recall that, for any given mesh $\mathcal{T}_h$, we can define a piecewise constant metric $\widetilde{M}_{\mathcal{T}_h}$, such that, $\widetilde{M}_{\mathcal{T}_k}|_K = \widetilde{M}_K = B_K^{-2} = R_K^T \Lambda_K^{-2} R_K$, for any $K \in \mathcal{T}_h$, the matrices being the ones defined in Section 2. With respect to this metric, any edge of triangle $K$ has unitary length. Indeed, for any $\vec{e} \in \mathcal{E}(K)$, we have

$$\begin{aligned} \vec{e}^{\,T} \widetilde{M}_K \vec{e} &= \vec{e}^{\,T} B_K^{-2} \vec{e} = \vec{e}^{\,T} B_K^{-1} Z_K^{-T} Z_K^{-1} B_K^{-1} \vec{e} = \vec{e}^{\,T} M_K^{-T} M_K^{-1} \vec{e} \\ &= ||M_K^{-1} \vec{e}||_2^2 = ||\vec{\widehat{e}}||_2^2 = 1, \end{aligned}$$

where $\vec{\widehat{e}} = T_K^{-1}(\vec{e})$.

For practical reasons, we approximate the quantities $\widetilde{\lambda}_1, \widetilde{\lambda}_2, \vec{\widetilde{r}}_1, \vec{\widetilde{r}}_2$ defining $\widetilde{M}_\Omega$ by piecewise constant functions over the triangulation $\mathcal{T}_h$, such that $\vec{\widetilde{r}}_i\big|_K = \vec{\widetilde{r}}_{i,K}$, $\widetilde{\lambda}_i\big|_K = \widetilde{\lambda}_{i,K}$, for any $K \in \mathcal{T}_h$ and with $i = 1, 2$.

Thus, we introduce the following *matching criterion*:

**Definition 7.1** *A mesh $\mathcal{T}_h$ matches a given metric $\widetilde{M}_\Omega$ if, for any $K \in \mathcal{T}_h$,*

$$\widetilde{M}_\Omega\big|_K = \widetilde{M}_{\mathcal{T}_k}\big|_K = \widetilde{M}_K,$$

*i.e. $\vec{\widetilde{r}}_{i,K} = \vec{r}_{i,K}$ and $\widetilde{\lambda}_{i,K} = \lambda_{i,K}$, for $i = 1, 2$.*

The determination of $\widetilde{M}_\Omega$ and, in view of the definition above, of the corresponding matching mesh, is usually carried out via an iterative procedure: starting from a given mesh $\mathcal{T}_h^k$, playing the role of a background mesh, i.e., a grid where the information concerning the new metric is stored and used to update the new mesh, by analyzing the solution on $\mathcal{T}_h^k$, we seek for an optimal metric $\widetilde{M}_\Omega^{k+1}$ (piecewise constant over $\mathcal{T}_h^k$) to drive the generation of a better, adapted grid $\mathcal{T}_h^{k+1}$. At each step of this procedure, to compute $\widetilde{M}_\Omega^{k+1}$, we start from the definition of the global estimator (26), and then, for convenience, we rewrite the local estimators $\eta_{K,T}$ as

$$\eta_{K,T}^2 = |K|^{3/2} \widetilde{\rho}_K(u_h) \Big[ s_K \left( \vec{r}_{1,K}^{\,T} \widetilde{G}_K(e_h^*) \vec{r}_{1,K} \right) + \frac{1}{s_K} \left( \vec{r}_{2,K}^{\,T} \widetilde{G}_K(e_h^*) \vec{r}_{2,K} \right) \Big]^{1/2}, \qquad (69)$$

where $\widetilde{\rho}_K(u_h) = \rho_K(u_h)|K|^{-1/2}$ and $\widetilde{G}_K(e_h^*) = G_K(e_h^*)|K|^{-1}$ are the scaled residual and matrix related to the reconstructed derivatives, respectively, the dependence on $k$ being dropped. This scaling is driven with the aim of making all the terms in the right-hand side of (69) approximately independent of the measure of triangle $K$, at least asymptotically (i.e., when the mesh is sufficiently fine), thus lumping this information only in a multiplicative constant.

After scaling, we resort to the 2.b) requirement mentioned above, i.e. for a fixed value of $\eta_{K,T}$, we minimize the number of triangles by maximizing $|K|$. This amounts to solving the following constrained minimization problem:

$$\begin{cases} \text{find } s_K \text{ and } \vec{r}_{1,K} \text{ such that} \\ I(s_K, \vec{r}_{1,K}) = s_K \big(\vec{r}_{1,K}^{\,T}\, \widetilde{G}_K(e_h^*)\, \vec{r}_{1,K}\big) + \dfrac{1}{s_K}\big(\vec{r}_{2,K}^{\,T}\, \widetilde{G}_K(e_h^*)\, \vec{r}_{2,K}\big) \quad \text{be minimized,} \\ \text{where } s_K \geq 1,\ \ \vec{r}_{1,K}, \vec{r}_{2,K} \in \mathbb{R}^2,\ \ \|\vec{r}_{1,K}\|_2 = \|\vec{r}_{2,K}\|_2 = 1,\ \ \text{and } \vec{r}_{1,K} \cdot \vec{r}_{2,K} = 0. \end{cases} \tag{70}$$

The solution to this problem is provided in the following

**Proposition 7.1** *The solution* $(\widetilde{s}_K, \widetilde{\vec{r}}_{1,K})$ *of* (70) *is such that* $\widetilde{\vec{r}}_{1,K}$ *is parallel to the eigenvector associated with the minimum eigenvalue of* $\widetilde{G}_K(e_h^*)$ *while*

$$\widetilde{s}_K = \sqrt{\frac{\max(\text{eig}(\widetilde{G}_K(e_h^*)))}{\min(\text{eig}(\widetilde{G}_K(e_h^*)))}},$$

$\text{eig}(\widetilde{G}_K(e_h^*))$ *being the set of the eigenvalues of* $\widetilde{G}_K(e_h^*)$.

**Proof.** Let us denote by $(\vec{v}_{1,K}, \sigma_{1,K})$ and $(\vec{v}_{2,K}, \sigma_{2,K})$ the two couples of orthonormal eigenvectors and eigenvalues of the symmetric positive semi-definite matrix $\widetilde{G}_K(e_h^*)$, where, without loss of generality, we assume $\sigma_{1,K} \geq \sigma_{2,K} (> 0)$. Let us expand $\vec{r}_{1,K}$ and $\vec{r}_{2,K}$ as

$$\vec{r}_{1,K} = a_1\, \vec{v}_{1,K} + a_2\, \vec{v}_{2,K}, \quad \vec{r}_{2,K} = -a_2\, \vec{v}_{1,K} + a_1\, \vec{v}_{2,K},$$

where $a_1^2 + a_2^2 = 1$. This gives,

$$\vec{r}_{1,K}^{\,T}\, \widetilde{G}_K(e_h^*)\, \vec{r}_{1,K} = \sigma_{1,K}\, a_1^2 + \sigma_{2,K}\, a_2^2, \quad \vec{r}_{2,K}^{\,T}\, \widetilde{G}_K(e_h^*)\, \vec{r}_{2,K} = \sigma_{2,K}\, a_1^2 + \sigma_{1,K}\, a_2^2.$$

Noticing that $(\vec{r}_{1,K}^{\,T}\, \widetilde{G}_K(e_h^*)\, \vec{r}_{1,K}) + (\vec{r}_{2,K}^{\,T}\, \widetilde{G}_K(e_h^*)\, \vec{r}_{2,K}) = \sigma_{1,K} + \sigma_{2,K}$, we are led to minimizing the quantity

$$I(s_K, \vec{r}_{1,K}) = \Big(s_K - \frac{1}{s_K}\Big)\, \vec{r}_{1,K}^{\,T}\, \widetilde{G}_K(e_h^*)\, \vec{r}_{1,K} + \frac{\sigma_{1,K} + \sigma_{2,K}}{s_K}$$

with respect to $\vec{r}_{1,K}$ and $s_K$. First notice that, for any given $s_K > 1$, the above expression is minimized when $\vec{r}_{1,K}^{\,T}\, \widetilde{G}_K(e_h^*)\, \vec{r}_{1,K}$ is minimum. This occurs when the Rayleigh quotient $\vec{r}_{1,K}^{\,T}\, \widetilde{G}_K(e_h^*)\, \vec{r}_{1,K}$ is equal to the minimum eigenvalue of $\widetilde{G}_K(e_h^*)$, i.e., when $\vec{r}_{1,K}^{\,T}\, \widetilde{G}_K(e_h^*)\, \vec{r}_{1,K} = \sigma_{2,K}$ and $\vec{r}_{1,K}$ is parallel to the eigenvector $\vec{v}_{2,K}$. When this is the case, it also holds $\vec{r}_{2,K}^{\,T}\, \widetilde{G}_K(e_h^*)\, \vec{r}_{2,K} = \sigma_{1,K}$ and $\vec{r}_{2,K}$ is parallel to $\vec{v}_{1,K}$. In turn, the minimization with respect to $s_K$ provides us with the optimal value of $s_K$

$$\widetilde{s}_K = \sqrt{\frac{\sigma_{1,K}}{\sigma_{2,K}}} \geq 1.$$

Then we consider some particular cases:

a) when $s_K = 1$, $I(1, \vec{r}_{1,K}) = \sigma_{1,K} + \sigma_{2,K}$, independently of $\vec{r}_{1,K}$, which can thus be chosen arbitrarily;

b) when $\sigma_{2,K} = 0$ we have to limit $\widetilde{s}_K$ to a suitable user-defined maximum value;

c) when $\sigma_{1,K} = \sigma_{2,K}$, all the Rayleigh quotients are equal to the common eigenvalue of $\widetilde{G}_K(e_h^*)$. In this case the solution of the minimization problem provides us with an arbitrary vector $\widetilde{\vec{r}}_{1,K}$ and with $\widetilde{s}_K = 1$, consistently with a). This corresponds to the isotropic case of an unstretched triangle.

$\square$

To define completely the metric we are left with computing the values of $\widetilde{\lambda}_{1,K}$ and $\widetilde{\lambda}_{2,K}$. For this purpose we use the equidistribution of the error. More precisely, since we can only act on the local estimators, we impose that $\eta_{K,T} = \tau$, for any $K \in \mathcal{T}_h$, where $\tau$ is a given tolerance. By combining the result of Proposition 7.1 with the equidistribution of the error, we single out the values of $\widetilde{\lambda}_{1,K}$ and $\widetilde{\lambda}_{2,K}$ as the solutions of the system

$$\begin{cases} \dfrac{\widetilde{\lambda}_{1,K}}{\widetilde{\lambda}_{2,K}} = \widetilde{s}_K \equiv q, \\ \widetilde{\lambda}_{1,K}\widetilde{\lambda}_{2,K} = \left[ \dfrac{\tau^4}{|\widehat{K}|^3 \, \widetilde{\rho}_K^2(u_h)\Big(\widetilde{s}_K\,\sigma_{2,K} + \dfrac{\sigma_{1,K}}{\widetilde{s}_K}\Big)} \right]^{1/3} \equiv p, \end{cases}$$

from which it follows that $\widetilde{\lambda}_{1,K} = \sqrt{pq}$, $\widetilde{\lambda}_{2,K} = \sqrt{p/q}$.

Finally, we recall that the global metric $\widetilde{M}_\Omega^{k+1}$ is obtained by letting $\vec{\widetilde{r}}_i|_K = \vec{\widetilde{r}}_{i,K}$ and $\widetilde{\lambda}_i|_K = \widetilde{\lambda}_{i,K}$, with $i = 1, 2$. Once the metric has been computed, the new mesh is built by a metric-driven mesh generator, e.g. BAMG [19], which receives as input the metric $\widetilde{M}_\Omega^{k+1}$ and returns the mesh $\mathcal{T}_h^{k+1}$ satisfying (within a certain tolerance) the matching condition.

## 7.2 Numerical Assessment

The procedure provided in Section 7.1 to get an adapted mesh satisfying criterion 2.b) together with an error equidistribution approach, is validated in this section. Moreover, to assess the robustness of the proposed anisotropic error estimator we study its behavior on some a priori chosen non-optimal grids. A comparison with the standard ZZ and the residual-based error estimators is also provided.

### 7.2.1 The first test case

We solve the Poisson problem in $\Omega = (0,1)^2$ with homogeneous Dirichlet boundary conditions, namely we choose $a_{ij} = \delta_{ij}$, $\gamma = 0$, and $\Gamma_N = \emptyset$ in (9), with $\delta_{ij}$ the Kronecker symbol. The forcing term $f$ is chosen such that the exact solution is given by

$$u = 4\left(1 - e^{-100x_1} - x_1(1 - e^{-100})\right) x_2\,(1 - x_2).$$

The solution $u$ exhibits an exponential layer along the boundary $x_1 = 0$, with an initial steepness of 100. The presence of the boundary layer justifies the use of an anisotropic mesh adaption technique.

Moving from the error estimator (26) and from an initial uniform mesh of about 1000 elements, we apply the procedure in Section 7.1 to get a new metric guaranteeing a prescribed accuracy $\tau = 10^{-3}$ of the approximate solution $u_h$ and an error equidistribution on the mesh elements, and the corresponding adapted mesh. A priori one would expect an adapted grid with the triangles stretched along the boundary layer to capture the directional features of the solution at hand. This is confirmed by Fig. 4 where the fourth adapted mesh, of about 5000 elements, is shown. The orientation and deformation of the mesh elements (shortest edges oriented across the direction of maximal variation of the solution) guarantee a reduction of the number of triangles, that is of the computational cost associated with the approximation of the problem at hand. A zoom of the mesh in correspondence with the boundary layer is also provided on the right of Fig. 4.

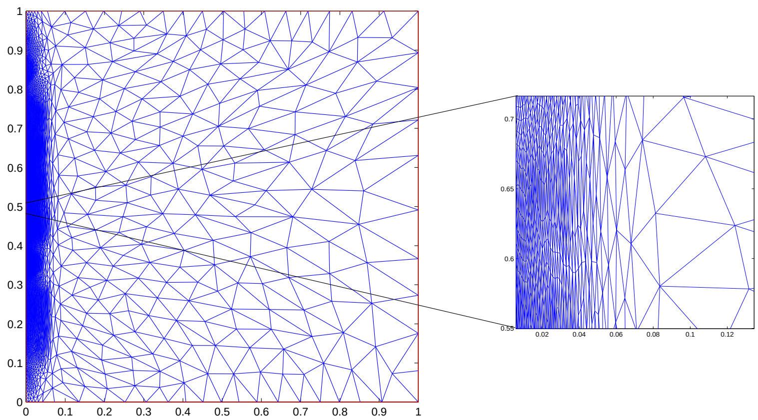

Figure 4: First test case: fourth adapted mesh obtained via the anisotropic ZZ error estimator (26).

In Fig. 5 we provide an adapted mesh obtained moving from the same initial mesh, but using another error estimator [12]. In more detail, we exploit an anisotropic counterpart of the dual-based error analysis in [4]. Also in this case, by suitably choosing the adjoint problem to be solved, we control the energy norm of the discretization error. By comparing the meshes in Figs. 4 and 5, we note that the distribution of the triangles is very similar. However, the grid in Fig. 5 has less elements (about 4000), the boundary layer being captured more sharply (compare the thickness of the refined areas near the side $x_1 = 0$). This sharpness, though, is balanced by the higher computational cost characterizing this second approach, due to the additional resolution of the dual problem.
We remark that a control of linear functionals of the discretization error, identifying physically meaningful quantities, is allowed by the dual-based analysis too. This approach yields meshes characterized by a distribution of triangles varying according to the quantity we are interested in. We refer to [12] for an example of such a technique.

To assess the robustness of the error estimator (26), let us study its behavior on a priori chosen meshes, following the criterion 1. in Section 7.1.
First, we consider a series of stretched meshes parametrized by a value $k$ such that, starting from a uniform $10 \times 10$ mesh of the domain $\Omega$, the new mesh is obtained by the transformation

$$x_1^{\text{new}} = \frac{\exp\left(k x_1^{\text{ini}}\right) - 1}{\exp\left(k\right) - 1}, \tag{71}$$

where $x_1^{\text{ini}}$ takes on the values of the horizontal coordinates of the nodes of the initial uniform grid. Notice that the meshes generated by the criterion (71) are correctly oriented (though not necessarily of the correct size), the triangles being stretched along the $x_2$-axis and gathered in correspondence with the side $x_1 = 0$.

Table 1 collects the most meaningful quantities related to this assessment. In particular, for the current mesh, from left to right, we find:

- the value $k$;
- the maximum and minimum stretching factor $s_K$;
- the total number `Nv` of vertices;

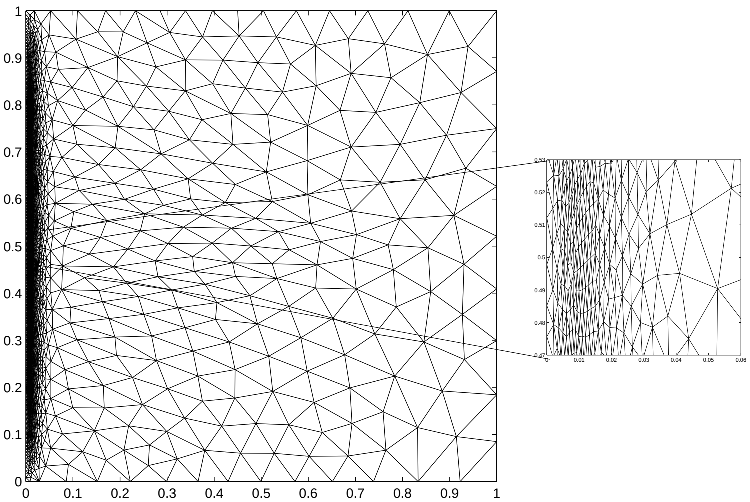

Figure 5: First test case: fourth adapted mesh obtained via an anisotropic dual-based approach.

- the value $\nu = \dfrac{||G^{ZZ}u_h - \nabla u||_{L^2(\Omega)}}{||\nabla u_h - \nabla u||_{L^2(\Omega)}}$, related to the condition (27);
- the energy norm $|||e_h|||$ of the discretization error;
- the effectivity index associated with the standard isotropic ZZ error estimator

$$\theta^{ZZ} = \frac{||G^{ZZ}u_h - \nabla u_h||_{L^2(\Omega)}}{|||e_h|||}\,;$$

- the effectivity index associated with the standard isotropic residual-based error estimator

$$\theta^{\text{Res}} = \frac{\Big(\sum_{K\in\mathcal{T}_h}\big(h_K^2||r_K(u_h)||^2_{L^2(K)} + \dfrac{h_K}{2}||R_K(u_h)||^2_{L^2(K)}\big)\Big)^{1/2}}{|||e_h|||}\,, \qquad (72)$$

  $r_K(u_h)$ and $R_K(u_h)$ being defined as in (14)-(15);
- the effectivity index associated with the anisotropic error estimator (26)

$$\theta^{A} = \frac{\eta_T}{|||e_h|||}\,.$$

Notice that the total number of mesh vertices is invariant, i.e. it does not depend on $k$.
Moreover, until the maximum stretching factor is about 26 (i.e. $k \leq 5$), the behavior of the energy norm of the discretization error as well as of the effectivity indexes $\theta^{ZZ}$ and $\theta^{A}$ is the expected one, $|||e_h|||$ diminishing and the indexes converging to constant values (about 1 and 4, respectively). On the other hand, the effectivity index $\theta^{\text{res}}$ associated with the residual-based error estimator gets larger and larger.

Table 1: First test case: values associated with the meshes generated by the criterion (71)

| $k$ | $\max_K s_K - \min_K s_K$ | Nv | $\nu$ | $\|\|\|e_h\|\|\|$ | $\theta^{ZZ}$ | $\theta^{\mathrm{res}}$ | $\theta^A$ |
|---|---|---|---|---|---|---|---|
| 1 | 1.73 – 1.73 | 121 | 0.9961 | 8.8780 | 0.0772 | 1.0416 | 0.3635 |
| 2.5 | 4.62 – 1.75 | 121 | 0.9057 | 3.4868 | 0.5470 | 12.7906 | 2.3530 |
| 5 | 26.2 – 1.75 | 121 | 0.7328 | 1.2096 | 0.9760 | 51.0799 | 4.3298 |
| 10 | 1480 – 1.77 | 121 | 0.8596 | 1.4224 | 0.8812 | 46.9202 | 4.0841 |
| 20 | $8.76 \cdot 10^6$ – 1.77 | 121 | 1.0568 | 2.7707 | 0.5798 | 23.1898 | 2.9348 |
| 30 | $6.46 \cdot 10^{10}$ – 2.60 | 121 | 0.9602 | 4.5719 | 0.3983 | 11.5777 | 2.8185 |
| 40 | $5.10 \cdot 10^{14}$ – 6.47 | 121 | 1.0206 | 5.4212 | 0.3636 | 11.1959 | 3.0221 |

For $10 \leq k \leq 40$, that is for extremely high values of the maximum stretching factor, a different and unexpected trend is shown by the quantities $|||e_h|||$, $\theta^{ZZ}$ and $\theta^A$, probably due to the lack of a sufficient number of mesh nodes along the $x_1$ axis, far from the boundary layer, or to the maximum very large aspect ratio, up to $10^{14}$. As for the quantity $\nu$ in the fourth column, we note that it is below the value 1, except for two cases only.

The same quantities collected in Table 1 are computed on a second series of stretched grids, obtained by refining the initial grid along the wrong direction, i.e. the $x_2$ axis and near the top side of the domain, by the relation

$$x_2^{\mathrm{new}} = \frac{\exp\left(k x_2^{\mathrm{ini}}\right) - 1}{\exp\left(k\right) - 1}, \tag{73}$$

where $x_2^{\mathrm{ini}}$ takes on the values of the vertical coordinates of the nodes of the initial uniform grid. This choice aims at comparing the error estimator (26) with the standard ZZ and residual-based estimators, in a very unfavorable situation. The results are summarized in Table 2.

Table 2: First test case: values associated with the meshes generated by the criterion (73)

| $k$ | $\max_K s_K - \min_K s_K$ | Nv | $\nu$ | $\|\|\|e_h\|\|\|$ | $\theta^{ZZ}$ | $\theta^{\mathrm{res}}$ | $\theta^A$ |
|---|---|---|---|---|---|---|---|
| 2.5 | 4.62 – 1.75 | 121 | 0.9995 | 8.9567 | 0.0781 | 1.2829 | 0.3652 |
| 5 | 26.2 – 1.75 | 121 | 0.9944 | 9.1328 | 0.0781 | 1.8650 | 0.3725 |
| 10 | 1480 – 1.77 | 121 | 0.9925 | 10.0010 | 0.0600 | 3.1409 | 0.3244 |
| 20 | $8.76 \cdot 10^6$ – 1.77 | 121 | 0.9957 | 11.8100 | 0.0314 | 4.7254 | 0.2265 |
| 30 | $6.46 \cdot 10^{10}$ – 2.60 | 121 | 0.9976 | 12.5180 | 0.0205 | 5.3389 | 0.1912 |
| 40 | $5.10 \cdot 10^{14}$ – 6.47 | 121 | 0.9987 | 12.7495 | 0.0138 | 5.5666 | 0.1686 |

The values of the energy norm $|||e_h|||$ are large and increase with $k$, due to the wrong choice of the meshes, while the effectivity indexes $\theta^{ZZ}$ and $\theta^A$ get smaller and smaller. This testifies that, for the Poisson problem at hand, the isotropic ZZ and the anisotropic error estimators underestimate the true error on such kind of grids. On the other hand, the effectivity index (72) seems to be stabilizing about the value 5.5. In this case, the quantity $\nu$ is less than 1 but very close to it, probably again due to the unfavorable choice of the mesh.

Finally, we evaluate the three error estimators above by approximating the Poisson problem at hand on structured grids obtained by subdividing the horizontal and vertical

sides of the domain by N1 and N2 uniform subintervals, respectively, and on criss-cross type meshes characterized by $\mathtt{N1} = \mathtt{N2}$ uniform subdivisions of all of the boundary edges. For both the structured and criss-cross meshes, the maximum and minimum values of the stretching factor $s_K$ are equal to 1.73 and 1, respectively.
The results are collected in Tables 3 and 4, respectively.

Table 3: First test case: values associated with the structured meshes

| N1 − N2 | Nv | $\nu$ | $\|\|\|e_h\|\|\|$ | $\theta^{ZZ}$ | $\theta^{\rm res}$ | $\theta^A$ |
|---|---|---|---|---|---|---|
| 20 − 3 | 84 | 0.9670 | 5.9649 | 0.3105 | 13.7252 | 1.5044 |
| 40 − 6 | 287 | 0.9093 | 3.5215 | 0.5428 | 20.6935 | 2.4027 |
| 80 − 12 | 1053 | 0.8718 | 1.8925 | 0.7636 | 24.9598 | 3.3841 |
| 160 − 24 | 4025 | 0.7865 | 0.9666 | 0.9056 | 27.6362 | 4.1149 |

In both cases, the energy norm of the discretization error reduces as the grid is refined, while the effectivity indexes $\theta^{ZZ}$ and $\theta^A$ get near the values 1 and 4, respectively. On the contrary, the estimates of the norm $|||e_h|||$ predicted by the residual-based error estimator are not reliable, as the large values of $\theta^{\rm res}$ suggest. In both cases, the values of $\nu$ are always less than 1 (except for the first mesh in Table 4), probably due to the regularity of the meshes.

Table 4: First test case: values associated with the criss-cross meshes

| N1 = N2 | Nv | $\nu$ | $\|\|\|e_h\|\|\|$ | $\theta^{ZZ}$ | $\theta^{\rm res}$ | $\theta^A$ |
|---|---|---|---|---|---|---|
| 4 | 41 | 1.0024 | 10.1142 | 0.0504 | 0.8272 | 0.2427 |
| 8 | 145 | 0.9786 | 6.5950 | 0.1920 | 3.2989 | 0.8779 |
| 16 | 545 | 0.9226 | 4.0655 | 0.4307 | 5.9274 | 1.8749 |
| 32 | 2113 | 0.8935 | 2.2421 | 0.6820 | 7.9493 | 2.9682 |
| 64 | 8321 | 0.8369 | 1.1542 | 0.8639 | 9.4810 | 3.8535 |

### 7.2.2 The second test case

We are still concerned with the solution of the Poisson problem in $\Omega = (0,1)^2$ with homogeneous Dirichlet boundary conditions. The choice $f = -20\sin(6\,\pi\,x_2)(-1 - 18\,\pi^2\,x_1 + 18\,\pi^2\,x_1^2)$ is made for the forcing term, so that the solution $u$ is given by

$$u = 10\,x_1\,(1 - x_1)\,\sin(6\,\pi\,x_2).$$

First, we solve the problem on a quasi-uniform mesh with average size 1/20. Then according to the criterion 2.b) cited in Section 7.1, we exploit the anisotropic information provided by the error estimator (26) to get an adapted mesh with an almost equidistributed error per element equal to $\tau = 10^{-3}$. This goal is reached after four iterations. The sequence of the four adapted grids is collected in Fig. 6 (left-right top-bottom). It appears that the directional features of the solution $u$ match the grids. The six more refined horizontal zones correspond to the regions of $\Omega$ where the maxima and minima of $u$ are reached.
Fig. 7 shows the discrete solution $u_h$ computed on the third adapted mesh. The two plots refer to different view points.
The same quantities considered in Tables 1-4 are now computed on the initial and on the

four adapted meshes, together with the total number `Nt` of mesh triangles. The corresponding values are gathered in Table 5. It can be noted that the anisotropic estimator is very little sensitive to the variations of the stretching factors. It is remarkable also that, in this example, the standard $ZZ$ error estimator has effectivity indexes very close to one. However it is not straightforward to extract some anisotropic information from this estimator to use in a predictive fashion. Finally, the residual-based error estimator turns out again not to be a reliable quantity, the corresponding effectivity index $\theta^{\mathrm{res}}$ having an increasing trend. Notice that the sequence of the values of $\nu$ is decreasing as the adaptation procedure progresses, assessing, in this case, the good properties of the ZZ recovery procedure.

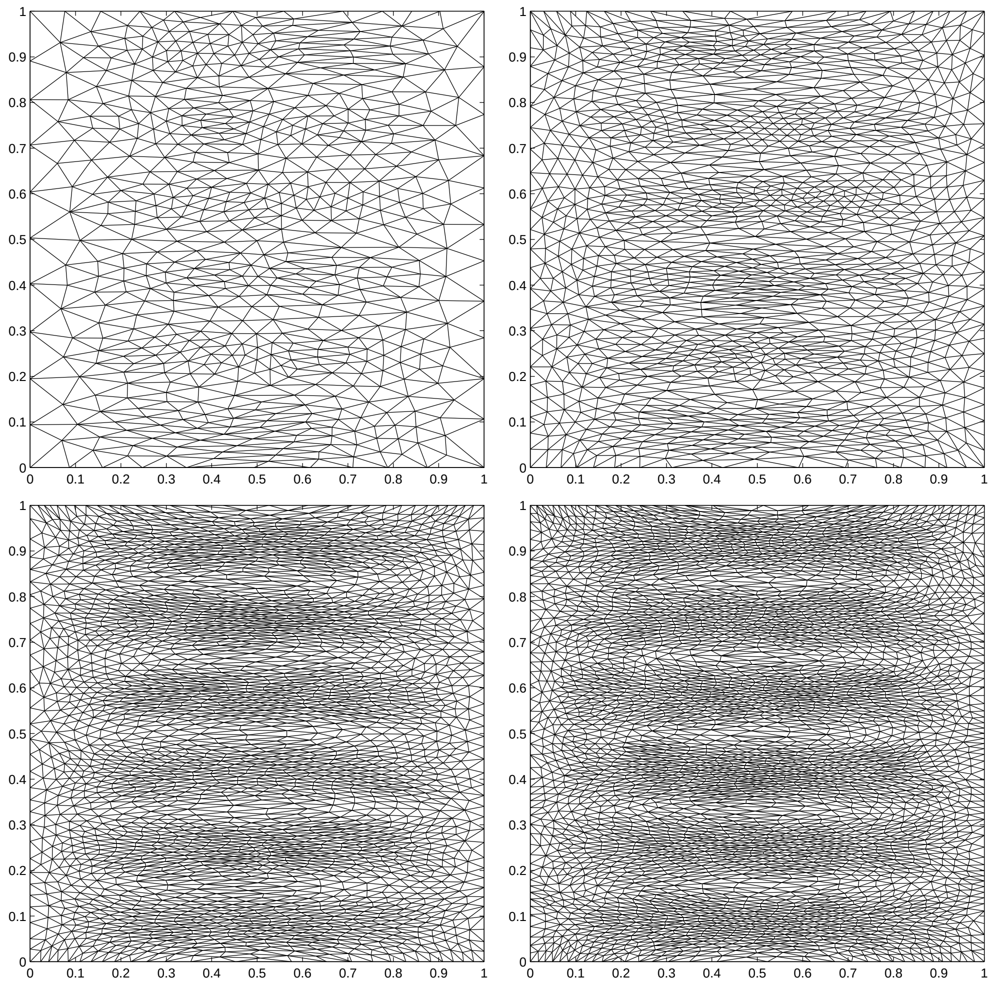

Figure 6: Second test case: sequence of the adapted meshes provided by the anisotropic ZZ error estimator (26) (left-right top-bottom).

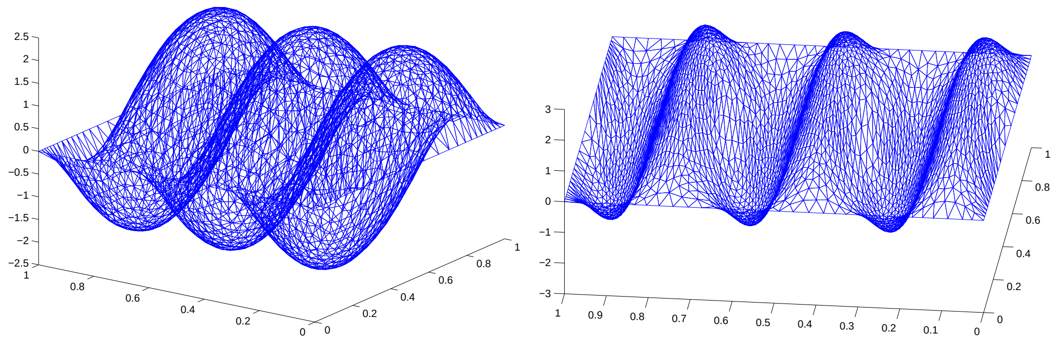

Figure 7: Second test case: discrete solution computed on the third adapted mesh.

Table 5: Second test case: values associated with the initial and the four adapted meshes

| $\max_K s_K - \min_K s_K$ | Nv | Nt | $\nu$ | $\|\|\|e_h\|\|\|$ | $\theta^{ZZ}$ | $\theta^{\rm res}$ | $\theta^A$ |
|---|---|---|---|---|---|---|---|
| $2.32 - 1.01$ | 665 | 1248 | 0.7083 | 5.3290 | 1.0504 | 7.69907 | 4.8269 |
| $15.6 - 1.03$ | 586 | 1128 | 0.6175 | 3.5560 | 1.0232 | 13.4678 | 4.9417 |
| $16.0 - 1.01$ | 1196 | 2307 | 0.4142 | 2.0337 | 1.0076 | 19.2115 | 4.9277 |
| $18.4 - 1.02$ | 2330 | 4533 | 0.3678 | 1.3077 | 1.0004 | 23.5636 | 4.9628 |
| $21.4 - 1.02$ | 3298 | 6416 | 0.3390 | 1.0928 | 0.9991 | 22.8249 | 4.9007 |

### 7.2.3 The third test case

As last test case we still deal with the solution of the Poisson problem in $\Omega = (0,1)^2$ completed with homogeneous Dirichlet boundary conditions. Now the forcing term $f$ is chosen such that the solution is

$$u = \sin(a\pi x_1)\,\sin(a\pi x_2), \tag{74}$$

with $a$ chosen equal to $1, 2$ and $4$. According to the criterion 1. in Section 7.1, we exploit such a test case to compare again the robustness of the anisotropic error estimator (26) with that of the standard ZZ and of the residual-based estimators. We use structured grids obtained by subdividing the horizontal and the vertical sides of the domain by `N1` and `N2` uniform subintervals, respectively.

Table 6: Third test case: values associated with structured meshes and for the choice $a = 1$ in (74)

| N1 − N2 | Nv | $\nu$ | $\|\|\|e_h\|\|\|$ | $\theta^{ZZ}$ | $\theta^{\rm res}$ | $\theta^A$ |
|---|---|---|---|---|---|---|
| $2 - 2$ | 9 | 1.0785 | 1.5352 | 0.6178 | 6.4211 | 2.5899 |
| $4 - 4$ | 25 | 0.8769 | 0.8427 | 0.9380 | 6.8819 | 3.8000 |
| $8 - 8$ | 81 | 0.5940 | 0.4323 | 1.0064 | 7.1731 | 4.2097 |
| $16 - 16$ | 289 | 0.3718 | 0.2176 | 1.0087 | 7.2769 | 4.3122 |
| $32 - 32$ | 1089 | 0.2338 | 0.1090 | 1.0048 | 7.3127 | 4.3374 |
| $64 - 64$ | 4225 | 0.1521 | 0.0545 | 1.0024 | 7.3260 | 4.3486 |

Table 7: Third test case: values associated with structured meshes and for the choice $a = 2$ in (74)

| N1 − N2 | Nv | $\nu$ | $\|\|\|e_h\|\|\|$ | $\theta^{ZZ}$ | $\theta^{\mathrm{res}}$ | $\theta^{A}$ |
|---|---|---|---|---|---|---|
| 2 − 2 | 9 | 1.0582 | 4.6980 | 0.3748 | 10.0847 | 2.5544 |
| 4 − 4 | 25 | 1.1147 | 3.0140 | 0.6251 | 6.4844 | 2.8587 |
| 8 − 8 | 81 | 0.9135 | 1.6807 | 0.9526 | 6.9236 | 4.0188 |
| 16 − 16 | 289 | 0.5767 | 0.8641 | 1.0070 | 7.2034 | 4.2996 |
| 32 − 32 | 1089 | 0.3342 | 0.4351 | 1.0061 | 7.2935 | 4.3525 |
| 64 − 64 | 4225 | 0.1955 | 0.2180 | 1.0029 | 7.3212 | 4.3564 |

Table 8: Third test case: values associated with structured meshes and for the choice $a = 4$ in (74)

| N1 − N2 | Nv | $\nu$ | $\|\|\|e_h\|\|\|$ | $\theta^{ZZ}$ | $\theta^{\mathrm{res}}$ | $\theta^{A}$ |
|---|---|---|---|---|---|---|
| 2 − 2 | 9 | 0.4811 | 2.9609 | 0.7136 | 7.8888 | 3.1110 |
| 4 − 4 | 25 | 1.1058 | 9.0507 | 0.5894 | 10.8892 | 3.5873 |
| 8 − 8 | 81 | 1.1328 | 6.0441 | 0.6266 | 6.5104 | 2.9673 |
| 16 − 16 | 289 | 0.9326 | 3.3556 | 0.9585 | 6.9436 | 4.1318 |
| 32 − 32 | 1089 | 0.5663 | 1.7274 | 1.0064 | 7.2183 | 4.3449 |
| 64 − 64 | 4225 | 0.3126 | 0.8702 | 1.0044 | 7.3018 | 4.3726 |

The results for the three different values of $a$ are collected in Tables 6-8. Besides a reduction of the norm $|||e_h|||$ of the discretization error, we notice that the values of the three effectivity indexes $\theta^{ZZ}$, $\theta^{\mathrm{res}}$ and $\theta^{A}$ stabilize around 1, 7.3 and 4.3, respectively. This trend states that, for this test case, except for a suitable scaling factor, the three error estimators provide reliable values for the energy norm of the discretization error, in the presence of the structured meshes chosen above. Finally, we note that, in all the three cases, the quantity $\nu$ is small, except for few cases, probably because of the regularity of the meshes and the smoothness of the solutions.

**Acknowledgments**
This work has been supported by the project MIUR 2001 “Numerical Methods in Fluid Dynamics and Electromagnetism”.

# References

[1] M. Ainsworth, J.T. Oden, A Posteriori Error Estimation in Finite Element Analysis (John Wiley & Sons, Inc., New-York, 2000).

[2] T. Apel, Anisotropic Finite Elements: Local Estimates and Applications, Book Series: Advances in Numerical Mathematics (Teubner, Stuttgart, 1999).

[3] S. Bartels, C. Carstensen, Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids. II: High order FEM, Math. Comp. 71 (2001) 971–994.

[4] R. Becker, R. Rannacher, An optimal control approach to a posteriori error estimation in finite element methods, Acta Numerica 10 (2001) 1–102.

[5] C. Carstensen, S. Bartels, Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids. I: Low order conforming, nonconforming, and mixed FEM, Math. Comp. 71 (2001) 945–969.

[6] C. Carstensen, S.A. Funken, Averaging technique for FE - a posteriori error control in elasticity. I: Conforming FEM, Comput. Methods Appl. Mech. Engrg. 190 (2001) 2483–2498.

[7] Ph. Ciarlet, The Finite Element Method for Elliptic Problems (North-Holland Publishing Company, Amsterdam, 1978).

[8] Ph. Clément, Approximation by finite element functions using local regularization, RAIRO Anal. Numér. 2 (1975) 77–84.

[9] E.F. D'Azevedo, R.B. Simpson, On optimal triangular meshes for minimizing the gradient error, Numer. Math. 59 (1991) 321–348.

[10] R. Duran, M.A. Muschietti, R. Rodríguez, On the asymptotic exactness of error estimators for linear triangular finite elements, Numer. Math. 59 (1991) 107–127.

[11] L. Formaggia, S. Perotto, New anisotropic a priori error estimates, Numer. Math. 89 (2001) 641–667.

[12] L. Formaggia, S. Perotto, Anisotropic error estimates for elliptic problems, Numer. Math. 94 (2003) 67–92.

[13] L. Formaggia, S. Micheletti, S. Perotto, Anisotropic mesh adaptation in Computational Fluid Dynamics: application to the advection-diffusion-reaction and the Stokes problems, to appear in Appl. Numer. Math.

[14] L. Formaggia, S. Perotto, P. Zunino, An anisotropic a-posteriori error estimate for a convection-diffusion problem, Comput. Visual. Sci. 4 (2001) 285–298.

[15] P.L. George, H. Borouchaki, Delaunay Triangulation and Meshing-Application to Finite Element (Editions Hermes, Paris, 1998).

[16] M.B. Giles, E. Süli, Adjoint methods for PDEs: a posteriori error analysis and postprocessing by duality, Acta Numerica 11 (2002) 145–236.

[17] G.H. Golub, C.F. Van Loan, Matrix Computations, 2nd ed. (The Johns Hopkins University Press, Baltimore, 1989).

[18] W.G. Habashi, M. Fortin, J. Dompierre, M.G. Vallet, Y. Bourgault, Anisotropic mesh adaptation: a step towards a mesh-independent and user-independent CFD, in Barriers and Challenges in Computational Fluid Dynamics, (Kluwer Acad. Publ., 1998) 99–117.

[19] F. Hecht, BAMG: bidimensional anisotropic mesh generator, *http://www-rocq.inria.fr/gamma/cdrom/www/bamg/eng.htm*, 1998.

[20] M. Krízek, P. Neittaanmäki, Superconvergence phenomenon in the finite element method arising from averaging gradients, Numer. Math. 45 (1984) 105–116.

[21] M. Krízek, P. Neittaanmäki, On a global superconvergence of the gradient of linear triangular elements, J. Comput. Appl. Math. 18 (1987) 221–233.

[22] G. Kunert, A Posteriori Error Estimation for Anisotropic Tetrahedral and Triangular Finite Element Meshes, Ph.D. thesis, Fakultät für Mathematik der Technischen Universität Chemnitz, 1999.

[23] G. Kunert, S. Nicaise, Zienkiewicz-Zhu error estimators on anisotropic tetrahedral and triangular finite element meshes, Preprint SFB393/01-20, Technishe Universität Chemnitz, 2001.

[24] A.M. Lakhany, I. Marek, J.R. Whiteman, Superconvergence results on mildly structured triangulations, Comput. Methods Appl. Mech. Engrg. 189 (2000) 1–75.

[25] J.L. Lions, E. Magenes, Non-Homogeneous Boundary Value Problem and Application, Volume I (Springer-Verlag, Berlin, 1972).

[26] S. Micheletti, S. Perotto, M. Picasso, Stabilized finite elements on anisotropic meshes: a priori error estimates for the advection-diffusion and Stokes problems, SIAM J. Numer. Anal. 41 (3) (2003) 1131–1162.

[27] S. Micheletti, S. Perotto, An anisotropic recovery-based a posteriori error estimator, Numerical Mathematics and Advanced Applications - Enumath 2001, Proceedings of the 4th European Conference on Numerical Mathematics and Advanced Applications. F. Brezzi, A. Buffa, S. Corsaro, A. Murli, ed., (Springer-Verlag, Italia, 2003) 731-741.

[28] S. Micheletti, S. Perotto, Anisotropic mesh adaption in CFD, MOX–Report No. 29, MOX – Department of Mathematics "F. Brioschi", Politecnico of Milano, submitted for the publication in Proceedings of Chicago Workshop on Adaptive Mesh Refinement Methods, Chicago, September 3-5, 2003.

[29] P. Morin, R.H. Nochetto, K.G. Siebert, Data oscillation and convergence of adaptive FEM, SIAM J. Numer. Anal. 38 (2) (2000) 466–488.

[30] J.T. Oden, S. Prudhomme, Goal-oriented error estimation and adaptivity for the finite element method, Computers Math. Applic. 41 (5-6) (2001) 735–756.

[31] M. Picasso, Numerical study of the effectivity index for an anisotropic error indicator based on Zienkiewicz-Zhu error estimator, Comm. Numer. Methods. Engrg. 19 (1) (2003) 13–23.

[32] M. Picasso, An anisotropic error indicator based on Zienkiewicz-Zhu error estimator: application to elliptic and parabolic problems, SIAM J. Sci. Comput. 24 (4) (2003) 1328–1355.

[33] A. Quarteroni, A. Valli, Numerical approximation of partial differential equations (Springer-Verlag, Berlin, 1994).

[34] L.R. Scott, S. Zhang, Finite element interpolation of non-smooth functions satisfying boundary conditions, Math. Comp. 54 (1990) 483–493.

[35] R.B. Simpson, Anisotropic mesh transformations and optimal error control, Appl. Numer. Math. 14 (1994) 183–198.

[36] R. Verfürth, A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques, Wiley - Teubner, New York, 1996.

[37] N. Yan, A. Zhou, Gradient recovery type a posteriori error estimation for finite element approximations on irregular meshes, Comput. Methods Appl. Mech. Engrg. 190 (2001) 4289–4299.

[38] Z. Zhang, J.Z. Zhu, Superconvergence of the derivative patch recovery technique and *a posteriori* error estimation, in the IMA Volumes in Mathematics and its Applications, I. Babuska, J.E. Flaherty, W.D. Henshaw, J.E. Hopcroft, J.E. Oliger, T. Tezduyar, ed., Modeling, Mesh Generation, and Adaptive Numerical Methods for Partial Differential Equations, Vol. 75 (Springer-Verlag, New York, 1995) 431–450.

[39] O.C. Zienkiewicz, Y.K. Cheung, The Finite Element Method in Structural and Continuum Mechanics. Numerical Solution of problems in structural and continuum mechanics (McGraw Hill, London, 1967).

[40] O.C. Zienkiewicz, J.Z. Zhu, A simple error estimator and adaptive procedure for practical engineering analysis, Int. J. Numer. Methods Eng. 24 (2) (1987) 337–357.

[41] O.C. Zienkiewicz, J.Z. Zhu, The superconvergent patch recovery (SPR) and adaptive finite element refinement, Comput. Methods Appl. Mech. Engrg. 101 (1-3) (1992) 207–224.

[42] O.C. Zienkiewicz, J.Z. Zhu, The superconvergent patch recovery and a posteriori error estimates. I: The recovery technique, Int. J. Numer. Methods Eng. 33 (7) (1992) 1331–1364.

[43] O.C. Zienkiewicz, J.Z. Zhu, The superconvergent patch recovery and a posteriori error estimates. II: Error estimates and adaptivity, Int. J. Numer. Methods Eng. 33 (7) (1992) 1365–1382.