

## A POSTERIORI ERROR ESTIMATES FOR A DISCONTINUOUS GALERKIN APPROXIMATION OF SECOND-ORDER ELLIPTIC PROBLEMS\*

OHANNES A. KARAKASHIAN<sup>†</sup> AND FREDERIC PASCAL<sup>‡</sup>

**Abstract.** Several a posteriori error estimators are introduced and analyzed for a discontinuous Galerkin formulation of a model second-order elliptic problem. In addition to residual-type estimators, we introduce some estimators that are couched in the ideas and techniques of domain decomposition. Results of numerical experiments are presented.

**Key words.** discontinuous Galerkin methods, a posteriori estimates

**AMS subject classifications.** 65N55, 65F10

**DOI.** 10.1137/S0036142902405217

**1. Introduction.** One of the important objectives of the numerical approximation of differential equations has been to obtain approximations whose error, as measured in some norm, falls in a given range, preferably as narrow as possible. In the finite element method, and specifically for elliptic boundary value problems, such a goal became possible with the advent of a posteriori estimates pioneered by Babuška and Rheinboldt [4, 5]. Acting on an approximation  $u_h$  calculated on a given mesh, such a posteriori estimates give lower and upper bounds on the error expressed in terms of contributions from individual triangles and interfaces. This makes it possible to calculate a new mesh by means of refinement and coarsening. For a survey of the vast amount of work spurred by the above two references, we refer the reader to the book by Verfürth [18]. More recently, attention has increasingly focused on the important issue of convergence, whereby a given tolerance is achieved after a finite number of refinement steps; cf., e.g., [10, 17, 15].

Our aim is to present a posteriori error estimates in the energy norm for a discontinuous Galerkin formulation of a simple second-order elliptic problem. In contrast to standard Galerkin methods, such work is still very rare. Indeed, we are aware only of [8] as taking the a posteriori approach; also, only the estimator (3.1) is treated, and with a different proof which relies on a Helmholtz-type decomposition of the gradient of the error, thus following a technique first used in the context of a posteriori estimates for nonconforming methods. See also [16] for a posteriori estimates in the  $L^2$  norm. Recall that in discontinuous Galerkin methods the trial and test spaces consist of piecewise totally discontinuous polynomials. That is, no continuity constraints are explicitly imposed on the trial and test functions across the element interfaces. As a consequence, weak formulations must include jump terms across interfaces, and typically penalty terms are (artificially) added to control the jump terms. Several variants of this approach exist; cf., e.g., [2, 9, 19]. For a nice survey of various discontinuous Galerkin methods, see [3].

---

\*Received by the editors April 8, 2002; accepted for publication (in revised form) April 3, 2003; published electronically December 17, 2003.

<http://www.siam.org/journals/sinum/41-6/40521.html>

<sup>†</sup>Department of Mathematics, The University of Tennessee, Knoxville, TN 37996 (ohannes@math.utk.edu). The research of this author was supported by the French Ministère de la Recherche for a “séjour scientifique de haut niveau” at the University of Paris-Sud.

<sup>‡</sup>Laboratoire de Mathématiques, Université de Paris-Sud, and Centre National de la Recherche Scientifique 91405, Orsay, France (Frederic.Pascal@math.u-psud.fr).

Discontinuous Galerkin methods have several advantages over other types of finite element methods. For example, the trial and test spaces are very easy to construct; they can naturally handle inhomogeneous boundary conditions and curved boundaries; and they allow the use of highly nonuniform and unstructured meshes. In addition, the fact that the mass matrices are block diagonal is an attractive feature in the context of time-dependent problems, especially if explicit time discretizations are used.

In this paper, we will concentrate on the construction and analysis of error estimators, postponing to a subsequent work the study of other important issues such as convergence of the adaptive scheme. In section 3 we present residual-type estimators whose form and analysis follow traditional lines, with the exception of some technical issues caused by the discontinuous nature of the finite element spaces.

In section 4 we present estimators requiring the solution of local problems. In a departure from more traditional techniques, ours flow from the ideas and techniques of domain decomposition, and specifically those expounded in [11]. In a nutshell, we view the computed solution  $u_h$  corresponding to a mesh  $\mathcal{T}_h$  as a “coarse-mesh” approximation to a more accurate approximation  $u'_h$  to  $u$ , with an eye towards using  $u'_h - u_h$  to estimate  $u - u_h$ . Obviously computing  $u'_h$  would prove too costly; instead, a good approximation thereof is obtained by adding to  $u_h$  the solutions of local problems, the supports of these local contributions playing the role of the subdomains. Indeed, our technique offers the tightest coupling yet known between a posteriori error estimation and domain decomposition, to the extent that the matrices involved in the solution of the local problems consist of the diagonal blocks of the global stiffness matrix that correspond to the individual triangles. A somewhat similar idea is found in [20] in the context of mortar finite elements. There are, however, substantial differences between the two approaches.

In section 5, we present results of numerical experiments focusing on the behavior of the effectivity indices as well as other characteristics of the various estimators.

**2. Preliminaries.** Let  $\Omega \subset \mathbf{R}^d$ ,  $d = 1, 2, 3$ , be a bounded domain. We consider the following model problem:

$$(2.1) \quad -\Delta u = f \quad \text{in } \Omega,$$

$$(2.2) \quad u = 0 \quad \text{on } \partial\Omega.$$

The treatment of second-order elliptic problems with more general coefficients and boundary conditions will be contained in a parallel work [13].

Throughout this paper, the standard space, norm, and inner product notation are adopted. Their definitions can be found in [1]. Also,  $c$  is used to denote a generic positive mesh-independent constant.

The discontinuous Galerkin method considered in this paper for discretizing problem (2.1)–(2.2) is the one proposed in [6] and [7, 12], where the biharmonic and Stokes problems, respectively, were considered.

Let  $\mathcal{T}_h = \{K_i : i = 1, 2, \dots, m_h\}$  be a family of star-like partitions (triangulations) of the domain  $\Omega$  parametrized by  $0 < h \leq 1$ . We assume the following:

- (i) The elements of  $\mathcal{T}_h$  satisfy the minimal angle condition. Specifically, there is a constant  $\theta_0 > 0$  such that  $h_K/\rho_K \geq \theta_0 \forall K \in \mathcal{T}_h$ , where  $h_K$  and  $\rho_K$  denote, respectively, the diameters of the circumscribed and inscribed balls to  $K$ .
- (ii)  $\mathcal{T}_h$  is locally quasi-uniform; that is, if two elements  $K_j$  and  $K_\ell$  are adjacent in the sense that  $\mu_{d-1}(\partial K_j \cap \partial K_\ell) > 0$ , then  $\text{diam}(K_j) \approx \text{diam}(K_\ell)$ .

Here  $\mu_{d-1}$  denotes the  $(d - 1)$ -dimensional Lebesgue measure. On  $\mathcal{T}_h$  we define the “energy space”  $E_h = \Pi_{K \in \mathcal{T}_h} H^2(K) \subset L^2(\Omega)$ . For  $r \geq 2$ , we define the finite element space  $V_h^r \subset E_h$  by  $V_h^r = \Pi_{K \in \mathcal{T}_h} P_{r-1}(K)$ , where  $P_{r-1}(K)$  denotes the space of polynomials of total degree  $r - 1$ .

Given the discontinuous nature of the piecewise polynomial functions, we define  $\mathcal{E}^I$  and  $\mathcal{E}^B$  to be the set of all interior and boundary edges (faces in the case  $d = 3$ ), respectively:

$$\begin{aligned} \mathcal{E}^I &= \{e = \partial K_j \cap \partial K_\ell, \quad \mu_{d-1}(\partial K_j \cap \partial K_\ell) > 0\}, \\ \mathcal{E}^B &= \{e = \partial K \cap \partial \Omega, \quad \mu_{d-1}(\partial K \cap \partial \Omega) > 0\}. \end{aligned}$$

We also set  $\mathcal{E} = \mathcal{E}^I \cup \mathcal{E}^B$ . We note that elements of  $\mathcal{E}^B$  may be curved. Also, if  $e \in \mathcal{E}^I$ , then  $e = \partial K^+ \cap \partial K^-$  for  $K^+, K^- \in \mathcal{T}_h$ . We may designate as  $K^+$  the triangle with the higher of the two indices.

Note that elements of the energy space  $E_h$  are not functions in the proper sense, and care must be applied in defining their values on  $\mathcal{E}$ . This is done in the sense of trace.

In order to construct a weak formulation for the problem (2.1)–(2.2), we introduce the bilinear form  $a_h^\gamma : E_h \times E_h \rightarrow \mathbf{R}$ :

$$\begin{aligned} (2.3) \quad a_h^\gamma(u, v) &= \sum_{K \in \mathcal{T}_h} (\nabla u, \nabla v)_K - \sum_{e \in \mathcal{E}^I} \left[ \langle \{\partial_n u\}, [v] \rangle_e + \langle \{\partial_n v\}, [u] \rangle_e - \gamma h_e^{-1} \langle [u], [v] \rangle_e \right] \\ &\quad - \sum_{e \in \mathcal{E}^B} \left[ \langle \partial_n u, v \rangle_e + \langle \partial_n v, u \rangle_e - \gamma h_e^{-1} \langle u, v \rangle_e \right], \end{aligned}$$

where  $h_e = \text{diam}(e)$  and

$$\begin{aligned} (u, v)_D &= \int_D u \cdot v \, dx, \\ \langle u, v \rangle_\Gamma &= \int_\Gamma uv \, ds, \quad \text{edge/surface integrals,} \quad |v|_\Gamma = \langle v, v \rangle_\Gamma^{1/2}, \\ [v]|_e &= v^+|_e - v^-|_e, \quad v^+ = v|_{K^+}, \quad v^- = v|_{K^-}, \quad e \in \mathcal{E}^I, \\ \{\partial_n v\}|_e &= \frac{\partial v^+}{\partial n^+} \Big|_e, \quad [\partial_n v]|_e = \frac{\partial v^+}{\partial n^+} \Big|_e - \frac{\partial v^-}{\partial n^+} \Big|_e, \quad e \in \mathcal{E}^I, \\ \partial_n v|_e &= \frac{\partial v^+}{\partial n^+} \Big|_e, \quad e \in \mathcal{E}^B. \end{aligned}$$

Some further comments on the nature of the form  $a_h^\gamma$  are in order:

- (a) The third and sixth terms have been added to symmetrize  $a_h^\gamma$ . Note that the former is zero for smooth  $u$ , while the latter is a known quantity since  $u|_{\partial \Omega}$  is given. The a priori estimates remain valid if these terms are removed.
- (b)  $\gamma$  is a positive (penalty) parameter that must be chosen appropriately in order for  $a_h^\gamma$  to be coercive.

*Remark 2.1.* There is an alternative formulation due to Arnold [2], which is obtained by setting  $\{\partial_n v\}|_e = \frac{1}{2}(\frac{\partial v^+}{\partial n^+} + \frac{\partial v^-}{\partial n^+})|_e$ . The results presented in this paper should be valid for Arnold’s formulation as well.

The form  $a_h^\gamma(\cdot, \cdot)$  is consistent with the Laplacian in the sense that if  $u \in H^2(\Omega)$ , then

$$(2.4) \quad a_h^\gamma(u, v) = -(\Delta u, v) - \sum_{e \in \mathcal{E}^B} \langle u, \partial_n v - \gamma h_e^{-1} v \rangle_e \quad \forall v \in E_h.$$

Thus, we define the discontinuous Galerkin approximation of  $u$  to be the element  $u_h^\gamma$  in  $V_h^r$  that satisfies

$$(2.5) \quad a_h^\gamma(u_h^\gamma, v) = F(v) := (f, v) \quad \forall v \in V_h^r.$$

The existence of a unique  $u_h^\gamma$  follows from Lemma 2.1.

We define the “energy” norm on  $E_h$  by

$$\|v\|_{1,h} = \left( \sum_{K \in \mathcal{T}_h} \|\nabla v\|_K^2 + \sum_{e \in \mathcal{E}^I} \left[ h_e |\{\partial_n v\}|_e^2 + h_e^{-1} |[v]|_e^2 \right] + \sum_{e \in \mathcal{E}^B} \left[ h_e |\partial_n v|_e^2 + h_e^{-1} |v|_e^2 \right] \right)^{1/2}.$$

Concerning the continuity and coercivity of the form  $a_h^\gamma$ , we have the following result (cf. [7]).

LEMMA 2.1. (i)

$$(2.6) \quad |a_h^\gamma(u, v)| \leq (1 + \gamma) \|u\|_{1,h} \|v\|_{1,h} \quad \forall u, v \in E_h.$$

(ii) *There exist positive constants  $\gamma_0$  and  $c_a$  such that for  $\gamma \geq \gamma_0$*

$$(2.7) \quad a_h^\gamma(v, v) \geq c_a \|v\|_{1,h}^2 \quad \forall v \in V_h^r.$$

Here  $\gamma_0$  depends only on  $r$  and the (aspect) ratios  $h_K/\rho_K$  of the elements. In view of condition (i) on the mesh,  $\gamma_0$  can grow only as a function of  $r$ . Numerical experiments reveal that  $\gamma_0 \approx 5$  for  $r = 2$  and  $\gamma_0 \approx 15$  for  $r = 3$ .

The proofs of the above rely on the following *trace* and *inverse* inequalities. Let  $D$  be a regular and starlike domain, and let  $\mu = \text{diam}(D)$ . Then

$$(2.8) \quad |v|_{\partial D}^2 \leq c_{tr} (\mu^{-1} \|v\|_D^2 + \mu \|\nabla v\|_D^2) \quad \forall v \in H^1(D).$$

Let  $|\cdot|_{j,D}$  denote the seminorm of  $H^j(D)$ . Then

$$(2.9) \quad |v|_{j,D} \leq c_{inv} \mu^{i-j} |v|_{i,D} \quad \forall v \in P_r, \quad 0 \leq i \leq j \leq 2,$$

the constant  $c_{inv}$  in (2.9) depending only on  $r$ .

We shall assume that the following approximation property holds: Let  $0 \leq m \leq r$ . Then there exists a constant  $c > 0$ , independent of  $\mathcal{T}_h$ , such that for any  $u \in H^m(\Omega)$  and  $K \in \mathcal{T}_h$  there exists  $\chi \in P_{r-1}(K)$  satisfying

$$(2.10) \quad |u - \chi|_{j,K} \leq ch_K^{m-j} |u|_{m,K}, \quad 0 \leq j \leq m.$$

It can be shown that the following error estimates hold (cf. [7]).

THEOREM 2.1. *Let  $u$  and  $u_h^\gamma$  be the solutions of (2.1)–(2.2) and (2.5), respectively, and suppose that  $u \in H^r(\Omega) \cap H_0^1(\Omega)$  with  $r \geq 2$ . Then there exists a positive constant  $c$ , which is independent of  $h$  and  $u$ , such that*

$$(2.11) \quad \|u - u_h^\gamma\|_{1,h} \leq c \left( \sum_{K \in \mathcal{T}_h} h_K^{2(r-1)} |u|_{r,K}^2 \right)^{1/2},$$

$$(2.12) \quad \|u - u_h^\gamma\| \leq ch^r |u|_{r,\Omega}.$$

**2.1. An approximation result.** For our first residual-type a posteriori estimate we will need to see how well an element of  $V_h^r$  can be approximated by elements of  $V_h^0 = V_h^r \cap H_0^1(\Omega)$ . The result, Theorem 2.2 below, can also be found in [14]. The proof we give here is constructive and differs entirely from the one given in [14]. We also consider separately in Theorem 2.3 the case when the mesh is nonconforming, i.e., is characterized by the presence of *hanging nodes*.

LEMMA 2.2. *Given  $N$  real numbers  $\{\alpha_1, \dots, \alpha_N\}$  let  $\beta = \frac{1}{N} \sum_{j=1}^N \alpha_j$ . Then,*

$$(2.13) \quad \sum_{j=1}^N |\alpha_j - \beta|^2 \leq C \sum_{j=1}^{N-1} |\alpha_{j+1} - \alpha_j|^2,$$

where  $C$  depends only on  $N$ .

*Proof.* For any  $j$ , the Cauchy–Schwarz inequality gives

$$|\alpha_j - \beta|^2 = \frac{1}{N^2} \left| \sum_{i=1}^N (\alpha_j - \alpha_i) \right|^2 \leq \frac{N-1}{N^2} \sum_{i=1}^N |\alpha_j - \alpha_i|^2.$$

Summing over  $j$ , we obtain  $\sum_{j=1}^N |\alpha_j - \beta|^2 \leq \frac{2(N-1)}{N} \sum_{j>i} |\alpha_j - \alpha_i|^2$ . The required result now follows, upon writing  $\alpha_j - \alpha_i = \sum_{k=i}^{j-1} (\alpha_{k+1} - \alpha_k)$  and using the arithmetic-geometric mean inequality.  $\square$

THEOREM 2.2. *Let  $\mathcal{T}_h$  be a conforming mesh consisting of triangles when  $d = 2$ , and tetrahedra when  $d = 3$ . Then for any  $v_h \in V_h^r$  there exists  $\chi \in V_h^0$  satisfying*

$$(2.14) \quad \sum_{K \in \mathcal{T}_h} \|\nabla(v_h - \chi)\|_K^2 \leq C \left( \sum_{e \in \mathcal{E}^I} h_e^{-1} |[v_h]|_e^2 + \sum_{e \in \mathcal{E}^B} h_e^{-1} |v_h|_e^2 \right)$$

for some constant  $C$  independent of  $h$  and  $v_h$  but which may depend on the constant  $\theta_0$  in assumption (i) on the mesh.

*Proof.* The main argument is quite natural. Given  $v_h \in V_h^r$ , we construct a function  $\chi \in V_h^0$  as follows: At every node of the mesh  $\mathcal{T}_h$  corresponding to a Lagrangian-type degree of freedom for  $V_h^r$ , the value of  $\chi$  is set to the average of the values of  $v_h$  at that node.

For each  $K \in \mathcal{T}_h$  let  $\mathcal{N}_K = \{x_K^{(j)}, j = 1, \dots, m\}$  be the Lagrange nodes (points) of  $K$  and  $\{\phi_K^{(j)}, j = 1, \dots, m\}$  the corresponding (local) basis functions satisfying  $\phi_K^{(j)}(x_K^{(i)}) = \delta_{ij}$ . Set  $\mathcal{N} = \cup_{K \in \mathcal{T}_h} \mathcal{N}_K$ . We view  $\mathcal{N}$  as the union of three disjoint classes:

$$\begin{aligned} \mathcal{N}_i &= \{\nu \in \mathcal{N} : \nu \text{ is interior to some element}\}, \\ \mathcal{N}_b &= \{\nu \in \mathcal{N} : \nu \in e \in \mathcal{E}^B\}, \\ \mathcal{N}_v &= \mathcal{N} \setminus (\mathcal{N}_i \cup \mathcal{N}_b). \end{aligned}$$

For each  $\nu \in \mathcal{N}$ , let  $\omega_\nu = \{K \in \mathcal{T}_h | \nu \in K\}$  and denote its cardinality by  $|\omega_\nu|$ . If  $\nu \in \mathcal{N}_i$ , then  $|\omega_\nu| = 1$ . On the other hand if  $\nu \in \mathcal{N}_b \cup \mathcal{N}_v$ , then  $|\omega_\nu|$  is bounded by a constant depending only on the constant  $\theta_0$ .

Now let  $\overset{0}{\mathcal{N}}$  be the collection of distinct Lagrange nodes  $\nu$  needed to construct a function  $\chi \in V_h^r$ . To each node  $\nu \in \overset{0}{\mathcal{N}}$  we associate the basis function  $\phi^{(\nu)}$  given by

$$\text{supp } \phi^{(\nu)} = \bigcup_{K \in \omega_\nu} K, \quad \phi^{(\nu)}|_K = \phi_K^{(j)}, \quad x_K^{(j)} = \nu.$$

We make it a point to include the boundary nodes  $\mathcal{N}_b$  in  $\overset{0}{\mathcal{N}}$  even though this is not necessary in view of the vanishing on  $\partial\Omega$  of the functions in  $V_h^r$ . We then can state the following characterization:  $\overset{0}{\mathcal{N}} \subseteq \mathcal{N}$  and the mesh  $\mathcal{T}_h$  is conforming if and only if  $\overset{0}{\mathcal{N}} = \mathcal{N}$ .

Now, given  $v_h \in V_h^r$ , written  $v_h = \sum_{K \in \mathcal{T}_h} \sum_{j=1}^m \alpha_K^{(j)} \phi_K^{(j)}$ , we define the function  $\chi \in V_h^r$  by (note that  $\overset{0}{\mathcal{N}} = \mathcal{N}$  since the mesh is conforming)

$$(2.15) \quad \chi = \sum_{\nu \in \overset{0}{\mathcal{N}}} \beta^{(\nu)} \phi^{(\nu)}, \quad \text{where } \beta^{(\nu)} = \begin{cases} 0 & \text{if } \nu \in \mathcal{N}_b, \\ \frac{1}{|\omega_\nu|} \sum_{x_K^{(j)} = \nu} \alpha_K^{(j)}, & \text{if } \nu \in \overset{0}{\mathcal{N}} \setminus \mathcal{N}_b. \end{cases}$$

Now set  $\beta_K^{(j)} = \beta^{(\nu)}$  whenever  $x_K^{(j)} = \nu$ .

A simple scaling argument shows that  $\|\nabla \phi_K^{(j)}\|_K^2 \leq ch_K^{d-2}$ . Hence

$$(2.16) \quad \begin{aligned} \sum_{K \in \mathcal{T}_h} \|\nabla(v_h - \chi)\|_K^2 &\leq cm \sum_{K \in \mathcal{T}_h} h_K^{d-2} \sum_{j=1}^m |\alpha_K^{(j)} - \beta_K^{(j)}|^2 \\ &\leq c \sum_{\nu \in \mathcal{N}} h_\nu^{d-2} \sum_{x_K^{(j)} = \nu} |\alpha_K^{(j)} - \beta^{(\nu)}|^2 \quad \left( h_\nu = \max_{K \in \omega_\nu} h_K \right) \\ &= c \sum_{\nu \in \mathcal{N}_v} h_\nu^{d-2} \sum_{x_K^{(j)} = \nu} |\alpha_K^{(j)} - \beta^{(\nu)}|^2 + c \sum_{\nu \in \mathcal{N}_b} h_\nu^{d-2} \sum_{x_K^{(j)} = \nu} |\alpha_K^{(j)}|^2. \end{aligned}$$

Note that there are no contributions from  $\mathcal{N}_i$ . We now temporarily focus on the case  $d = 2$ . For  $\nu \in \mathcal{N}_v$ , we enumerate the elements of  $\omega_\nu$  as  $\{K_1, \dots, K_{|\omega_\nu|}\}$  so that any consecutive pair  $K_i, K_{i+1}$  in that list share an edge. Then from Lemma 2.2, with some constant  $c$  depending only on  $|\omega_\nu|$  and thus on  $\theta_0$ , we have

$$(2.17) \quad \sum_{x_K^{(j)} = \nu} |\alpha_K^{(j)} - \beta^{(\nu)}|^2 \leq c \sum_{i=1}^{|\omega_\nu|-1} |\alpha_{K_i}^{(j_i)} - \alpha_{K_{i+1}}^{(j_{i+1})}|^2.$$

For  $d = 3$ , it may not be possible to enumerate  $\omega_\nu$  in such a way. However, by allowing some repetitions of its elements, we can write  $\omega_\nu = \{K_{\ell_1}, \dots, K_{\ell_{n(\nu)}}\}$  for some  $n(\nu)$ , so that in this case also  $K_{\ell_i}$  and  $K_{\ell_{i+1}}$  share a face or an edge. Having done so, by applying Lemma 2.2 to the list obtained by removing all repetitions of elements of  $\omega_\nu$  and then using the arithmetic-geometric mean inequality, we obtain

$$(2.18) \quad \sum_{x_K^{(j)} = \nu} |\alpha_K^{(j)} - \beta^{(\nu)}|^2 \leq c \sum_{i=1}^{n(\nu)-1} |\alpha_{K_{\ell_i}}^{(j_{\ell_i})} - \alpha_{K_{\ell_{i+1}}}^{(j_{\ell_{i+1}})}|^2.$$

Using (2.17) if  $d = 2$ , or (2.18) if  $d = 3$ , from (2.16) we have

$$(2.19) \quad \sum_{K \in \mathcal{T}_h} \|\nabla(v_h - \chi)\|_K^2 \leq c \sum_{e \in \mathcal{E}^I} \sum_{\nu \in e} h_\nu^{d-2} \left| \alpha_{K^+}^{(j_\nu^+)} - \alpha_{K^-}^{(j_\nu^-)} \right|^2 + c \sum_{\nu \in \mathcal{N}_b} \sum_{x_K^{(j)} = \nu} h_\nu^{d-2} \left| \alpha_K^{(j)} \right|^2,$$

with  $x_{K^+}^{(j_\nu^+)} = x_{K^-}^{(j_\nu^-)} = \nu$ . Note that  $\alpha_{K^+}^{(j_\nu^+)} - \alpha_{K^-}^{(j_\nu^-)}$  is the jump in the values of  $v_h$  at  $\nu$  across  $e$ . Also, since the mesh  $\mathcal{T}_h$  is locally quasi-uniform, it follows that

$$(2.20) \quad \sum_{\nu \in e} h_\nu^{d-2} \left| \alpha_{K^+}^{(j_\nu^+)} - \alpha_{K^-}^{(j_\nu^-)} \right|^2 \leq ch_e^{d-2} \|[v_h]\|_{L^\infty(e)}^2 \leq ch_e^{-1} \|[v_h]\|_e^2,$$

where the constant  $c$  depends on the number of nodes in  $e$ .

Similarly, it can be shown that for  $\nu \in e \in \mathcal{E}^B$ ,

$$(2.21) \quad \sum_{x_K^{(j)} = \nu} h_\nu^{d-2} \left| \alpha_K^{(j)} \right|^2 \leq ch_e^{-1} |v_h|_e^2.$$

The required result now follows from (2.19)–(2.21).  $\square$

We now consider the case when the mesh is nonconforming. We make the following observations, using the notation established in Theorem 2.2:

- (i) The hanging nodes are precisely the members of  $\mathcal{N} \setminus \mathcal{N}^0$ .
- (ii) A hanging node cannot be a member of  $\mathcal{N}_b$  or  $\mathcal{N}_i$ .
- (iii) For every hanging node  $\nu$  there is a nonempty proper subset  $\tilde{\omega}_\nu$  of  $\omega_\nu$  such that if  $K \in \tilde{\omega}_\nu$ , then  $\nu$  is not a local node of  $K$ . For  $d = 2$ ,  $|\tilde{\omega}_\nu| = 1$ .

We shall also require that  $\mathcal{T}_h$  be obtained from a conforming mesh  $\mathcal{T}_h^0$  via a finite number of refinement/coarsening steps. In particular, we assume that there is a mapping  $Level : \mathcal{T}_h \rightarrow \mathcal{N}$ , the set of nonnegative integers, such that

- (iv)  $Level(K) = 0 \forall K \in \mathcal{T}_h^0$  ( $\mathcal{T}_h^0 \subseteq \mathcal{T}_h$ ).
- (v) If  $K \in \omega_\nu \setminus \tilde{\omega}_\nu$  and  $\tilde{K} \in \tilde{\omega}_\nu$  are as in (iii) above, then  $Level(K) > Level(\tilde{K})$ .

An example of the mapping  $Level$  can be constructed for  $d = 2$  as follows. Suppose that we *refine* a given triangle (the *father*) by cutting it in the usual way (see, e.g., Figure 3.2) into four triangles of equal area (the *sons*). On the other hand, we *coarsen* the mesh by merging four sons of the same father. Then we define  $Level(K)$ ,  $K \in \mathcal{T}_h$ , by  $|K| = |K^0| (\frac{1}{4})^{Level(K)}$ , where  $K^0$  is the triangle in  $\mathcal{T}_h^0$  that contains  $K$  and where  $|\cdot|$  denotes area.

We have the following result.

**THEOREM 2.3.** *Let  $\mathcal{T}_h$  be a nonconforming mesh consisting of triangles when  $d = 2$  and tetrahedra when  $d = 3$ . We shall also assume that  $\mathcal{T}_h$  can be described in terms of the mapping  $Level$  as discussed above. Then (2.14) holds, but the constant  $C$  may also depend on  $L_{max} = \max\{Level(K), K \in \mathcal{T}_h\}$ .*

*Proof.* Noting that an element of  $V_h^0$  is still defined by its values at the nodes in  $\mathcal{N}^0$ , we define the approximant  $\chi \in V_h^0$  of  $v_h$  via (2.15). This uniquely determines the values of  $\chi$  at the hanging nodes, so we let  $\beta^{(\nu)} = \chi(\nu)$  for  $\nu \in \mathcal{N} \setminus \mathcal{N}^0$ . In a similar fashion, we introduce the quantities  $\tilde{\alpha}_{\tilde{K}}^{(\nu)} = (v_h|_{\tilde{K}})(\nu)$ ,  $\tilde{K} \in \tilde{\omega}_\nu$ ,  $\nu \in \mathcal{N} \setminus \mathcal{N}^0$ .

Proceeding as in (2.16), we obtain

$$\begin{aligned}
 \sum_{K \in \mathcal{T}_h} \|\nabla(v_h - \chi)\|_K^2 &\leq c \sum_{\ell=0}^{L_{max}} \sum_{K \in \mathcal{T}_h^\ell} h_K^{d-2} \sum_{j=1}^m |\alpha_K^{(j)} - \beta_K^{(j)}|^2 \\
 (2.22) \quad &\leq c \sum_{\nu \in \mathcal{N}^0} h_\nu^{d-2} \sum_{x_K^{(j)} = \nu} |\alpha_K^{(j)} - \beta^{(\nu)}|^2 + c \sum_{\nu \in \mathcal{N} \setminus \mathcal{N}^0} h_\nu^{d-2} \sum_{\substack{x_K^{(j)} = \nu \\ K \in \omega_\nu \setminus \tilde{\omega}_\nu}} |\alpha_K^{(j)} - \beta^{(\nu)}|^2,
 \end{aligned}$$

where  $\mathcal{T}_h^\ell = \{K \in \mathcal{T}_h \mid Level(K) = \ell\}$ . The first sum on the right-hand side of (2.22) can be handled as in the conforming case using steps (2.17)–(2.21). As for the second sum, for a given  $x_K^{(j)} = \nu$ ,  $K \in \omega_\nu \setminus \tilde{\omega}_\nu$ , we choose  $\tilde{K} \in \tilde{\omega}_\nu$  such that  $K$  and  $\tilde{K}$  share an edge or a face. A crucial observation is that  $Level(\tilde{K}) < Level(K)$ . Then

$$|\alpha_K^{(j)} - \beta^{(\nu)}|^2 \leq 2|\alpha_K^{(j)} - \tilde{\alpha}_{\tilde{K}}^{(\nu)}|^2 + 2|\tilde{\alpha}_{\tilde{K}}^{(\nu)} - \beta^{(\nu)}|^2.$$

Now  $|\alpha_K^{(j)} - \tilde{\alpha}_{\tilde{K}}^{(\nu)}|$  is the jump in the values of  $v_h$  at the node  $\nu$  which belongs to the interface  $e$  between  $K$  and  $\tilde{K}$ , and thus  $h_\nu^{d-2}|\alpha_K^{(j)} - \tilde{\alpha}_{\tilde{K}}^{(\nu)}|^2$  can be bounded by  $ch_e^{-1}|[v_h]_e|^2$  as was done in (2.20). On the other hand,

$$\tilde{\alpha}_{\tilde{K}}^{(\nu)} - \beta^{(\nu)} = (v_h|_{\tilde{K}})(\nu) - (\chi|_{\tilde{K}})(\nu) = \sum_{j=1}^m (\alpha_{\tilde{K}}^{(j)} - \beta_{\tilde{K}}^{(j)})\phi_{\tilde{K}}^{(j)}.$$

Thus,

$$|\tilde{\alpha}_{\tilde{K}}^{(\nu)} - \beta^{(\nu)}|^2 \leq \sum_{j=1}^m |\alpha_{\tilde{K}}^{(j)} - \beta_{\tilde{K}}^{(j)}|^2 \cdot \sum_{j=1}^m |\phi_{\tilde{K}}^{(j)}(\nu)|^2 \leq c \sum_{j=1}^m |\alpha_{\tilde{K}}^{(j)} - \beta_{\tilde{K}}^{(j)}|^2$$

for some constant  $c$  independent of  $h$ . Gathering these results, we have

$$\begin{aligned}
 \sum_{\nu \in \mathcal{N} \setminus \mathcal{N}^0} h_\nu^{d-2} \sum_{\substack{x_K^{(j)} = \nu \\ K \in \omega_\nu \setminus \tilde{\omega}_\nu}} |\alpha_K^{(j)} - \beta^{(\nu)}|^2 &= \sum_{\nu \in \mathcal{N} \setminus \mathcal{N}^0} h_\nu^{d-2} \sum_{\ell=0}^{L_{max}} \sum_{K \in \mathcal{T}_h^\ell} \sum_{\substack{x_K^{(j)} = \nu \\ K \in \omega_\nu \setminus \tilde{\omega}_\nu}} |\alpha_K^{(j)} - \beta^{(\nu)}|^2 \\
 &\leq c \sum_{e \in \mathcal{E}^I} h_e^{-1} |[v_h]_e|^2 + c \sum_{\ell=0}^L \sum_{K \in \mathcal{T}_h^\ell} h_K^{d-2} \sum_{j=1}^m |\alpha_K^{(j)} - \beta_K^{(j)}|^2 \\
 (2.23) \quad &\leq c \sum_{e \in \mathcal{E}^I} h_e^{-1} |[v_h]_e|^2 + c \sum_{\nu \in \mathcal{N}^0} h_\nu^{d-2} \sum_{x_K^{(j)} = \nu} |\alpha_K^{(j)} - \beta^{(\nu)}|^2 \\
 &\quad + c \sum_{\nu \in \mathcal{N} \setminus \mathcal{N}^0} h_\nu^{d-2} \sum_{\ell=0}^L \sum_{K \in \mathcal{T}_h^\ell} \sum_{\substack{x_K^{(j)} = \nu \\ K \in \omega_\nu \setminus \tilde{\omega}_\nu}} |\alpha_K^{(j)} - \beta^{(\nu)}|^2
 \end{aligned}$$

for some  $L$  that satisfies  $0 \leq L < L_{max}$ . Repeating this argument a finite number of times, the last sum in (2.23) (over the hanging nodes) will be eventually replaced by  $c \sum_{e \in \mathcal{E}^I} h_e^{-1} |[v_h]_e|^2 + c \sum_{\nu \in \mathcal{N}^0} h_\nu^{d-2} \sum_{x_K^{(j)} = \nu} |\alpha_K^{(j)} - \beta^{(\nu)}|^2$ . As we mentioned earlier, the latter term can be bounded by  $c(\sum_{e \in \mathcal{E}^I} h_e^{-1} |[v_h]_e|^2 + \sum_{e \in \mathcal{E}^B} h_e^{-1} |v_h|_e^2)$  just as in the conforming case. This concludes the proof.  $\square$

**3. A posteriori estimates.** This section is devoted to residual-type a posteriori estimates. The estimators as well as the exposition follow the lines found in Verfürth [18], with the exception of the technical details stemming from the discontinuous nature of  $V_h^r$ . We also note that our estimators (3.11) and (3.12) are entirely local.

Again for the sake of simplifying the exposition, and in this section only, we shall assume that  $f$  is a piecewise polynomial function on the mesh  $\mathcal{T}_h$ . Given that we have decided not to worry about quadrature errors, this is not an unreasonable assumption, since any given quadrature rule used to evaluate  $(f|_K, v)$  cannot distinguish between  $f|_K$  and the Lagrange interpolant of  $f$  at the quadrature points in  $K$ .

**THEOREM 3.1.** *Let  $e = u - u_h^\gamma$ . Then*

$$(3.1) \quad \sum_{K \in \mathcal{T}_h} \|\nabla e\|_K^2 \leq c \left\{ \sum_{K \in \mathcal{T}_h} h_K^2 \|f + \Delta u_h^\gamma\|_K^2 + \sum_{e \in \mathcal{E}^I} h_e |\partial_n u_h^\gamma|_e^2 + \gamma^2 \sum_{e \in \mathcal{E}^I} h_e^{-1} |[u_h^\gamma]|_e^2 + \gamma^2 \sum_{e \in \mathcal{E}^B} h_e^{-1} |u_h^\gamma|_e^2 \right\}.$$

*Proof.* From (2.4) and (2.5) there follows the orthogonality relation  $a_h^\gamma(e, v_h) = 0 \forall v_h \in V_h^r$ . Now for  $v \in E_h$  and  $v_h \in V_h^r$ , let  $\eta = v - v_h$ . We have

$$(3.2) \quad \begin{aligned} a_h^\gamma(e, v) &= a_h^\gamma(e, \eta) = (f, \eta) - a_h^\gamma(u_h^\gamma, \eta) \\ &= (f, \eta) - \left\{ \sum_{K \in \mathcal{T}_h} (\nabla u_h^\gamma, \nabla \eta)_K \right. \\ &\quad - \sum_{e \in \mathcal{E}^I} \left[ \langle \{\partial_n u_h^\gamma\}, [\eta] \rangle_e + \langle \{\partial_n \eta\}, [u_h^\gamma] \rangle_e - \gamma h_e^{-1} \langle [u_h^\gamma], [\eta] \rangle_e \right] \\ &\quad \left. - \sum_{e \in \mathcal{E}^B} \left[ \langle \partial_n u_h^\gamma, \eta \rangle_e + \langle \partial_n \eta, u_h^\gamma \rangle_e - \gamma h_e^{-1} \langle u_h^\gamma, \eta \rangle_e \right] \right\}. \end{aligned}$$

Now, integrating by parts, we see that

$$(3.3) \quad \begin{aligned} \sum_{K \in \mathcal{T}_h} (\nabla u_h^\gamma, \nabla \eta)_K &= \sum_{K \in \mathcal{T}_h} (-\Delta u_h^\gamma, \eta)_K + \sum_{K \in \mathcal{T}_h} \langle \partial_n u_h^\gamma, \eta \rangle_{\partial K} \\ &= \sum_{K \in \mathcal{T}_h} (-\Delta u_h^\gamma, \eta)_K + \sum_{e \in \mathcal{E}^I} \left[ \langle \{\partial_n u_h^\gamma\}, [\eta] \rangle_e + \langle [\partial_n u_h^\gamma], \eta^* \rangle_e \right] \\ &\quad + \sum_{e \in \mathcal{E}^B} \langle \partial_n u_h^\gamma, \eta \rangle_e, \end{aligned}$$

where  $\eta^* = \eta^-$  for Baker’s method and  $\eta^* = \frac{1}{2}(\eta^+ + \eta^-)$  for Arnold’s method. Now, using (3.3) in (3.2), we obtain

$$(3.4) \quad \begin{aligned} a_h^\gamma(e, v) &= \sum_{K \in \mathcal{T}_h} (f + \Delta u_h^\gamma, \eta)_K \\ &\quad + \sum_{e \in \mathcal{E}^I} \left[ \langle \{\partial_n \eta\}, [u_h^\gamma] \rangle_e - \langle [\partial_n u_h^\gamma], \eta^* \rangle_e - \gamma h_e^{-1} \langle [u_h^\gamma], [\eta] \rangle_e \right] \\ &\quad + \sum_{e \in \mathcal{E}^B} \left[ \langle \partial_n \eta, u_h^\gamma \rangle_e - \gamma h_e^{-1} \langle u_h^\gamma, \eta \rangle_e \right]. \end{aligned}$$

From the definition of  $a_h^\gamma(e, v)$  and using (3.4), we get

$$\begin{aligned}
 & \sum_{K \in \mathcal{T}_h} (\nabla e, \nabla v)_K + \gamma \sum_{e \in \mathcal{E}^I} h_e^{-1} \langle [e], [v] \rangle_e + \gamma \sum_{e \in \mathcal{E}^B} h_e^{-1} \langle e, v \rangle_e = a_h^\gamma(e, v) \\
 & \quad + \sum_{e \in \mathcal{E}^I} \left[ \langle \{\partial_n e\}, [v] \rangle_e + \langle \{\partial_n v\}, [e] \rangle_e \right] \\
 & \quad + \sum_{e \in \mathcal{E}^B} \left[ \langle \partial_n e, v \rangle_e + \langle \partial_n v, e \rangle_e \right] \\
 (3.5) \quad & = \sum_{K \in \mathcal{T}_h} (f + \Delta u_h^\gamma, \eta)_K + \sum_{e \in \mathcal{E}^I} \left[ \langle \{\partial_n \eta\}, [u_h^\gamma] \rangle_e - \langle [\partial_n u_h^\gamma], \eta^* \rangle_e \right. \\
 & \quad \left. + \langle \{\partial_n e\}, [v] \rangle_e + \langle \{\partial_n v\}, [e] \rangle_e - \gamma h_e^{-1} \langle [u_h^\gamma], [\eta] \rangle_e \right] \\
 & \quad + \sum_{e \in \mathcal{E}^B} \left[ \langle \partial_n \eta, u_h^\gamma \rangle_e + \langle \partial_n e, v \rangle_e + \langle \partial_n v, e \rangle_e - \gamma h_e^{-1} \langle u_h^\gamma, \eta \rangle_e \right].
 \end{aligned}$$

First note that

$$\langle \{\partial_n \eta\}, [u_h^\gamma] \rangle_e + \langle \{\partial_n v\}, [e] \rangle_e = - \langle \{\partial_n v_h\}, [u_h^\gamma] \rangle_e, \quad e \in \mathcal{E}^I,$$

and

$$\langle \partial_n \eta, u_h^\gamma \rangle_e + \langle \partial_n v, e \rangle_e = - \langle \partial_n v_h, u_h^\gamma \rangle_e, \quad e \in \mathcal{E}^B.$$

We will choose  $v_h$  to be piecewise constant on  $\mathcal{T}_h$ . Thus these four terms are zero. Hence (3.5) reduces to

$$\begin{aligned}
 & \sum_{K \in \mathcal{T}_h} (\nabla e, \nabla v)_K + \gamma \sum_{e \in \mathcal{E}^I} h_e^{-1} \langle [e], [v] \rangle_e + \gamma \sum_{e \in \mathcal{E}^B} h_e^{-1} \langle e, v \rangle_e = \sum_{K \in \mathcal{T}_h} (f + \Delta u_h^\gamma, \eta)_K \\
 (3.6) \quad & + \sum_{e \in \mathcal{E}^I} \left[ - \langle [\partial_n u_h^\gamma], \eta^* \rangle_e + \langle \{\partial_n e\}, [v] \rangle_e - \gamma h_e^{-1} \langle [u_h^\gamma], [\eta] \rangle_e \right] \\
 & + \sum_{e \in \mathcal{E}^B} \left[ \langle \partial_n e, v \rangle_e - \gamma h_e^{-1} \langle u_h^\gamma, \eta \rangle_e \right].
 \end{aligned}$$

At this point, we set  $v = e$  and observe that

$$\begin{aligned}
 & \sum_{e \in \mathcal{E}^I} \langle \{\partial_n e\}, [e] \rangle_e + \sum_{e \in \mathcal{E}^B} \langle \partial_n e, e \rangle_e = - \sum_{e \in \mathcal{E}^I} \langle \{\partial_n e\}, [u_h^\gamma] \rangle_e - \sum_{e \in \mathcal{E}^B} \langle \partial_n e, u_h^\gamma \rangle_e \\
 (3.7) \quad & = - \sum_{e \in \mathcal{E}^I} \langle \{\partial_n e\}, [u_h^\gamma - \chi] \rangle_e - \sum_{e \in \mathcal{E}^B} \langle \partial_n e, u_h^\gamma - \chi \rangle_e
 \end{aligned}$$

for any  $\chi \in V_h^r$ . Since  $a_h^\gamma(e, u_h^\gamma - \chi) = 0$ , we replace the terms  $\sum_{e \in \mathcal{E}^I} \langle \{\partial_n e\}, [e] \rangle_e + \sum_{e \in \mathcal{E}^B} \langle \partial_n e, e \rangle_e$  on the right-hand side of (3.6) with

$$\begin{aligned}
 & - \sum_{K \in \mathcal{T}_h} (\nabla e, \nabla (u_h^\gamma - \chi))_K - \sum_{e \in \mathcal{E}^I} \left[ \langle \{\partial_n (u_h^\gamma - \chi)\}, [u_h^\gamma] \rangle_e - \gamma h_e^{-1} |[u_h^\gamma]_e|^2 \right] \\
 & \quad - \sum_{e \in \mathcal{E}^B} \left[ \langle \partial_n (u_h^\gamma - \chi), u_h^\gamma \rangle_e - \gamma h_e^{-1} |u_h^\gamma|_e^2 \right]
 \end{aligned}$$

to obtain

$$\begin{aligned}
 \sum_{K \in \mathcal{T}_h} \|\nabla e\|_K^2 &= \sum_{K \in \mathcal{T}_h} (f + \Delta u_h^\gamma, \eta)_K - \sum_{e \in \mathcal{E}^I} \left[ \langle [\partial_n u_h^\gamma], \eta^* \rangle_e + \gamma h_e^{-1} \langle [u_h^\gamma], [\eta] \rangle_e \right] \\
 &\quad - \gamma \sum_{e \in \mathcal{E}^B} h_e^{-1} \langle u_h^\gamma, \eta \rangle_e - \sum_{K \in \mathcal{T}_h} (\nabla e, \nabla(u_h^\gamma - \chi))_K \\
 (3.8) \quad &\quad - \sum_{e \in \mathcal{E}^I} \langle \{\partial_n(u_h^\gamma - \chi)\}, [u_h^\gamma] \rangle_e - \sum_{e \in \mathcal{E}^B} \langle \partial_n(u_h^\gamma - \chi), u_h^\gamma \rangle_e.
 \end{aligned}$$

Here we have used the facts that  $\eta = e - v_h$ ,  $[e]|_e = -[u_h]|_e \ \forall e \in \mathcal{E}^I$ , and  $e|_e = -u_h|_e \ \forall e \in \mathcal{E}^B$ .

We now obtain bounds for the terms on the right-hand side of (3.8). Those that contain  $\eta$  are bounded by  $\frac{1}{2}$  times

$$\begin{aligned}
 (3.9) \quad &\frac{1}{\epsilon_1} \sum_{K \in \mathcal{T}_h} h_K^2 \|f + \Delta u_h^\gamma\|_K^2 + \frac{1}{\epsilon_2} \sum_{e \in \mathcal{E}^I} h_e |\partial_n u_h^\gamma|_e^2 + \frac{1}{\epsilon_3} \gamma \sum_{e \in \mathcal{E}^I} h_e^{-1} |[u_h^\gamma]|_e^2 \\
 &\quad + \frac{1}{\epsilon_4} \gamma \sum_{e \in \mathcal{E}^B} h_e^{-1} |u_h^\gamma|_e^2 + \epsilon_1 \sum_{K \in \mathcal{T}_h} h_K^{-2} \|\eta\|_K^2 + \epsilon_2 \sum_{e \in \mathcal{E}^I} h_e^{-1} |\eta^*|_e^2 \\
 &\quad + \epsilon_3 \gamma \sum_{e \in \mathcal{E}^I} h_e^{-1} |[\eta]|_e^2 + \epsilon_4 \gamma \sum_{e \in \mathcal{E}^B} h_e^{-1} |\eta|_e^2
 \end{aligned}$$

for any  $\epsilon_i > 0$ ,  $i = 1, \dots, 4$ . To estimate the “ $\eta$ ” terms in (3.9) we choose as  $v_h$  the best piecewise constant approximation of  $e$ . From (2.10) this gives

$$h_K^{-2} \|\eta\|_K^2 = h_K^{-2} \|e - v_h\|_K^2 \leq c \|\nabla e\|_K^2.$$

Also, using the trace inequality (2.8) and (2.10), we obtain

$$\begin{aligned}
 h_e^{-1} (|\eta^*|_e^2 + |[\eta]|_e^2) &\leq c \sum_{K=K^+, K^-} h_e^{-1} (h_K^{-1} \|\eta\|_K^2 + h_K \|\nabla \eta\|_K^2) \\
 &\leq c \sum_{K=K^+, K^-} h_e^{-1} h_K \|\nabla e\|_K^2.
 \end{aligned}$$

The local quasiuniformity of the mesh implies that  $h_e \approx h_{K^+} \approx h_{K^-}$ . Thus  $h_e^{-1} h_K \leq c$ . A similar bound holding for  $\sum_{e \in \mathcal{E}^B} h_e^{-1} |\eta|_e^2$ , we can now hide the “ $\eta$ ” terms in the left-hand side of (3.8) by taking the  $\epsilon$ ’s sufficiently small. In particular, we must take  $\epsilon_3 \approx 1/\gamma$  and  $\epsilon_4 \approx 1/\gamma$ .

To obtain (3.1), we need to estimate the terms containing  $u_h^\gamma - \chi$ . Indeed these are bounded by

$$\begin{aligned}
 (3.10) \quad &\epsilon \sum_{K \in \mathcal{T}_h} \|\nabla e\|_K^2 + \frac{1}{\epsilon} \sum_{K \in \mathcal{T}_h} \|\nabla(u_h^\gamma - \chi)\|_K^2 + \sum_{e \in \mathcal{E}^I} h_e |\{\partial_n(u_h^\gamma - \chi)\}|_e^2 \\
 &\quad + \sum_{e \in \mathcal{E}^I} h_e^{-1} |[u_h^\gamma]|_e^2 + \sum_{e \in \mathcal{E}^B} h_e |\partial_n(u_h^\gamma - \chi)|_e^2 + \sum_{e \in \mathcal{E}^B} h_e^{-1} |u_h^\gamma|_e^2.
 \end{aligned}$$

Using the estimates (2.8) and (2.9), we see that the two terms in (3.10) that contain  $\partial_n(u_h^\gamma - \chi)$  are bounded by  $\sum_{K \in \mathcal{T}_h} \|\nabla(u_h^\gamma - \chi)\|_K^2$ . In view of Theorem 2.2, the latter is bounded by  $\sum_{e \in \mathcal{E}^I} h_e^{-1} |[u_h^\gamma]|_e^2 + \sum_{e \in \mathcal{E}^B} h_e^{-1} |u_h^\gamma|_e^2$ . Using this fact completes the proof.  $\square$

THEOREM 3.2. *Suppose that  $f$  is a piecewise polynomial on  $\mathcal{T}_h$ . Then*  
 (i) *for each  $K \in \mathcal{T}_h$ ,*

$$(3.11) \quad h_K^2 \|f + \Delta u_h^\gamma\|_K^2 \leq c \|\nabla e\|_K^2;$$

(ii) *for  $e = K^+ \cap K^- \in \mathcal{E}^I$ ,*

$$(3.12) \quad h_e |[\partial_n u_h^\gamma]_e|^2 \leq c(\|\nabla e\|_{K^+}^2 + \|\nabla e\|_{K^-}^2).$$

*Proof.* To estimate  $\|f + \Delta u_h^\gamma\|_K$ , we set  $v_h = 0$  and  $v|_K = (f + \Delta u_h^\gamma) b_K$ , where  $b_K$  is the ‘‘bubble’’ function  $27\lambda_1\lambda_2\lambda_3$  expressed in terms of the barycentric coordinates of  $K$ ; we extend  $v$  to the outside of  $K$  by zero. Using this  $v$  in (3.6), we obtain

$$\int_K (f + \Delta u_h^\gamma)^2 b_K dx = (\nabla e, \nabla((f + \Delta u_h^\gamma) b_K))_K.$$

Now since  $b_K > 0$  on  $\text{int}(K)$ ,  $(\int_K (\cdot)^2 b_K dx)^{1/2}$  defines a norm on  $L^2(K)$ , equivalent to the  $L^2$  norm on  $P_m(K)$  for any fixed  $m$ . Thus, there exists a constant  $c > 0$  such that

$$(3.13) \quad \int_K (f + \Delta u_h^\gamma)^2 b_K dx \geq c \|f + \Delta u_h^\gamma\|_K^2.$$

Since  $\|b_K\|_{L^\infty(K)} = 1$ , a scaling argument can be used to show that, while the constant  $c$  may depend on  $r$  and the degree of  $f$ , it is independent of  $h_K$ . On the other hand, using the inverse inequality (2.9), we have

$$\begin{aligned} (\nabla e, \nabla((f + \Delta u_h^\gamma) b_K))_K &\leq \|\nabla e\|_K \|\nabla((f + \Delta u_h^\gamma) b_K)\|_K \\ &\leq c\epsilon \|f + \Delta u_h^\gamma\|_K^2 + \frac{1}{\epsilon} h_K^{-2} \|\nabla e\|_K^2. \end{aligned}$$

This gives (i). We next estimate  $h_e |[\partial_n u_h^\gamma]_e|^2$ . Let  $e = \partial K^+ \cap \partial K^-$  and suppose that  $e$  is a full edge of both  $K^+$  and  $K^-$ . (See Remark 3.1 below.) Extend  $[\partial_n u_h^\gamma]$  to a function  $\phi$  defined over  $\tilde{K} = K^+ \cup K^-$  by extending by constants along lines normal to  $e$ ; see Figure 3.1. Also, let  $b$  denote the bubble function on  $\tilde{K}$  given by

$$b|_{K^+} = 4\lambda_1^+ \lambda_2^+, \quad b|_{K^-} = 4\lambda_1^- \lambda_2^-.$$

Let  $v = \phi b$  and set  $v_h = 0$ . Using this  $v$  in (3.6), we get

$$\begin{aligned} h_e \int_e [\partial_n u_h^\gamma]^2 b ds &= h_e \sum_{K=K^+, K^-} \left[ (f + \Delta u_h^\gamma, \phi b)_K - (\nabla e, \nabla(\phi b))_K \right] \\ &\leq h_e \sum_{K=K^+, K^-} \left[ \|f + \Delta u_h^\gamma\|_K \|\phi b\|_K + \|\nabla e\|_K \|\nabla(\phi b)\|_K \right] \\ (3.14) \quad &\leq \frac{1}{2\epsilon} \sum_{K=K^+, K^-} \left\{ h_K^2 \|f + \Delta u_h^\gamma\|_K^2 + \|\nabla e\|_K^2 \right\} \\ &\quad + \frac{\epsilon}{2} \sum_{K=K^+, K^-} \left\{ \|\phi b\|_K^2 + h_K^2 \|\nabla(\phi b)\|_K^2 \right\}. \end{aligned}$$

Using arguments similar to those leading to (3.13), we obtain

$$(3.15) \quad \int_e [\partial_n u_h^\gamma]^2 b ds \geq c |[\partial_n u_h^\gamma]_e|^2$$

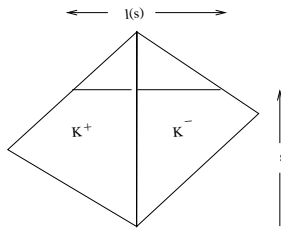


FIG. 3.1.

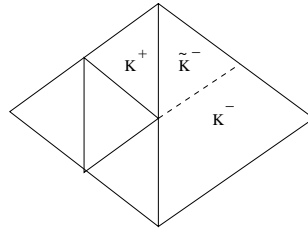


FIG. 3.2.

for some positive constant  $c$  depending only on  $r$ . Moreover,

$$(3.16) \quad \|\phi b\|_{\tilde{K}}^2 \leq \|\phi\|_{\tilde{K}}^2 = \int_e [\partial_n u_h^\gamma]^2 l(s) ds \leq h_e |[\partial_n u_h^\gamma]|_e^2,$$

where  $l(s)$  is as in Figure 3.1. Using the inverse inequality (2.9), we see that

$$(3.17) \quad \sum_{K=K^+, K^-} h_K^2 \|\nabla(\phi b)\|_K^2 \leq c \|\phi b\|_{\tilde{K}}^2 \leq ch_e |[\partial_n u_h^\gamma]|_e^2.$$

The required estimate now follows from (3.14)–(3.17).  $\square$

*Remark 3.1.* If  $e$  is not a full edge of one of the triangles, say  $K^-$ , then we can work with  $\tilde{K}^-$  instead; see Figure 3.2.

**4. Estimates based on the solution of local problems.** In this section, we shall introduce and analyze a posteriori estimates that are based on domain decomposition techniques proposed in [11]. The approach consists of viewing the computed solution  $u_h^\gamma$  as the coarse-mesh approximation to some function which is arguably a more accurate approximation to  $u$ . Before suggesting some choices for this quantity, let us say that there will be no attempt to compute it directly, but rather to approximate it by adding to  $u_h^\gamma$  a function obtained through the solution of “local” problems. For simplicity, we restrict the exposition to  $d = 2$  and assume that  $\mathcal{T}_h$  is a conforming mesh of triangles. Also, we should note that the results of [11] concern Baker’s formulation only; however, we believe that similar results can be obtained for Arnold’s formulation.

**4.1. A nonoverlapping approach.** To begin, let  $\mathcal{T}_{h/2}$  be the mesh obtained by cutting every  $K \in \mathcal{T}_h$  into four equal triangles. In a similar way, we may define  $\mathcal{T}_{h'} := \mathcal{T}_{h/2^p}$ ,  $h' := h/2^p$  by repeating this process  $p$  times. On the latter, we define a finite element space  $V' := V_{h'}^{r'}$  of discontinuous piecewise polynomial functions of degree less than or equal to  $r' - 1$ , where  $r' \geq r$ . This way,  $V_h^r$  is a subspace of  $V'$ . On  $V' \times V'$  we define the bilinear form  $a' := a_{h'}^{\gamma'}$ , just as in the definition of  $a_h^\gamma$  in (2.3). It is crucial for the analysis that  $a_h^\gamma$  be the restriction of  $a'$  to  $V_h^r$  in the sense that

$$(4.1) \quad a_h^\gamma(v, w) = a'(v, w) \quad \forall v, w \in V_h^r.$$

*Remark 4.1.* By comparing the penalty terms in  $a_h^\gamma$  and  $a'$ , we see that (4.1) requires the condition  $\gamma'(h')^{-1} = \gamma h^{-1}$ , which, in view of the fact that  $h' = h/2^p$ , is equivalent to  $\gamma' = \gamma 2^{-p}$ . Now since the coercivity of the forms  $a'$  and  $a_h^\gamma$  can be guaranteed only if  $\gamma' \geq \gamma'_0(r')$  and  $\gamma \geq \gamma_0(r)$ , respectively (see Lemma 2.1(i)), we see that  $\gamma$  must be chosen sufficiently large in order to have  $\gamma 2^{-p} \geq \gamma'_0$ . This does not

present any theoretical difficulties, since  $\gamma$  can take on arbitrarily large values without having any result discussed in this work break down. On the other hand, the quality of the a priori and a posteriori estimates may suffer, as is the case with (3.1). (In this respect, see the discussion at the beginning of section 5 and Figures 5.5 and 5.6.) In practice, however, we anticipate that  $r' = r + 1$  and/or  $h' = h/2$  should be sufficient. This was indeed the case in all our numerical experiments.

For each  $K \in \mathcal{T}_h$ , we consider the “local” space  $V'(K)$  obtained by restricting  $V'$  to  $K$ . By extending the elements of  $V'(K)$  by zero to the rest of  $\Omega$ ,  $V'(K)$  becomes a subspace of  $V'$ . Indeed, the latter is the direct sum of these local subspaces. On  $V'(K) \times V'(K)$  we introduce the bilinear form  $a'_K(\cdot, \cdot)$  as the restriction of  $a'(\cdot, \cdot)$  to  $V'(K) \times V'(K)$  (see (4.4) in [11]). As such,  $a'_K$  inherits the symmetry and coercivity of  $a'$  on  $V'(K)$ . In particular, for any  $\gamma' \geq \gamma'_0$  there holds

$$(4.2) \quad a'_K(v, v) \geq c \|v\|_{1,K}^2 \quad \forall v \in V'(K),$$

where  $\|\cdot\|_{1,K}$  denotes the restriction of the  $\|\cdot\|_{1,h'}$  norm to  $V'(K)$ . Adopting the terminology of [11], we consider  $\mathcal{T}_h$  as the coarse mesh of  $\mathcal{T}_{h'}$ . Also, each  $K \in \mathcal{T}_h$  is considered as a subdomain in  $\mathcal{T}_{h'}$ . In other words,  $\mathcal{T}_h$  is both the coarse mesh and the subdomain partition of  $\mathcal{T}_{h'}$ .

Now let  $u' := u_{h'}^{\gamma'} \in V'$  be the discontinuous Galerkin approximation of  $u$  in the space  $V'$ ; i.e.,

$$(4.3) \quad a'(u', v) = (f, v) \quad \forall v \in V'.$$

At this point, we observe that, by virtue of (4.1), (2.5) and (4.3) imply the following orthogonality relation:

$$(4.4) \quad a'(u' - u_h^\gamma, v) = 0 \quad \forall v \in V_h^\gamma.$$

Next, let the functions  $\{\eta_K \in V'(K) \mid K \in \mathcal{T}_h\}$  be given as the solutions of the local problems

$$(4.5) \quad a'_K(\eta_K, v) = (f, v) - a'(u_h^\gamma, v) \quad \forall v \in V'(K).$$

The functions  $\{\eta_K\}$  can be computed independently of each other and in parallel. Moreover, the function  $\eta := \sum_{K \in \mathcal{T}_h} \eta_K$  approximates  $\zeta := u' - u_h^\gamma$  in the following sense.

**THEOREM 4.1.** *There exist positive constants  $C_1$  and  $C_2$  such that*

$$(4.6) \quad C_1 \|\eta\|_{1,h'} \leq \|\zeta\|_{1,h'} \leq C_2 \frac{h}{h'} \|\eta\|_{1,h'}.$$

*Proof.* Since  $(f, v) = a'(u', v)$ ,  $v \in V'$ , from (4.5) we have

$$(4.7) \quad a'_K(\eta_K, v) = a'(\zeta, v) \quad \forall v \in V'(K).$$

Thus,

$$(4.8) \quad \sum_{K \in \mathcal{T}_h} a'_K(\eta_K, \eta_K) = a'(\zeta, \eta) \leq c \|\zeta\|_{1,h'} \|\eta\|_{1,h'}.$$

From (4.2) it follows that  $\sum_{K \in \mathcal{T}_h} a'_K(\eta_K, \eta_K) \geq c \sum_{K \in \mathcal{T}_h} \|\eta_K\|_{1,K}^2$ . On the other hand, it is easy to see that  $\sum_{K \in \mathcal{T}_h} \|\eta_K\|_{1,K}^2 \geq \|\eta\|_{1,h'}^2$ . Thus, the first half of (4.6) follows.

To prove the second inequality, let  $\zeta = \sum_{K \in \mathcal{T}_h} \zeta_K, \zeta_K \in V'(K)$ . Let  $\zeta_0$  be the piecewise constant function on  $\mathcal{T}_h$  defined by

$$\zeta_0|_K := \zeta_{K,0} = \frac{1}{|K|} \int_K \zeta_K \, dx.$$

Now since  $\zeta_0 \in V_h^r$ , it follows from (4.7) and (4.4) that

$$a'_K(\eta_K, \zeta_K - \zeta_{K,0}) = a'(\zeta, \zeta_K - \zeta_{K,0}) = a'(\zeta, \zeta_K).$$

Summing over  $K \in \mathcal{T}_h$ , we obtain

$$(4.9) \quad \sum_{K \in \mathcal{T}_h} a'_K(\eta_K, \zeta_K - \zeta_{K,0}) = a'(\zeta, \zeta) \geq c \|\zeta\|_{1,h'}^2.$$

Now, it follows from the Cauchy-Schwarz inequality that

$$(4.10) \quad \begin{aligned} \sum_{K \in \mathcal{T}_h} a'_K(\eta_K, \zeta_K - \zeta_{K,0}) &\leq \left( \sum_{K \in \mathcal{T}_h} a'_K(\eta_K, \eta_K) \right)^{1/2} \\ &\times \left( \sum_{K \in \mathcal{T}_h} a'_K(\zeta_K - \zeta_{K,0}, \zeta_K - \zeta_{K,0}) \right)^{1/2}. \end{aligned}$$

Also from (4.8) it follows that

$$(4.11) \quad \left( \sum_{K \in \mathcal{T}_h} a'_K(\eta_K, \eta_K) \right)^{1/2} \leq c \|\zeta\|_{1,h'}^{1/2} \|\eta\|_{1,h'}^{1/2}.$$

On the other hand, with the interface bilinear form  $I'(\cdot, \cdot)$  defined in (4.5) of [11],

$$\begin{aligned} I'(u, v) = \sum_{e' \in \mathcal{E}^I} \left\{ \langle \{\partial_n u\}, v^- \rangle_{e'} + \langle \{\partial_n v\}, u^- \rangle_{e'} \right. \\ \left. - \gamma'(h')^{-1} [\langle u^+, v^- \rangle_{e'} + \langle v^+, u^- \rangle_{e'}] \right\} \quad \forall u, v \in V', \end{aligned}$$

we have

$$(4.12) \quad \begin{aligned} \sum_{K \in \mathcal{T}_h} a'_K(\zeta_K - \zeta_{K,0}, \zeta_K - \zeta_{K,0}) &= a'(\zeta - \zeta_0, \zeta - \zeta_0) - I'(\zeta - \zeta_0, \zeta - \zeta_0) \\ &\leq 2a'(\zeta, \zeta) + 2a'(\zeta_0, \zeta_0) + |I'(\zeta - \zeta_0, \zeta - \zeta_0)|. \end{aligned}$$

In [11] it is proved that  $a'(\zeta_0, \zeta_0)$  and  $|I'(\zeta - \zeta_0, \zeta - \zeta_0)|$  are bounded by  $c \frac{h}{h'} a'(\zeta, \zeta)$ . Thus from (4.9)–(4.12) it follows that

$$\|\zeta\|_{1,h'}^2 \leq c \left( \frac{h}{h'} \right)^{1/2} \|\zeta\|_{1,h'}^{3/2} \|\eta\|_{1,h'}^{1/2},$$

from which the second inequality of (4.6) follows.  $\square$

We shall use the equivalence just proved to obtain estimates for  $e = u - u_h^\gamma$ . Letting  $e' = u - u'$ , we have  $e = e' + \zeta$ . We now argue as follows: It is reasonable to expect that  $e'$  is much smaller than  $e$ , say in the energy norm; therefore  $e$  and  $\zeta$

are nearly equal. Since  $\zeta$  is not computed, we shall approximate it, and hence  $e$ , by  $\eta$ , where the latter is obtained by solving local problems. To quantify matters, since  $a'(e', v) = 0 \forall v \in V'$  and  $\zeta \in V'$ , we obtain

$$(4.13) \quad a'(e, e) = a'(e', e') + a'(\zeta, \zeta).$$

It follows from this and (2.7) that  $a'(e', e') = \epsilon a'(e, e)$  for some  $0 < \epsilon < 1$ . Based on a priori estimates, it is reasonable to expect that  $\epsilon = O(\frac{(h')^{r'-1}}{h^{r-1}}) \ll 1$ . Thus,

$$a'(e, e) = \frac{1}{1 - \epsilon} a'(\zeta, \zeta).$$

In view of the equivalence between  $\eta$  and  $\zeta$  provided by Theorem 4.1, we can use  $a'(\eta, \eta)$  to obtain lower and upper bounds for  $a'(e, e)$ .

**4.2. An overlapping approach.** Let  $\Omega = \cup_{e \in \mathcal{E}} \Omega_e$  be an overlapping decomposition of  $\Omega$ , where each  $\Omega_e$  is the following union of the triangles in  $\mathcal{T}_h$ : If  $e \in \mathcal{E}^I$ , then  $e = \partial K^+ \cap \partial K^-$  and  $\Omega_e = K^+ \cup K^-$ ; else if  $e \in \mathcal{E}^B$ , then  $e = \partial K \cap \partial \Omega$  and  $\Omega_e = K$ .

On  $\mathcal{T}_h$ , we define a finite element space  $V' := V_h^{r'}$  of discontinuous piecewise polynomial functions of degree less than or equal to  $r' - 1$ , where  $r' \geq r + 1$ ; let us recall that  $r \geq 2$  is fixed. We construct a subspace decomposition of this latter finite element space by defining the subspaces  $\{V'_e\}_{e \in \mathcal{E}}$  associated with the subdomains  $\{\Omega_e\}_{e \in \mathcal{E}}$  by

$$V'_e = \{v_h \in V', v_h = 0 \text{ in } \Omega \setminus \bar{\Omega}_e\}.$$

Thus the following decomposition, which is not direct, holds:

$$(4.14) \quad V' = V_h^r + \sum_{e \in \mathcal{E}} V'_e.$$

On  $V' \times V'$ , we define the bilinear form  $a' := a_h^{\gamma'}$  as in the definition of  $a_h^\gamma$  in (2.3), and again it is crucial for the analysis that  $a_h^{\gamma'}$  be the restriction of  $a'$  to  $V_h^r$  in the sense that

$$(4.15) \quad a_h^{\gamma'}(v, w) = a'(v, w) \quad \forall v, w \in V_h^r.$$

For each edge  $e \in \mathcal{E}$ , we consider on  $V'_e \times V'_e$  the symmetric and coercive bilinear form  $a'_e(\cdot, \cdot)$  as the restriction of  $a'(\cdot, \cdot)$  to  $V'_e \times V'_e$  (see (4.4) in [11]),

$$(4.16) \quad a'_e(v, w) = a'(v, w) \quad \forall v, w \in V'_e.$$

Following [11], we are able to define the additive operator  $T = T_0 + \sum_{e \in \mathcal{E}} T_e$ , where  $T_0$  is a projection operator from  $V'$  to  $V_h^r$  defined by

$$(4.17) \quad a_h^{\gamma'}(T_0 u, v) = a'(T_0 u, v) = a'(u, v) \quad \forall v \in V_h^r$$

and  $T_e$  is a projection operator from  $V'$  to  $V'_e$  defined by

$$(4.18) \quad a'_e(T_e u, v) = a'(T_e u, v) = a'(u, v) \quad \forall v \in V'_e.$$

Lemmas 5.1–5.5 in [11] can easily be adapted to the present case with  $H = h$ , and  $\delta \sim h$ , and Theorem 5.7 in [11] then reads as follows.

THEOREM 4.2. *There exist positive constants  $c_1, c_2$ , which are independent of  $h$  and of the number of edges, such that there holds the estimate*

$$(4.19) \quad c_1 a'(v, v) \leq a'(Tv, Tv) \leq c_2 a'(v, v) \quad \forall v \in V'.$$

Now let  $u' := u_h^\gamma \in V'$  be the discontinuous Galerkin approximation of  $u$  in the space  $V'$ ; i.e.,

$$(4.20) \quad a'(u', v) = (f, v) \quad \forall v \in V'.$$

Next, let the functions  $\eta_e \in V_e'$  be given as the solutions of the local problems

$$(4.21) \quad a_e'(\eta_e, v) = (f, v) - a'(u_h^\gamma, v) \quad \forall v \in V_e'.$$

The functions  $\{\eta_e\}_{e \in \mathcal{E}}$  can be computed independently of each other and in parallel, and the function  $\eta := \sum_{e \in \mathcal{E}} \eta_e$  approximates  $\zeta := u' - u_h^\gamma$  in the following sense.

THEOREM 4.3. *There exist positive constants  $C_1$  and  $C_2$  such that*

$$(4.22) \quad C_1 \|\zeta\|_{1,h} \leq \|\eta\|_{1,h} \leq C_2 \|\zeta\|_{1,h}.$$

*Proof.* Let us prove that  $\eta = T\zeta$ . Indeed, from (4.20) and (4.21) we get

$$(4.23) \quad a_e'(\eta_e, v) = a'(u' - u_h^\gamma, v) \quad \forall v \in V_e',$$

which means from the definition (4.18) of  $T_e$  that  $\eta_e = T_e \zeta \forall e \in \mathcal{E}$ .

Now, by virtue of (4.15), (2.5) and (4.20) imply the orthogonality relation

$$(4.24) \quad a'(u' - u_h^\gamma, v) = 0 \quad \forall v \in V_h^r,$$

which means from the definition of  $T_0$  that  $T_0 \zeta = 0$ .

Thus (4.22) follows from Theorem 4.2.  $\square$

We conclude, in a similar way as in section 4.1, that we can use  $a'(\eta, \eta)$  to obtain lower and upper bounds for  $a'(e, e)$ .

*Remark 4.2.* As for the nonoverlapping approach, we could define the finite element space  $V' := V_{h'}^r$  with  $h' = h/2^p$ , where the mesh is obtained in dimension 2 by cutting  $p$  times every triangle into four equal triangles.

**5. Numerical results in one dimension.** In this section, we present numerical results obtained from the one-dimensional (1-d) model problem

$$(5.1) \quad -\frac{d^2 u}{dx^2} = f, \quad 0 < x < 1, \quad u(0) = u(1) = 0,$$

with the exact solution  $u(x) = e^{-\alpha(x-\frac{1}{2})^2}$ ,  $\alpha = 100$ .

For the sake of brevity, we shall consider in these numerical experiments only the weak formulation (2.3) due to Baker; for some comparisons with the Arnold formulation, we refer to the technical report [13].

**5.1. Convergence with respect to  $\gamma$  and  $h$ .** For a number of years we have been interested in the behavior of the discontinuous Galerkin approximations as a function of the penalty parameter  $\gamma$ , and indeed we had a proof (unpublished) that

$$(5.2) \quad \lim_{\gamma \rightarrow \infty} u_h^\gamma = u_h^G,$$

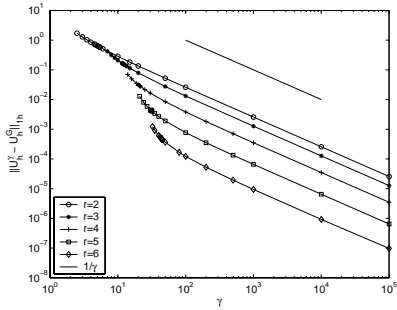


FIG. 5.1.  $\|u_h^\gamma - u_h^G\|_{1,h}$  versus  $\gamma$ .

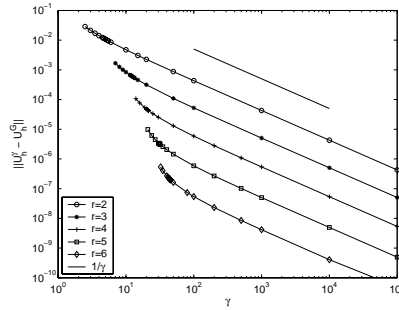


FIG. 5.2.  $\|u_h^\gamma - u_h^G\|$  versus  $\gamma$ .

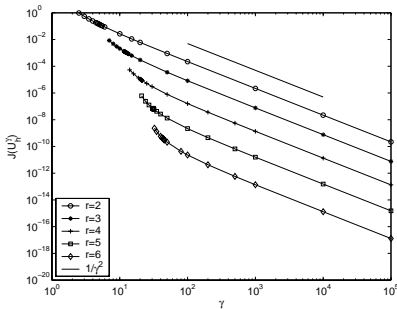


FIG. 5.3.  $J(u_h^\gamma)$  versus  $\gamma$ .

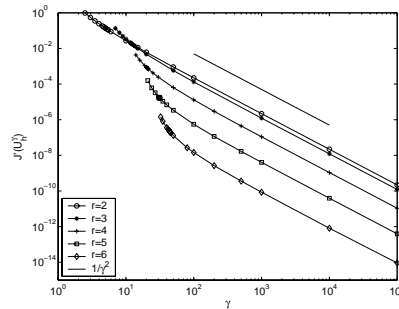


FIG. 5.4.  $J'(u_h^\gamma)$  versus  $\gamma$ .

where  $u_h^G$  is the standard Galerkin approximation of  $u$  defined by

$$(5.3) \quad (\nabla u_h^G, \nabla \chi) = (f, \chi) \quad \forall \chi \in V_h^r.$$

A more recent proof can be found in [14]. Besides its intrinsic value, this result can be used to show that the discontinuous Galerkin method can yield more accurate results than the standard Galerkin version for a range of values of  $\gamma$ , as shown later in Figure 5.5.

In the first experiments, the domain  $[0, 1]$  is divided into a uniform mesh of 20 subintervals. The approximations  $u_h^\gamma$ , solution of (2.5), and  $u_h^G$ , solution of (5.3), are computed using piecewise polynomials of degree up to 5. Figures 5.1 and 5.2 show the difference between  $u_h^\gamma$  and  $u_h^G$  in the energy norm  $\|\cdot\|_{1,h}$  and the  $L^2$  norm, respectively, as a function of  $\gamma$ . These plots highlight the convergence (5.2) according to the rate  $O(\frac{1}{\gamma})$ . Similarly, Figure 5.3 shows the behavior of the jump in  $u_h^\gamma$ ,

$$(5.4) \quad J(u_h^\gamma) \equiv \sum_{e \in \mathcal{E}^I} h_e^{-1} |[u_h^\gamma]|_e^2 + \sum_{e \in \mathcal{E}^B} h_e^{-1} |u_h^\gamma|_e^2,$$

as a function of  $\gamma$ . Note that  $J(u_h^\gamma) = J(u_h^\gamma - u_h^G)$ . We see that  $J(u_h^\gamma)$  behaves as  $\frac{1}{\gamma^2}$  when  $\gamma$  tends to infinity. In the same way, one can observe in Figure 5.4 that the jump of the derivative

$$(5.5) \quad J'(u_h^\gamma) \equiv \sum_{e \in \mathcal{E}^I} h_e |\{\partial_n(u_h^\gamma - u_h^G)\}|_e^2 + \sum_{e \in \mathcal{E}^B} h_e |\partial_n(u_h^\gamma - u_h^G)|_e^2$$

behaves also as  $\frac{1}{\gamma^2}$ .

We now study the difference between the discontinuous Galerkin and exact solutions of the problem. For  $h = \frac{1}{20}$  and  $r = 2$ ,  $\|u_h^\gamma - u\|$  and  $|u_h^\gamma - u|_{1,h} \equiv (\sum_{K \in \mathcal{T}_h} \|\nabla(u_h^\gamma - u)\|_K^2)^{1/2}$  are plotted in Figures 5.5 and 5.6, respectively. We see that these converge to values (represented by the dashed lines) which, in view of (5.2), must be  $\|u_h^G - u\|$  and  $\|\nabla(u_h^G - u)\|$ , respectively. We also observe that there exists an optimal value  $\gamma_{opt}$  of the penalty parameter  $\gamma$ , for which  $\|u_h^\gamma - u\|$  is minimized. From these and other numerical experiments not reported here (see the report [13] for other values of  $r$ ), we claim that, in the case of the Baker formulation, this optimal value does not depend either on the mesh size or on  $u$  (or, from an equivalent point of view, on the function  $f$ ), but depends on the degree of the polynomial approximations: It follows approximately the rule

$$(5.6) \quad \gamma_{opt} = (r - 1)(r + 3),$$

as can be seen in Figure 5.7, where the circles represent the numerical value of  $\gamma_{opt}$  for different values of  $r - 1$  and the continuous line represents the  $(r - 1)(r + 3)$  function.

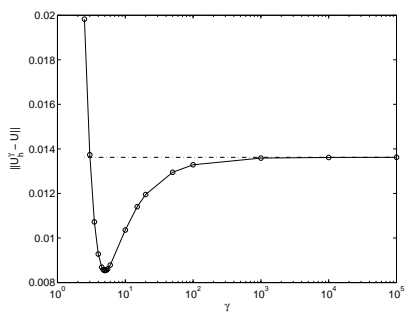


FIG. 5.5.  $\|u_h^\gamma - u\|$  versus  $\gamma$ .

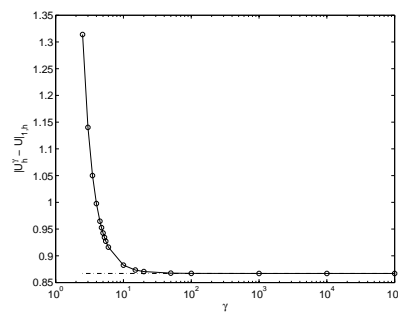


FIG. 5.6.  $|u_h^\gamma - u|_{1,h}$  versus  $\gamma$ .

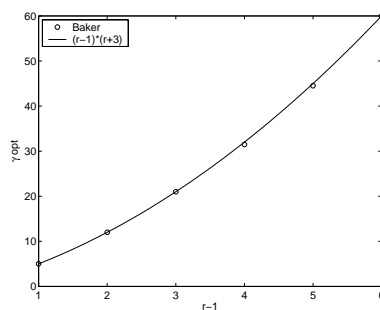


FIG. 5.7.  $\gamma_{opt}$  versus  $r - 1$ .

For the parameter  $\gamma$  chosen approximately equal to  $\gamma_{opt}$ , we now investigate the convergence of  $u_h^\gamma$  to  $u$  on a sequence of uniformly refined meshes. The differences  $\|u_h^\gamma - u\|$  and  $\|u_h^\gamma - u\|_{1,h}$  are plotted in Figures 5.8 and 5.9, respectively, for piecewise polynomials of degree  $r - 1 = 1, 2, 3, 4$ . The observed rates of convergence of  $O(h^r)$  and  $O(h^{r-1})$ , respectively, conform to the a priori estimates expressed in (2.12) and

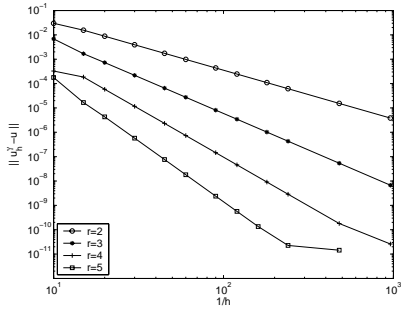


FIG. 5.8.  $\|u_h^\gamma - u\|$  versus  $1/h$ .

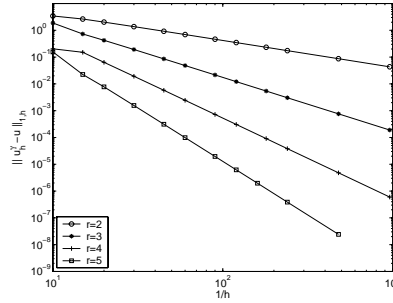


FIG. 5.9.  $\|u_h^\gamma - u\|_{1,h}$  versus  $1/h$ .

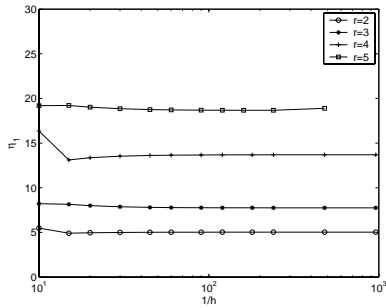


FIG. 5.10.  $\eta_1$  versus  $1/h$ .

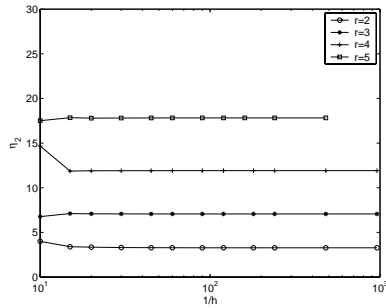


FIG. 5.11.  $\eta_2$  versus  $1/h$ .

in (2.11): The distorted line for  $r = 5$  is due to the limitations of computations in double precision.

**5.2. Effectivity indices.** In order to judge the quality of the various error estimators presented above, we compute for each an effectivity index, defined as the ratio of the estimator to the exact error. First, we study three estimators featured in Theorems 3.1 and 3.2. Specifically, in Figure 5.10, the effectivity index  $\eta_1$  of the first estimator corresponding to Theorem 3.1 and formula (3.1) is plotted as a function of  $1/h$  and for values of  $r - 1$  between 1 and 4:

$$(5.7) \quad \eta_1^2 = \frac{\sum_{K \in \mathcal{T}_h} h_K^2 \|f + \Delta u_h^\gamma\|_K^2 + \sum_{e \in \mathcal{E}^I} h_e |\partial_n u_h^\gamma|_e^2 + \gamma^2 \sum_{e \in \mathcal{E}^I} h_e^{-1} |[u_h^\gamma]|_e^2 + \gamma^2 \sum_{e \in \mathcal{E}^B} h_e^{-1} |u_h^\gamma|_e^2}{\sum_{K \in \mathcal{T}_h} \|\nabla e\|_K^2}.$$

In Figure 5.11, we plot the effectivity index  $\eta_2$  of the second estimator, which corresponds to Theorem 3.2(i) and to (3.11):

$$(5.8) \quad \eta_2^2 = \frac{\sum_{K \in \mathcal{T}_h} h_K^2 \|f + \Delta u_h^\gamma\|_K^2}{\sum_{K \in \mathcal{T}_h} \|\nabla e\|_K^2}.$$

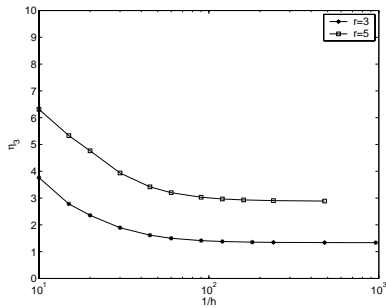


FIG. 5.12.  $\eta_3$  versus  $1/h$  for even  $r - 1$ .

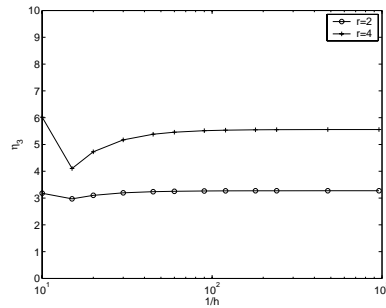


FIG. 5.13.  $\eta_3$  versus  $1/h$  for odd  $r - 1$ .

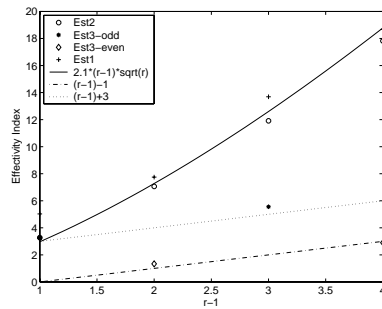


FIG. 5.14. *Effectivity indices versus  $r - 1$ .*

Finally, Figures 5.12 and 5.13 represent for odd and even degrees of polynomials, respectively, the effectivity index  $\eta_3$  of the estimator associated with Theorem 3.2(ii) and (3.12):

$$(5.9) \quad \eta_3^2 = \frac{\sum_{e \in \mathcal{E}^I} h_e \left| [\partial_n u_h^\gamma] \right|_e^2}{\sum_{K \in \mathcal{T}_h} \|\nabla e\|_K^2}.$$

It is seen that, as  $h$  decreases, these indices converge to values larger than 1. We also observe that  $\eta_1$  and  $\eta_2$  attain their respective asymptotic values rather quickly. On the other hand, while  $\eta_3$  is somewhat slower in that respect, it is still nearly constant over a wide range of values of  $h$ .

Additionally, the asymptotic values depend strongly on  $r$ . Since it is desirable to have effectivity indices close to 1, we tried to find simple laws describing this dependence. As evidenced by Figure 5.14, the following functions seem to “fit” the asymptotic values reasonably well:

$$(5.10) \quad \eta_1 \sim \eta_2 \sim 2.1(r - 1)\sqrt{r},$$

$$(5.11) \quad \eta_3 \sim r - 2 \quad \text{if } r - 1 \text{ is even,}$$

$$(5.12) \quad \eta_3 \sim r + 2 \quad \text{if } r - 1 \text{ is odd.}$$

Indeed, dividing the above estimators by the corresponding asymptotic values should result in effectivity indices that are very close to 1.

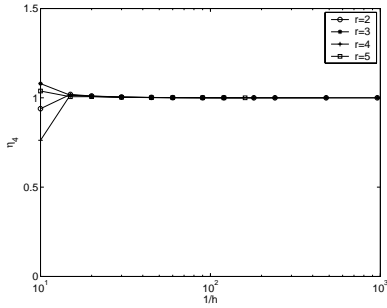


FIG. 5.15. Effectivity index of the nonoverlapping approach-based estimator with  $h' = h$  and  $r' = r + 1$  versus  $h$ .

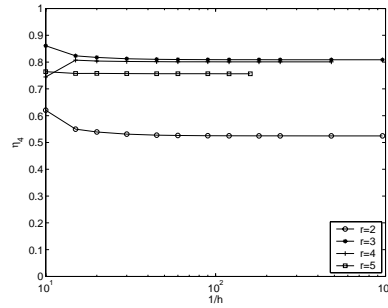


FIG. 5.16. Effectivity index of the nonoverlapping approach-based estimator with  $h' = h/2$  and  $r' = r + 1$  versus  $h$ .

The previous experiments are repeated with the two error estimators based on the nonoverlapping and overlapping domain decomposition approaches. More precisely, we define  $\eta_4$  by

$$(5.13) \quad \eta_4 = \frac{\|\eta\|_{1,h'}}{\|e\|_{1,h'}},$$

where  $\eta = \sum_{K \in \mathcal{T}_h} \eta_K$  and  $\eta_K$  is the solution of the local problem (4.5). Among the various values of parameters that are possible, we chose the two combinations  $h' = h$ ,  $r' = r + 1$  and  $h' = h/2$ ,  $r' = r + 1$ , the results being reported in Figures 5.15 and 5.16, respectively. In the former case, we observe that the index is exactly equal to 1 even for relatively large values of  $h$  and does not depend on  $r$  or on  $r' > r$ . In the latter case, the index is slightly less than 1 and depends on  $r$  and not on  $r' \geq r + 1$ .

In the same way,

$$(5.14) \quad \eta_5 = \frac{\|\eta\|_{1,h'}}{\|e\|_{1,h'}},$$

where  $\eta = \sum_{K \in \mathcal{T}_h} \eta_K$  and  $\eta_K$  is the solution of the local problem (4.21). The results for the case  $h' = h$  and  $r' = r + 1$  are reported in Figure 5.17. We observe that the effectivity index in this case is equal to 2, which is also the number of triangles in a subdomain  $\Omega_e$ . In the case of  $h' = h/2$  and  $r' = r + 1$ , this index is slightly higher than 1, as can be seen in Figure 5.18.

If any conclusions can be drawn after such a limited number of experiments, they would be that, while the estimators  $\eta_4$  and  $\eta_5$  based on the ideas of domain decomposition seem to be very robust, the estimators  $\eta_2$  and  $\eta_3$  are, in contrast, less expensive to implement, and offer the added advantage of being entirely local.

**5.3. Adaptive mesh strategy.** In order to gauge the efficiency of the a posteriori error estimates that we have derived, we present here two  $h$ -adaptive methods for approximating the solution of problem (5.1). We based our numerical experiments on the second estimate, which corresponds to Theorem 3.2(i) and to the effectivity index  $\eta_2$  plotted in Figure 5.11.

Both strategies modify the mesh by refinement of some marked elements while keeping the degree of the polynomials constant. The goal is to generate a mesh in a finite number of steps such that a given tolerance is met by the approximate solution

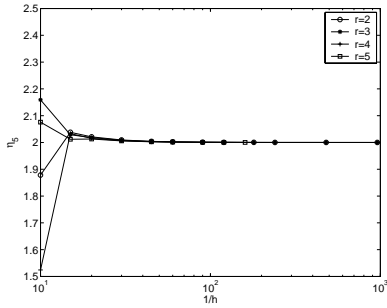


FIG. 5.17. Effectivity index of the overlapping approach-based estimator with  $h' = h$  and  $r' = r + 1$  versus  $h$ .

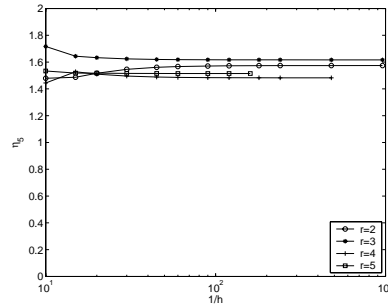


FIG. 5.18. Effectivity index of the overlapping approach-based estimator with  $h' = h/2$  and  $r' = r + 1$  versus  $h$ .

on this mesh. To do this, some optimality criteria have to be imposed. The first technique is based on the convergent adaptive algorithm proposed in [10] for solving Poisson’s equation and used, for instance, in [17]. In order to minimize the total number of degrees of freedom, this strategy equidistributes the given tolerance ( $tol$ ) on each element. Consequently, the local error of the optimal mesh  $\mathcal{T}_h$  satisfies  $\eta_K(u_h^\gamma)^2 \sim \frac{tol^2}{m_h}$ , where  $m_h$  is the number of elements and  $u_h^\gamma$  the discontinuous Galerkin solution on  $\mathcal{T}_h$ . Since the number of iterations required to get this optimal mesh is quite large, we derived a second strategy, which turned out to require less cpu-time. We shall next describe these two strategies and apply them to problem (5.1), with the following exact solution:

$$u(x) = (1 - x) \left( \tan^{-1} \left( \alpha \left( x - \frac{1}{2} \right) \right) + \tan^{-1} \left( \frac{\alpha}{2} \right) \right), \quad \alpha = 100.$$

Given an error tolerance  $tol$  and a coarse mesh  $\mathcal{T}_0$ , let  $u_0^\gamma$  denote the discontinuous Galerkin solution on  $\mathcal{T}_0$ . In this study, for simplicity reasons, we are not considering the effect of data oscillations as in [15]. Let  $k = 0$ . The first strategy involves the following steps:

- (i) compute the local indicator  $\eta_K^k(u_k^\gamma)$  such that

$$(5.15) \quad \eta_K^k(u_k^\gamma)^2 = \frac{h_K^{k+2} \|f + \Delta u_k^\gamma\|_K^2}{r(r-1)^2};$$

- (ii) compute the total error estimate

$$(5.16) \quad \eta^k(u_k^\gamma) = \left( \sum_{K \in \mathcal{T}_k} \eta_K^k(u_k^\gamma)^2 \right)^{1/2};$$

- (iii) select a set  $\hat{\mathcal{T}}_k$  of “marked” elements to be refined such that, for a given parameter  $\theta$  (fixed in our experiments to 0.5),

$$(5.17) \quad \left( \sum_{K \in \hat{\mathcal{T}}_k} \eta_K^k(u_k^\gamma)^2 \right)^{1/2} \geq \theta \eta^k(u_k^\gamma);$$

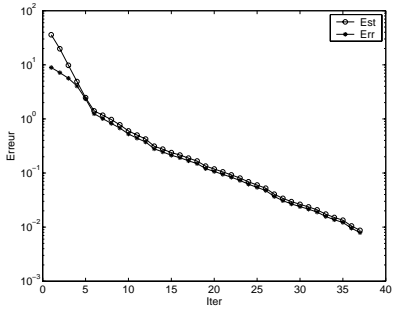


FIG. 5.19. Strategy 1: estimate and exact error versus  $h$ -adaptive iterations ( $r = 2$ ).

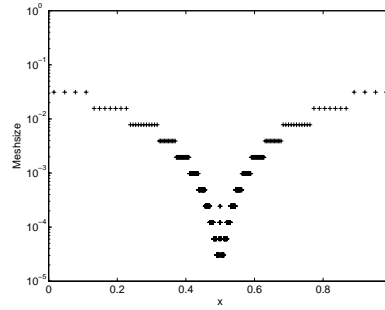


FIG. 5.20. Strategy 1: mesh size versus  $x$  ( $r = 2$ ).

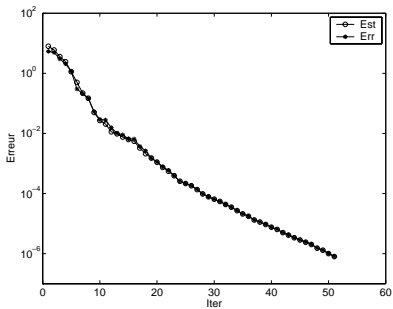


FIG. 5.21. Strategy 1: estimate and exact error versus  $h$ -adaptive iterations ( $r = 4$ ).

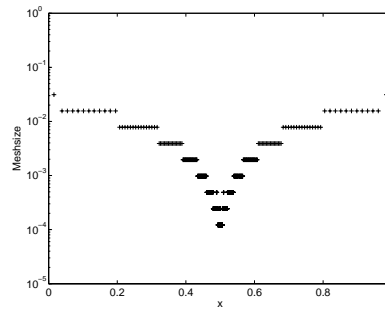


FIG. 5.22. Strategy 1: mesh size versus  $x$  ( $r = 4$ ).

- (iv) obtain a refined mesh  $\mathcal{T}_{k+1}$  by dividing each element  $K \in \hat{\mathcal{T}}_k$  (in two parts for a 1-d problem);
- (v) compute the discontinuous Galerkin solution on  $\mathcal{T}_{k+1}$ ;
- (vi)  $k \leftarrow k + 1$  and go to step (i).

The algorithm is stopped when  $\eta^k(u_k^\gamma) \leq tol$  in step (ii). In practice, for computing marked elements in step (iii), we follow the procedure proposed in [10]. Let us remark that, for changing to the other estimates, formula (5.15) just has to be adapted in step (i).

For  $r = 2$  and  $tol = 0.01$ , this strategy required 37 iterations to reach the optimal mesh, which has 1411 elements and whose distribution of mesh size is plotted in Figure 5.20. For  $r = 4$  and  $tol = 10^{-6}$ , 50 iterations were necessary, the final mesh has 471 elements, and the mesh size distribution is plotted in Figure 5.22. In both cases, the error estimate is an accurate approximation of the exact error (in energy norm), as can be seen in Figures 5.19 and 5.21.

Now let us observe that, to get  $\sum_K \|\nabla e\|_K^2 \leq tol^2$ , it is sufficient to distribute (not to equidistribute) the errors as follows:

$$(5.18) \quad \|\nabla e\|_K \leq \sqrt{\frac{|K|}{|\Omega|}} tol.$$

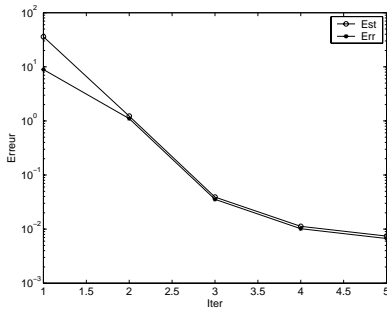


FIG. 5.23. Strategy 2: estimate and exact error versus h-adaptive iterations ( $r = 2$ ).

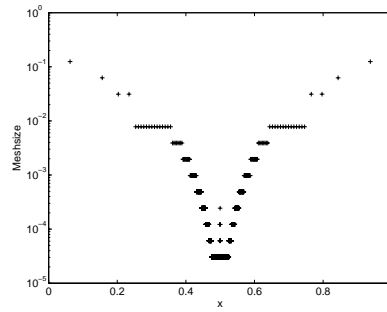


FIG. 5.24. Strategy 2: mesh size versus  $x$  ( $r = 2$ ).

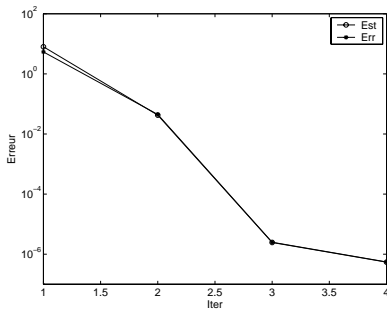


FIG. 5.25. Strategy 2: estimate and exact error versus h-adaptive iterations ( $r = 4$ ).

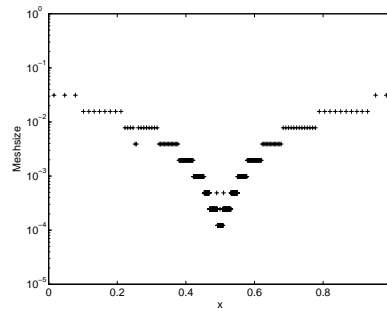


FIG. 5.26. Strategy 2: mesh size versus  $x$  ( $r = 4$ ).

Therefore, we developed a strategy in which an element  $K$ , whose local error estimate  $\eta_K$  is larger than  $\sqrt{|K|/|\Omega|} tol$ , has to be refined as many times as necessary to reduce the local error by the amount  $\sqrt{|K|} tol / \sqrt{|\Omega|} \eta_K$ . Let us recall that from the a priori estimation (2.11) the rate of convergence in the energy norm is  $O(h^{r-1})$ . Thereafter, the number of times the element has to be divided can be estimated to be

$$(5.19) \quad nbr = \frac{\log \left( \frac{\sqrt{|\Omega|} \eta_K}{\sqrt{|K|} tol} \right)}{\log 2^{r-1}}.$$

In one dimension, this is equivalent to determining into how many segments,  $nbs$ , the element  $K$  has to be divided:

$$(5.20) \quad nbs = \left( \frac{\sqrt{|\Omega|} \eta_K}{\sqrt{|K|} tol} \right)^{\frac{1}{r-1}}.$$

The second strategy consists in the following steps:

- (i) compute the local indicator  $\eta_K^k(u_k^\gamma)$  given by (5.15),
- (ii) compute the total error estimate  $\eta^k(u_k^\gamma)$  according to (5.16),
- (iii) compute for each element the nearest power of 2 of  $nbs$  defined in (5.20),
- (iv) obtain a refined mesh  $\mathcal{T}_{k+1}$  by dividing each element by this power of 2 for a 1-d problem,

(v) compute the discontinuous Galerkin solution on  $\mathcal{T}_{k+1}$ ,

(vi)  $k \leftarrow k + 1$  and go to step (i).

The algorithm is stopped when  $\eta^k(u_k^\gamma) \leq \text{tol}$  in step (ii).

For  $r = 2$  and  $\text{tol} = 0.01$ , in only 5 steps this strategy reaches the mesh such that  $\eta^k \leq \text{tol}$ . The cpu-time is then significantly reduced. However, this time the number of elements is not optimal anymore and is equal to 2316 elements. For  $r = 4$  and  $\text{tol} = 10^{-6}$ , we get the given tolerance in 4 iterations, and the final mesh has 573 elements. The distribution of mesh size plotted in Figures 5.24 and 5.26 is almost the same as in the first strategy, and, except for the first iteration, the a posteriori estimate gives a good approximation of the exact error, as shown in Figures 5.23 and 5.25.

## REFERENCES

- [1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [2] D. N. ARNOLD, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.
- [3] D. ARNOLD, F. BREZZI, B. COCKBURN, AND D. MARINI, *Discontinuous Galerkin methods for elliptic problems*, in Proceedings of the International Symposium on the Discontinuous Galerkin Method, B. Cockburn, G. E. Karniadakis, C.-W. Shu, eds., Springer Lecture Notes in Comput. Sci. Engrg. 11, Springer-Verlag, Berlin, 2000, pp. 89–101.
- [4] I. BABUŠKA AND W. C. RHEINBOLDT, *Error estimates for adaptive finite element computations*, SIAM J. Numer. Anal., 15 (1978), pp. 736–754.
- [5] I. BABUŠKA AND W. C. RHEINBOLDT, *A posteriori error estimates for the finite element method*, Internat. J. Numer. Methods Engrg., 12 (1978), pp. 1597–1615.
- [6] G. BAKER, *Finite element methods for elliptic equations using nonconforming elements*, Math. Comp., 31 (1977), pp. 45–59.
- [7] G. A. BAKER, W. N. JUREIDINI, AND O. A. KARAKASHIAN, *Piecewise solenoidal vector fields and the Stokes problem*, SIAM J. Numer. Anal., 27 (1990), pp. 1466–1485.
- [8] R. BECKER, P. HANSBO, AND M. LARSON, *Energy norm a posteriori error estimation for discontinuous Galerkin methods*, Comput. Methods Appl. Mech. Engrg., to appear.
- [9] J. DOUGLAS, JR., AND T. DUPONT, *Interior Penalty Procedures for Elliptic and Parabolic Galerkin Methods*, Lecture Notes in Phys. 58, Springer, Berlin, 1976.
- [10] W. DÖRFLER, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.
- [11] X. FENG AND O. A. KARAKASHIAN, *Two-level additive Schwarz methods for a discontinuous Galerkin approximation of second order elliptic problems*, SIAM J. Numer. Anal., 39 (2001), pp. 1343–1365.
- [12] O. A. KARAKASHIAN AND W. N. JUREIDINI, *A nonconforming finite element method for the stationary Navier–Stokes equations*, SIAM J. Numer. Anal., 35 (1998), pp. 93–120.
- [13] O. KARAKASHIAN AND F. PASCAL, *A Priori and A Posteriori Estimates for Discontinuous Galerkin Method*, Technical report, in preparation.
- [14] M. LARSON AND A. NIKLASSON, *Conservation Properties for the Continuous and Discontinuous Galerkin Methods*, Chalmers Finite Element Center Preprint 2000-08.
- [15] P. MORIN, R. H. NOCHETTO, AND K. G. SIEBERT, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000) pp. 466–488.
- [16] B. RIVIÈRE AND M. WHEELER, *A posteriori error estimates and mesh adaptation strategy for discontinuous Galerkin methods applied to diffusion problems*, Comput. Math. Appl., to appear.
- [17] A. SCHMIDT AND K. G. SIEBERT, *A posteriori estimators for the h-p version of the finite element method in 1D*, Appl. Numer. Math., 35 (2000), pp. 43–66.
- [18] R. VERFÜRTH, *A Review of A Posteriori Error Estimation and Adaptive Mesh Refinement Techniques*, Wiley-Teubner, New York, 1995.
- [19] M. F. WHEELER, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Numer. Anal., 15 (1978), pp. 152–161.
- [20] B. I. WOHLMUTH, *Hierarchical a posteriori error estimators for mortar finite element methods with Lagrange multipliers*, SIAM J. Numer. Anal., 36 (1999), pp. 1636–1658.