STOPPING CRITERIA FOR ADAPTIVE FINITE ELEMENT SOLVERS

MARIO ARIOLI*, EMMANUIL H. GEORGOULIS†, AND DANIEL LOGHIN‡

Abstract. We consider a family of practical stopping criteria for linear solvers for adaptive finite element methods for symmetric elliptic problems. A contraction property between two consecutive levels of refinement of the adaptive algorithm is shown when the a family of smallness criteria for the corresponding linear solver residuals are assumed on each level or refinement. More importantly, based on known and new results for the estimation of the residuals of the conjugate gradient method, we show that the smallness criteria give rise to practical stopping criteria for the iterations of the linear solver, which guarantees that the (inexact) adaptive algorithm converges. A series of numerical experiments highlights the practicality of the theoretical developments.

1. Introduction. Adaptive finite element methods (AFEM), based on a-posteriori error estimates, have become established procedures for computing efficient and accurate approximations to the solution of partial differential equations, especially for the case of elliptic problems (see, e.g., [38, 1] and the references therein). Recent years have seen a growing interest in establishing convergence results for AFEM algorithms. The focus has been mainly second order elliptic problems; for this class of problems results are currently available for conforming finite element methods [17, 32, 8, 37, 35, 14, 33, 10, 7], for classical non-conforming methods [13] and for discontinuous Galerkin methods [27, 25, 9].

The typical algorithm is essentially a simple four-step procedure: solve–estimate–mark–refine. Namely, given a subdivision of the computational domain, one solves the finite element problem (which involves the assembly of the stiffness matrix and the solution of the resulting linear system), then computes an a posteriori error estimate; based on this estimate, a marking strategy is applied to identify for further refinement a set of elements in the current subdivision. Typically, it is assumed that the solution of the linear systems on each level is exact; the few exceptions [35, 20] are discussed below. For large problems, especially in three dimensions, the exact solution of the linear systems may not be obtained efficiently, if at all, by using sparse direct solvers. In such cases, a competitive, or only, alternative is provided by iterative linear solvers, such as the Conjugate Gradient (CG) method, which compute approximations to the exact finite element solution at each iteration of the linear solver. The key issue in this context is accuracy: a highly accurate approximation of the solution to the linear system at each level is inefficient and, most likely, unnecessary, while a poor approximation may affect the quality of the a posteriori estimators, thereby yielding different, possibly extensive refinements. The latter, in turn, affects the quality of the solution at the next step and, potentially, the convergence of the adaptive algorithm itself.

The question of accuracy of an iterative method employed to solve a finite element system of equations has already been considered for the case of a single problem, see, e.g., [2, 16, 4], where a priori convergence properties of FEM gave rise to heuristic stopping criteria for iterative methods employed to solve the linear systems. For a sequence of problems in an adaptive context, we note the recent contributions [26, 34], where a posteriori bounds have been employed as stopping criteria for iterative solvers. Also, in the seminal work [35], the extent of inexactness due to variational crimes (quadrature or inexactness in the solution of the linear system) has been taken into account in the convergence of adaptive algorithms for elliptic problems.

It is now known that, in order to bring the finite element error measured in the energy norm below a certain tolerance in an efficient way, one needs to monitor the norm of the algebraic residual dual to the energy norm. This task is non-trivial. In particular, the algebraic residual of the resulting linear

*Rutherford Appleton Laboratory, Chilton, Didcot, Oxfordshire, OX11 0QX, United Kingdom (M.Arioli@rl.ac.uk).

†Department of Mathematics, University of Leicester, University Road, Leicester, LE1 7RH, United Kingdom (Emmanuil.Georgoulis@mcs.le.ac.uk).

‡School of Mathematics, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom (d.loghin@bham.ac.uk).

system is *not* available during the computation. To this end, in [26, 34, 35], a smallness criterion was proposed requiring that the dual-to-the-energy norm of the algebraic residual is bounded by a multiple of the residual a posteriori error estimators. [34] discusses the practical implementation by providing *estimates* for the dual-to-the-energy norm of the algebraic residual by relating it to the Euclidean norm, or to a preconditioner-induced norm, while [26] employs additionally the Hestenes and Stiefel estimate (see section 5). Therefore, the integration of stopping criteria in an adaptive procedure in [26, 34] has only been effected in an empirical manner. [35] shows that if the above smallness criterion is satisfied, the standard adaptive FEM algorithm converges with optimal complexity. However, in [35], there is no discussion on the practical implementation of the dual-to-the-energy norm of the algebraic residual.

In light of these recent results, it is desirable to provide *practical* stopping criteria for iterative methods which *guarantee* the convergence of the adaptive procedure. The contribution of this work is two-fold. First, it is shown that, given an exact convergent adaptive strategy, one can compute a sequence of inexact solutions which will yield a convergent adaptive algorithm as long as the algebraic residual satisfies a family of abstract certain smallness criteria. This family of smallness criteria, which includes (but extends) the ones considered in [26, 34, 35], involves both the energy-dual norm of the algebraic residual and the a posteriori error estimate. The proof is based on proving a contraction property for the adaptive algorithm in the spirit of [14]. Second, it is shown how to estimate and implement the smallness criteria *practically* within an adaptive context, yielding an automatic stopping criterion for the accuracy of iterative solutions for the adaptive finite element algorithm, where each level is solved iteratively using the Conjugate Gradient method. A crucial development in this work is the incorporation of computable *bounds* of the algebraic residual in the dual-to-the-energy norm. This, together with the careful calculation of the constants appearing in the smallness criterion, yields a *practical convergent adaptive algorithm*. This is in contrast to [2, 16, 4, 26, 34] where estimates (and not, necessarily, bounds) of the algebraic residual in dual-to-the-energy norm are considered, thereby not guaranteeing the theoretical convergence of the adaptive algorithm. Finally, we note the forthcoming work [20], where convergence and quasi-optimality of AFEM with a different stopping criterion is considered.

We shall only consider iterative solution methods of Krylov subspace type. This class of methods has been analyzed extensively over the last three decades and convergence and applicability are largely understood. Furthermore, specific results are available for linear systems arising from finite element discretizations; in particular, it has been shown that convergence should be monitored in the energy norm for reasons of efficiency [2, 16, 4, 3]. This applies even more stringently in the case of adaptive finite element algorithms where the size of the problem grows with each step. Moreover, one expects an adaptive, possibly hierarchical procedure, such as AFEM, to provide certain a posteriori information or even recycling capabilities at each step which would aid the iterative method at the next step. While this is stated in [26], we show that this is indeed the case and that computational efficiency can be dramatically improved.

For the sake of simplicity of exposition we illustrate our approach on a standard symmetric second-order elliptic problem. We stress, however, that generalizations of the results (in various settings) presented in this work are expected to be possible. In particular, in order to focus on the practical implementation of the adaptive algorithm, we do not consider the question of proving optimality of the adaptive algorithm in this work. We believe, however, that the ideas of the proofs of quasi-optimality from [14] can be carried over to the setting considered here also.

The paper is organized as follows. In Section 2 we introduce the model problem and the adaptive finite element algorithm in exact form. Section 3 contains an analysis of the inexact version of the AFEM method. In particular, we derive a sufficient criterion to ensure convergence of the inexact algorithm. Section 4 discusses the practical implementation of this criterion which requires the evaluation of the dual norm of the algebraic residual. This can be achieved by employing the Conjugate Gradient method for solving the linear systems. Finally, in Section 5 we illustrate the efficiency of our approach on a range of two- and three-dimensional problems.

2. Problem formulation. Here, we introduce notation and a description of the model problem together with an posteriori error bound of residual type. We also include a description of an adaptive finite element algorithm together with the corresponding convergence result.

2.1. Model problem and the finite element method. The standard Lebesgue spaces are denoted by $L^p(\omega)$, $1 \leq p \leq +\infty$, $\omega \subset \mathbb{R}^d$; when $p = 2$ the corresponding inner product is denoted by $\langle \cdot, \cdot, \rangle_\omega$ and norm $\| \cdot \|_\omega$; when $\omega = \Omega$, we shall drop the subindex for brevity, writing $\langle \cdot, \cdot, \rangle$ and $\| \cdot \|$, respectively. We also denote by $H_0^1(\omega)$ the standard Sobolev space of functions with zero trace on $\partial\omega$.

Let Ω be a bounded open polyhedral domain in $\mathbb{R}^d$, $d = 2, 3$ and let $\partial\Omega$ denote its boundary. We consider the second order equation

$$-\nabla \cdot (a\nabla u) = f \quad \text{in } \Omega, \tag{2.1}$$

where $a \in [L^\infty(\Omega)]^{d\times d}$ is a positive definite tensor and $f \in L^2(\Omega)$. For simplicity of the presentation, we impose homogeneous Dirichlet boundary condition $u = 0$ on $\partial\Omega$, although this appears not to be an essential restriction. We shall denote by $|| \cdot ||_a := \|\sqrt{a}\nabla(\cdot)\|$ the, so-called, energy norm.

Let $\mathcal{T}$ be a conforming subdivision of Ω into disjoint simplicial elements $\kappa \in \mathcal{T}$. We assume that the subdivision $\mathcal{T}$ is shape-regular (see, e.g., p.124 in [15]) and that it is constructed via affine mappings F_κ, where $F_\kappa : \hat{\kappa} \to \kappa$, with non-singular Jacobian, where $\hat{\kappa}$ is the reference simplex.

For a nonnegative integer r, we denote by $\mathcal{P}_r(\hat{\kappa})$, the set of all polynomials of total degree at most r, defined on $\hat{\kappa}$. We consider the finite element space

$$\mathcal{V} := \{V \in H_0^1(\Omega) : \ V|_\kappa \circ F_\kappa \in \mathcal{P}_r(\hat{\kappa}),\, \kappa \in \mathcal{T}\}. \tag{2.2}$$

By Γ we denote the union of all $(d-1)$-dimensional element faces associated with the subdivision $\mathcal{T}$ (including the boundary). Further we decompose Γ into two disjoint subsets $\Gamma = \partial\Omega \cup \Gamma_{\text{int}}$, where $\Gamma_{\text{int}} := \Gamma\backslash\partial\Omega$. We define $h_\kappa := (\mu_d(\kappa))^{1/d}$, $\kappa \in \mathcal{T}$, where μ_d is the d-dimensional Lebesgue measure. Also, for two (generic) elements κ^+, κ^- sharing a face $e := \partial\kappa^+ \cap \partial\kappa^- \subset \Gamma_{\text{int}}$ we define $h_e := \mu_{d-1}(e)$. We collect these quantities into the element-wise constant function $\mathbf{h} : \Omega \to \mathbb{R}$, with $\mathbf{h}|_\kappa = h_\kappa$, $\kappa \in \mathcal{T}$ and $\mathbf{h}|_e = h_e$, $e \in \Gamma$. The families of meshes constructed by the algorithms presented in this work will be conforming and shape-regular.

We assume throughout that the diffusion tensor a is element-wise constant; this has been done in the interest of simplicity of the exposition only. This restriction can be lifted by combining the a posteriori bounds from [14] (where the case of variable diffusivity is presented) with the developments described below.

The finite element method reads:

$$\text{find } U \in \mathcal{V} \text{ such that} \quad a(U, V) = l(V) \quad \forall V \in \mathcal{V}, \tag{2.3}$$

where the bilinear form $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ and the linear form $l : H_0^1(\Omega) \to \mathbb{R}$ are given by

$$a(w, v) := \int_\Omega a\nabla w \cdot \nabla v \, \mathrm{d}x \qquad \text{and} \qquad l(v) := \int_\Omega fv \, \mathrm{d}x, \tag{2.4}$$

respectively, for $w, v \in H_0^1(\Omega)$.

We note the projection property of the finite element method in the following lemma.

LEMMA 2.1. *Let u be the (weak) solution to problem (2.1) with homogeneous Dirichlet boundary conditions and $U \in \mathcal{V}$ be the finite element solution to (2.3). Then, for any $V \in \mathcal{V}$, we have*

$$||u - U||_a^2 = ||u - V||_a^2 - ||V - U||_a^2. \tag{2.5}$$

Proof. We have, respectively,

$$\begin{aligned} ||u-U||_a^2 &= a(u-U,u-U) = a(u-V,u-V) + a(u-V,V-U) \\ &= ||u-V||_a^2 + a(U-V,V-U) = ||u-V||_a^2 - ||U-V||_a^2, \end{aligned}$$

using Galerkin orthogonality. □

Let now $\{\phi_i\}_{1\le i\le N}$ denote a set of basis functions for $\mathcal{V}$ so that

$$U = \sum_{i=1}^{N} \mathbf{u}_i \phi_i,$$

and let $A_{ij} = a(\phi_j, \phi_i)$, $\mathbf{b}_k = l(\phi_k)$, $i,j,k = 1,\cdots,N$. With this notation, the linear system corresponding to (2.3) is

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

where $A \in \mathbb{R}^{N\times N}$ is the stiffness matrix corresponding to a set of basis functions $\{\phi_i\}_{1\le i\le N}$.

2.2. A posteriori error bounds of residual type. For every face $e \in \Gamma_{\rm int}$, we define the *jump* across e of a scalar function w, defined in an open neighbourhood of e, by

$$[w](x) = \lim_{t\to 0} \Big(w(x - t\mathbf{n}_e) - w(x + t\mathbf{n}_e) \Big),$$

for $x \in e$, where $\mathbf{n}_e$ denotes a normal vector to e. (Note that the jump is only uniquely defined up to a sign, which is unimportant for the discussion below.) For any subset $\mathcal{M} \subset \mathcal{T}$ (i.e., $\mathcal{M}$ is a collection of elements of $\mathcal{T}$), we define the local estimator by

$$\eta_{\mathcal{T}}(U,\mathcal{M}) := \Big(\sum_{\kappa\in\mathcal{M}} \Big(h_\kappa^2 \|f + \nabla\cdot(a\nabla U)\|_\kappa^2 + \sum_{e\in\Gamma_{\rm int}\cap\partial\kappa} h_e \|[a\nabla U\cdot\mathbf{n}_e]\|_e^2 \Big) \Big)^{1/2}.$$

We recall a standard residual-type a posteriori reliability bound of the energy-norm error (see, e.g., [38, 1]).

THEOREM 2.2. *Let $u \in H_0^1(\Omega)$ be the solution to (2.1), with homogeneous Dirichlet boundary conditions, $U \in \mathcal{V}$ be the finite element approximation, associated with the mesh $\mathcal{T}$. Then there exists a positive constant $C_{2.2}$, independent of $\mathcal{T}$, $\mathbf{h}$, u and U, so that*

$$||u-U||_a^2 \le C_{2.2}\, \eta_{\mathcal{T}}^2(U,\mathcal{T}). \tag{2.7}$$

2.3. AFEM algorithms. We describe now briefly the adaptive finite element algorithm analyzed in [35, 14]. Henceforth, all objects indexed by $m \in \mathbb{N}$ refer to the object in the m-th iteration of the adaptive algorithm.

Each iteration of the algorithm comprises four steps which are summarized in the workflow below:

$$\text{SOLVE} \to \text{ESTIMATE} \to \text{MARK} \to \text{REFINE} \tag{2.8}$$

The first step involves computing the finite element solution U_m^e to the problem:

$$\text{find } U_m^e \in \mathcal{V}_m \quad \text{such that} \quad a(U_m^e, V_m) = l(V_m) \quad \forall V_m \in \mathcal{V}_m. \tag{2.9}$$

(The superscript "e" in this section signifies that here we refer to the "exact" algorithm, i.e., all calculations are performed exactly on all levels.)

In the second step, for each element $\kappa \in \mathcal{T}_m$ we calculate the local estimators

$$\eta^2_{\mathcal{T}_m}(U^e_m, \kappa) := h^2_\kappa \|f + \nabla \cdot (a \nabla U^e_m)\|^2_\kappa + \sum_{e \in \Gamma_{\text{int},m} \cap \partial\kappa} h_e \|[a \nabla U^e_m \cdot \mathbf{n}]\|^2_e.$$

The third step identifies a subset $\mathcal{M}_m$ of elements of the mesh $\mathcal{T}_m$ which are marked for refinement, using a Dörfler-type marking strategy (see [17]); more precisely, for a user-defined parameter $0 < \theta \leq 1$ (from now on termed the *Dörfler marking parameter*), we find $\mathcal{M}_m$, subset of $\mathcal{T}_m$, such that

$$\eta^2_{\mathcal{T}_m}(U^e_m, \mathcal{M}_m) \geq \theta \; \eta^2_{\mathcal{T}_m}(U^e_m, \mathcal{T}_m); \tag{2.10}$$

the elements $\kappa \in \mathcal{M}_m$ are called *marked elements*.

In the fourth step, the marked elements are refined by newest vertex bisection (see, e.g., [14] for details); if the resulting mesh is not conforming, it is made into a conforming mesh by sufficient additional refinement (again by newest vertex bisection) of elements possessing hanging nodes. These elements are then added to the marked elements $\mathcal{M}_m$, thereby arriving to the new mesh $\mathcal{T}_{m+1}$.

The adaptive algorithm constructs a sequence of objects $\{\mathcal{T}_m, \mathcal{V}_m, U^e_m\}_{m \geq 1}$, starting with a (given) conforming mesh $\mathcal{T}_0$; the corresponding pseudo-code is given below.

Algorithm 1. AFEM algorithm

Set parameter $0 < \theta \leq 1$. Set $m = 0$.
While convergence criterion not satisfied
1. Solve exactly (2.9) to obtain U^e_m.
2. Compute local estimators $\eta_{\mathcal{T}_m}(U^e_m, \kappa)$, $\kappa \in \mathcal{T}_m$.
3. Mark elements $\mathcal{M}_m$ for refinement in $\mathcal{T}_m$ using (2.10).
4. Refine $\mathcal{M}_m$ to obtain new mesh $\mathcal{T}_{m+1}$. Set $m \leftarrow m + 1$.

End

The above procedure assumes that exact integration is employed and that the solution of the corresponding linear system in the first step is exact at each iteration of the adaptive algorithm. This may not be possible for large problems and the next section describes an inexact version of this algorithm. We conclude this section with the main convergence result in [14].

THEOREM 2.3. *There exist constants $\xi > 0$ and $0 < \alpha < 1$ such that*

$$\|u - U^e_{m+1}\|^2_a + \xi \eta^2_{\mathcal{T}_{m+1}}(U^e_{m+1}, \mathcal{T}_{m+1}) \leq \alpha \left(\|u - U^e_m\|^2_a + \xi \eta^2_{\mathcal{T}_m}(U^e_m, \mathcal{T}_m) \right).$$

The above result can be extended to the inexact case, as shown next.

3. Inexact AFEM algorithms. We introduce now an inexact AFEM algorithm with the same workflow (2.8) as the exact algorithm described above. In general, the subdivisions arising in an inexact context are different from the exact case; we will denote the subdivisions and associated finite element spaces by $\tilde{\mathcal{T}}_m$ and by $\tilde{\mathcal{V}}_m$, respectively.

In the first step we solve inexactly (2.9) to obtain an *inexact finite element solution* $\tilde{U}_m \in \tilde{\mathcal{V}}_m$, which is an approximation to the *exact finite element solution* U_m *on the space* $\tilde{\mathcal{V}}_m$ (i.e., U_m is the solution to the finite element problem (2.3) posed on the finite element space $\mathcal{V} = \tilde{\mathcal{V}}_m$). The theory presented in this section is not dependent on the specific nature of inexactness of $\tilde{U}_m$, so we keep the discussion in an abstract setting.

In the remainder, we shall use the abbreviations $\eta_m(V,\mathcal{M}) \equiv \eta_{\tilde{\mathcal{T}}_m}(V,\mathcal{M})$ for $V \in \tilde{\mathcal{V}}_m$, $\mathcal{M} \subset \tilde{\mathcal{T}}_m$ and $\eta_m(V) \equiv \eta_m(V,\tilde{\mathcal{T}}_m)$, for $m = 0, 1, \ldots$. We shall assume that the inexact solution $\tilde{U}_m$ satisfies the following inequality:

$$||\tilde{U}_{m-1} - U_{m-1}||_a^2 + \mu||\tilde{U}_m - U_m||_a^2 \leq \nu\eta_{m-1}^2(\tilde{U}_{m-1}), \tag{3.1}$$

for some values μ and ν (to be made precise later).

In the second step we calculate local estimators $\eta_m^2(\tilde{U}_m,\kappa)$ based on $\tilde{U}_m$, for each element $\kappa \in \tilde{\mathcal{T}}_m$.

The third step uses the same Dörfler-type marking strategy as in the exact case, with U_m replaced by $\tilde{U}_m$: we find $\tilde{\mathcal{M}}_m$, subset of $\tilde{\mathcal{T}}_m$, such that

$$\eta_m^2(\tilde{U}_m,\tilde{\mathcal{M}}_m) \geq \theta\ \eta_m^2(\tilde{U}_m). \tag{3.2}$$

Finally, the marked elements are refined as in the exact case, by newest vertex bisection

The adaptive algorithm constructs a sequence of objects $\{\tilde{\mathcal{T}}_m, \tilde{\mathcal{V}}_m, \tilde{U}_m\}_{m\geq 1}$, starting with a (given) conforming mesh $\mathcal{T}_0$ and it is summarised as follows.

Algorithm 2. Inexact AFEM

Set parameters $0 < \theta \leq 1$, μ and ν. Initialise $\tilde{U}_0$. Set $m = 1$.
While convergence criterion not satisfied
1. Solve inexactly (2.9) to obtain $\tilde{U}_m$ so that (3.1) is satisfied.
2. Compute local estimators $\eta_{\tilde{\mathcal{T}}_m}(\tilde{U}_m,\kappa)$,$\kappa \in \tilde{\mathcal{T}}_m$.
3. Mark elements $\tilde{\mathcal{M}}_m$ for refinement in $\tilde{\mathcal{T}}_m$ using (2.10).
4. Refine $\tilde{\mathcal{M}}_m$ to obtain new mesh $\tilde{\mathcal{T}}_{m+1}$. Set $m \leftarrow m+1$.

End

3.1. Convergence of inexact adaptive finite element solver. We first need the following lemmata.

LEMMA 3.1. *Let $V, W \in \mathcal{V}$. Then, there exists $C_{3.1} > 1$, depending only on the polynomial degree r and the shape-regularity of the mesh $\mathcal{T}$, such that*

$$\eta^2(V) \leq (1+\gamma)\,\eta^2(W) + C_{3.1}(1+\gamma^{-1})||W - V||_a^2, \tag{3.3}$$

for any $\gamma > 0$.

Proof. Using standard inverse estimates, one can show that

$$\|\mathbf{h}\nabla\cdot(a\nabla(W-V))\|^2 + \|\mathbf{h}^{1/2}[a\nabla(W-V)\cdot\mathbf{n}]\|_{\Gamma_{\rm int}}^2 \leq C||W-V||_a^2, \tag{3.4}$$

for a constant $C > 1$, depending on r and on the shape-regularity of $\mathcal{T}$. Using the triangle inequality, Young's inequality and the bound (3.4), we have

$$\begin{aligned}\eta(V)^2 =& \|\mathbf{h}(f + \nabla\cdot(a\nabla(W-V+V)))\|^2 + \|\mathbf{h}^{1/2}[a\nabla(W-V+V)\cdot\mathbf{n}]\|_{\Gamma_{\rm int}}^2 \\ \leq& (1+\gamma)\big(\|\mathbf{h}(f+\nabla\cdot(a\nabla W)\|^2 + \|\mathbf{h}^{1/2}[a\nabla W\cdot\mathbf{n}]\|_{\Gamma_{\rm int}}^2\big) \\ &+ C(1+\gamma^{-1})||W-V||_a^2.\end{aligned} \tag{3.5}$$

□

The following result is inspired from [14, Corollary 4.4].

LEMMA 3.2. *For any $V_{m+1} \in \mathcal{V}_{m+1}$, we have*

$$\eta_{m+1}^2(V_{m+1}) \leq (1-\tau\theta)(1+\delta)\eta_m^2(\tilde{U}_m) + (1+\delta^{-1})C_{3.1}||V_{m+1} - \tilde{U}_m||_a^2, \tag{3.6}$$

where $\theta \in (0,1)$ is the Dörfler marking parameter, and $\tau = 1 - 2^{-1/d}$.

Proof. Using standard inverse estimates, one can show that

$$\|\mathbf{h}_{m+1}\nabla\cdot(a\nabla(V_{m+1}-\tilde{U}_m))\|^2 + \|\mathbf{h}_{m+1}^{1/2}[a\nabla(V_{m+1}-\tilde{U}_m)\cdot\mathbf{n}]\|^2_{\Gamma_{\mathrm{int},m+1}} \leq C||V_{m+1}-\tilde{U}_m||_a^2, \tag{3.7}$$

for C as in (3.7), with $\mathbf{h}_m$ and $\Gamma_{\mathrm{int},m}$ denoting the mesh-size function and the union of interior faces on the mesh $\tilde{\mathcal{T}}_m$, $m = 0, 1, \ldots$, respectively. Using the triangle inequality, Young's inequality and the bound (3.4), we have

$$\begin{aligned}
\eta_{m+1}^2(V_{m+1}) =& \|\mathbf{h}_{m+1}(f+\nabla\cdot(a\nabla V_{m+1}))\|^2 + \|\mathbf{h}_{m+1}^{1/2}[a\nabla V_{m+1}\cdot\mathbf{n}]\|^2_{\Gamma_{\mathrm{int},m+1}} \\
\leq& (1+\delta)\big(\|\mathbf{h}_{m+1}(f+\nabla\cdot(a\nabla\tilde{U}_m)\|^2 + \|\mathbf{h}_{m+1}^{1/2}[a\nabla\tilde{U}_m\cdot\mathbf{n}]\|^2_{\Gamma_{\mathrm{int},m+1}}\big) \\
&+ (1+\delta^{-1})C||V_{m+1}-\tilde{U}_m||_a^2 \\
=& (1+\delta)\eta_{m+1}^2(V_{m+1}) + (1+\delta^{-1})C||V_{m+1}-\tilde{U}_m||_a^2,
\end{aligned} \tag{3.8}$$

for any $\delta > 0$.

An element $\kappa \in \tilde{\mathcal{M}}_m$ is bisected into elements

$$\mathcal{R}_\kappa := \{\kappa' \in \tilde{\mathcal{T}}_{m+1} : \kappa' \subset \kappa\},$$

due to the refinement strategy of the adaptive algorithm. Observing that, for all $\kappa' \in \mathcal{R}_\kappa$, we have $h_{\kappa'} \leq 2^{-1/d}h_\kappa$, we deduce

$$\begin{aligned}
\eta_{m+1}^2(\tilde{U}_m) &= \eta_m^2(\tilde{U}_m, \tilde{\mathcal{T}}_m \setminus \tilde{\mathcal{M}}_m) + \eta_{m+1}^2(\tilde{U}_m, \{\mathcal{R}_\kappa : \kappa \in \tilde{\mathcal{M}}_m\}) \\
&\leq \eta_m^2(\tilde{U}_m, \tilde{\mathcal{T}}_m \setminus \tilde{\mathcal{M}}_m) + 2^{-1/d}\eta_m^2(\tilde{U}_m, \tilde{\mathcal{M}}_m),
\end{aligned} \tag{3.9}$$

since $[a\nabla\tilde{U}_m\cdot\mathbf{n}] = 0$ almost everywhere on $\Gamma_{\mathrm{int},m+1}\setminus\Gamma_{\mathrm{int},m}$. Combining (3.5) with (3.9), the result readily follows by making use of the marking strategy (2.10). ∎

THEOREM 3.3. *Let u, θ, $\tilde{U}_m$ and $\tilde{U}_{m+1}$, $m \geq 1$ as above, be such that*

$$\mu_1||\tilde{U}_m - U_m||_a^2 + \mu_2||\tilde{U}_{m+1} - U_{m+1}||_a^2 \leq \nu_1\eta_m^2(\tilde{U}_m) + \nu_2\eta_{m+1}^2(\tilde{U}_{m+1}), \tag{3.10}$$

with

$$\mu_1 := \Big(\frac{1+\psi^{-1}}{1+\psi} + \epsilon\xi\big(1+2C_{2.2}C_{3.1}\big) - 1\Big), \quad \mu_2 := \Big(\frac{1+\gamma^{-1}}{(1+\gamma)(1+\psi)} + 1\Big), \tag{3.11}$$

$\nu_1 := \beta_1\xi$, $\nu_2 := \beta_2\xi$, where $0 < \epsilon < 1$, $\xi := \big(C_{3.1}(1+\gamma)(1+\psi)(1+\delta^{-1})\big)^{-1}$, and $0 \leq \beta_i < 1$, $i = 1, 2$, the positive constants γ, δ, ψ and ϵ are chosen so that $(1-\tau\theta)(1+\delta) + 2\epsilon C_{2.2} + \beta_1 + \beta_2 < 1$. Then, there exist a constant $0 < \alpha < 1$, depending only on the shape regularity of $\tilde{\mathcal{T}}_1$ and on the marking parameter θ, such that

$$||u - \tilde{U}_{m+1}||_a^2 + \zeta\eta_{m+1}^2(\tilde{U}_{m+1}) \leq \alpha\big(||u-\tilde{U}_m||_a^2 + \zeta\eta_m^2(\tilde{U}_m)\big), \tag{3.12}$$

for $\zeta := (1-\beta_2)\xi$.

Proof. The projection property of the finite element method (2.5) for the (exact) finite element solution U_m with $V = \tilde{U}_m \in \tilde{\mathcal{V}}_m$, for all $m \geq 1$, reads

$$||u - \tilde{U}_m||_a^2 = ||u - U_m||_a^2 + ||\tilde{U}_m - U_m||_a^2. \tag{3.13}$$

Also, (2.5) for U_{m+1} with $V = \tilde{U}_m$ (noting that $\tilde{U}_m \in \tilde{\mathcal{V}}_{m+1}$, as the adaptive algorithm involves only refinement of elements), for all $m \geq 1$, yielding

$$||u - U_{m+1}||_a^2 = ||u - \tilde{U}_m||_a^2 - ||\tilde{U}_m - U_{m+1}||_a^2. \tag{3.14}$$

Combining (3.13) (for $m+1$) with (3.14), we deduce

$$||u - \tilde{U}_{m+1}||_a^2 = ||u - \tilde{U}_m||_a^2 - ||\tilde{U}_m - U_{m+1}||_a^2 + ||\tilde{U}_{m+1} - U_{m+1}||_a^2. \tag{3.15}$$

The property (2.5) still holds if we interpret U_m as the finite element approximation in $\tilde{\mathcal{V}}_m$ to the variational problem:

$$\text{find } U_{m+1} \in \tilde{\mathcal{V}}_{m+1} \quad \text{such that} \quad a(U_{m+1}, V_{m+1}) = l(V_{m+1}) \quad \forall V_{m+1} \in \tilde{\mathcal{V}}_{m+1}, \tag{3.16}$$

since $\tilde{\mathcal{V}}_m \subset \tilde{\mathcal{V}}_{m+1}$. This, in particular, implies the orthogonality property

$$||U_{m+1} - \tilde{U}_m||_a^2 = ||U_{m+1} - U_m||_a^2 + ||\tilde{U}_m - U_m||_a^2. \tag{3.17}$$

Combining (3.15) with (3.17), we arrive to

$$\begin{aligned} ||u - \tilde{U}_{m+1}||_a^2 =& ||u - \tilde{U}_m||_a^2 + ||\tilde{U}_{m+1} - U_{m+1}||_a^2 \\ &- ||U_{m+1} - U_m||_a^2 - ||\tilde{U}_m - U_m||_a^2. \end{aligned} \tag{3.18}$$

We now focus on estimating the term $\eta_{m+1}^2(\tilde{U}_{m+1})$. The bound (3.6) for $V_{m+1} = \tilde{U}_{m+1}$ yields

$$\eta_{m+1}^2(\tilde{U}_{m+1}) \leq (1 - \tau\theta)(1 + \delta)\eta_m^2(\tilde{U}_m) + (1 + \delta^{-1})C_{3.1}||\tilde{U}_{m+1} - \tilde{U}_m||_a^2. \tag{3.19}$$

Combining the triangle and Young's (i.e., $(a+b)^2 \leq (1+\gamma)a^2 + (1+\gamma^{-1})b^2$ for $\gamma > 0$) inequalities, we deduce

$$\begin{aligned} ||\tilde{U}_m - \tilde{U}_{m+1}||_a^2 \leq& (1 + \gamma^{-1})||U_{m+1} - \tilde{U}_{m+1}||_a^2 + (1 + \gamma)||U_{m+1} - \tilde{U}_m||_a^2 \\ \leq& (1 + \gamma^{-1})||U_{m+1} - \tilde{U}_{m+1}||_a^2 \\ &+ (1 + \gamma)(1 + \psi)||U_{m+1} - U_m||_a^2 \\ &+ (1 + \gamma)(1 + \psi^{-1})||U_m - \tilde{U}_m||_a^2, \end{aligned} \tag{3.20}$$

for $\gamma, \psi > 0$. Using (3.20) to estimate the third term on the right-hand side of (3.19), along with (3.18), and setting $\xi = \big(C_{3.1}(1+\gamma)(1+\psi)(1+\delta^{-1})\big)^{-1}$, we arrive to the bound

$$\begin{aligned} ||u - \tilde{U}_{m+1}||_a^2 + \xi\eta_{m+1}^2(\tilde{U}_{m+1}) \leq& ||u - \tilde{U}_m||_a^2 + \xi(1 - \tau\theta)(1 + \delta)\eta_m^2(\tilde{U}_m) \\ &+ \mu_2||\tilde{U}_{m+1} - U_{m+1}||_a^2 \\ &+ \Big(\frac{1 + \psi^{-1}}{1 + \psi} - 1\Big)||\tilde{U}_m - U_m||_a^2. \end{aligned} \tag{3.21}$$

To estimate further the first term on the right-hand side of (3.21), we use (3.13), working as follows:

$$\begin{aligned} ||u - \tilde{U}_m||_a^2 =& (1 - \epsilon\xi)||u - \tilde{U}_m||_a^2 + \epsilon\xi||u - U_m||_a^2 + \epsilon\xi||\tilde{U}_m - U_m||_a^2 \\ \leq& (1 - \epsilon\xi)||u - \tilde{U}_m||_a^2 + \epsilon\xi C_{2.2}\eta_m^2(U_m) + \epsilon\xi||\tilde{U}_m - U_m||_a^2 \\ \leq& (1 - \epsilon\xi)||u - \tilde{U}_m||_a^2 + 2\epsilon\xi C_{2.2}\eta_m^2(\tilde{U}_m) + \epsilon\xi\big(1 + 2C_{2.2}C_{3.1}\big)||\tilde{U}_m - U_m||_a^2, \end{aligned} \tag{3.22}$$

for any $0 < \epsilon < 1$, using Theorem 2.2 and (3.3) (with $V = U_m$, $W = \tilde{U}_m$ and $\gamma = 1$). Applying (3.22) on (3.21), we arrive to

$$\begin{aligned} ||u - \tilde{U}_{m+1}||_a^2 + \xi\eta_{m+1}^2(\tilde{U}_{m+1}) \leq (1 - \epsilon\xi)||u - \tilde{U}_m||_a^2 \\ + \xi\big((1 - \tau\theta)(1 + \delta) + 2\epsilon C_{2.2}\big)\eta_m^2(\tilde{U}_m) \\ + \mu_1||\tilde{U}_m - U_m||_a^2 + \mu_2||\tilde{U}_{m+1} - U_{m+1}||_a^2. \end{aligned} \tag{3.23}$$

Hypothesis (3.10) then implies

$$\begin{aligned} ||u - \tilde{U}_{m+1}||_a^2 + \xi\eta_{m+1}^2(\tilde{U}_{m+1}) \leq (1 - \epsilon\xi)||u - \tilde{U}_m||_a^2 + \beta_2\xi\eta_{m+1}^2(\tilde{U}_{m+1}) \\ + \xi\big((1 - \tau\theta)(1 + \delta) + 2\epsilon C_{2.2} + \beta_1\big)\eta_m^2(\tilde{U}_m). \end{aligned} \tag{3.24}$$

for $0 \leq \beta_i < 1$, $i = 1, 2$. Using the hypothesis that the positive constants γ, δ, ψ and ϵ are chosen so that $(1 - \tau\theta)(1 + \delta) + 2\epsilon C_{2.2} + \beta_1 + \beta_2 < 1$, we arrive to

$$||u - \tilde{U}_{m+1}||_a^2 + (1 - \beta_2)\xi\eta_{m+1}^2(\tilde{U}_{m+1}) \leq \alpha||u - \tilde{U}_m||_a^2 + \alpha(1 - \beta_2)\xi\eta_m^2(\tilde{U}_m), \tag{3.25}$$

choosing

$$\alpha := \max\big\{1 - \epsilon\xi, \frac{(1 - \tau\theta)(1 + \delta) + 2\epsilon C_{2.2} + \beta_1}{1 - \beta_2}\big\}.$$

□

We note that this is a generic result of inexactness in the computation of the solutions of each step of AFEM. Inexactness can result from a number of reasons, such as inexact linear solvers, quadrature errors, etc. The first result on convergence of AFEM taking into account inexactness appeared in [35]. The criterion (3.10) presented above includes the case of the respective criterion of admissible inexactness in [35] (possibly with different constants), by selecting $\mu_1 = 0$ and $\nu_1 = 0$.

It is important to note, however, that (3.10) is not an a-posteriori criterion as such, as it involves knowledge of U_m, which is not available in practice. For the case of inexactness due to the approximate solution of the linear systems, (3.10) involves algebraic errors alone. As such, it represents essentially an *adaptive stopping criterion* for an iterative method, provided we can bound these algebraic errors from above. In the next section we show how (3.10) can be implemented in a guaranteed fashion for the Conjugate Gradient method, thereby yielding a practical convergent AFEM with inexact solvers.

We conclude this section by noting a practically important case of Theorem 3.3.

COROLLARY 3.4. *Let u, $\tilde{U}_m$ and $\tilde{U}_{m+1}$, $m \geq 1$ as above, be such that*

$$||\tilde{U}_m - U_m||_a^2 + \mu||\tilde{U}_{m+1} - U_{m+1}||_a^2 \leq \nu\eta_m^2(\tilde{U}_m), \tag{3.26}$$

with

$$\mu := \frac{1 + \xi C_{3.1}(1 + \gamma^{-1})}{\epsilon\xi\big(1 + 2C_{2.2}C_{3.1}\big)}, \quad \nu := \frac{\beta}{\epsilon\big(1 + 2C_{2.2}C_{3.1}\big)}, \tag{3.27}$$

where $0 < \epsilon < 1$, $\xi := \big(2C_{3.1}(1 + \gamma)(1 + \delta^{-1})\big)^{-1}$, and β, γ, δ and ϵ are chosen small enough, so that $(1 - \tau\theta)(1 + \delta) + 2\epsilon C_{2.2} + \beta < 1$. Then, there exist a constant $0 < \alpha < 1$, depending only on the shape regularity of $\tilde{\mathcal{T}}_1$ and on the marking parameter θ, such that

$$||u - \tilde{U}_{m+1}||_a^2 + \xi\eta_{m+1}^2(\tilde{U}_{m+1}) \leq \alpha\big(||u - \tilde{U}_m||_a^2 + \xi\eta_m^2(\tilde{U}_m)\big). \tag{3.28}$$

Proof. The proof follows immediately from Theorem 3.3, by setting $\psi = 1$ and $\beta_1 = b$ and $\beta_2 = 0$. ∎

This result may appear somewhat surprising, in that it is sufficient to control the inexactness error at step k of AFEM from the a-posteriori error estimator of step $k-1$. Apart from potential theoretical implications, this has the practical advantage in that, to assess the validity of (3.26), there is no need to compute the a-posteriori estimator again after each linear solver iteration, as is the case when $\nu_2 \neq 0$ in (3.10).

We note that (3.26) may not be satisfied for each step of AFEM. If this is the case, one can revert to (3.10) with $\nu_2 \neq 0$ for the next step of AFEM. Nevertheless, in all the numerical results presented below (3.26) is always satisfied.

4. Stopping criteria for the Conjugate Gradient method. We now turn to the practical implementation of criterion (3.10). Our choice of solver is the Conjugate Gradient (CG) method, due to its convenient properties (see below). At each step m we need to solve iteratively a linear system of the form (2.6)

$$A_m \mathbf{u}_m = \mathbf{b}_m. \tag{4.1}$$

The matrices A_m have size $N_m \times N_m$ with $\{N_m\}_m$ an increasing sequence. We denote by $\mathbf{u}_m^k \in \mathbb{R}^{N_m}$ the k-th CG iterate at step m of the adaptive algorithm and by U_m^k the corresponding function in $\tilde{\mathcal{V}}_m$. We denote the residual by $\mathbf{r}_m^k := \mathbf{b}_m - A_m \mathbf{u}_m^k$ and note that the energy norm of the error can be expressed as a dual norm of the residual:

$$\|U_m - U_m^k\|_a = \|\mathbf{u}_m - \mathbf{u}_m^k\|_{A_m} = \|\mathbf{r}_m^k\|_{A_m^{-1}},$$

where $\langle x, y\rangle_A := x^T A y$, $x, y \in \mathbb{R}^N$, $A \in \mathbb{R}^{N\times N}$, denotes the standard inner product weighted by A in $\mathbb{R}^N$, with the corresponding norm $\|x\|_A := \sqrt{\langle x, x\rangle_A}$. A candidate for $\tilde{U}_m$ is U_m^k for any k for which criterion (3.10) is satisfied. Our aim is to find an automatic way of choosing k so that the overall solution process is computationally efficient.

It is well-known that the Conjugate Gradient method minimises the energy norm of the error, namely

$$\mathbf{u}_m^k = \arg \min_{\mathbf{u} \in \mathcal{K}_k(\mathbf{r}_m^0, A_m)} \|\mathbf{u}_m - \mathbf{u}\|_{A_m},$$

where $\mathcal{K}_k(\mathbf{r}_m^0, A_m) := \left\{\mathbf{r}_m^0, A_m \mathbf{r}_m^0, \cdots, A_m^{k-1}\mathbf{r}_m^0\right\}$ is the Krylov subspace of dimension k. Thus, the energy norm of the error decreases monotonically and criterion (3.10) will be satisfied for all U_m^k with $k > k^*$ for some k^*. This makes CG an attractive choice for enforcing criterion (3.10) efficiently. In addition, there are various established numerical techniques that provide bounds or estimates for the energy norm of the error at each step. We note that these properties do not hold in general, and that for non-symmetric problems, the best choice of iterative method remains unclear.

4.1. Error bounds for the Conjugate Gradient method. The CG method is included below for the generic problem

$$A\mathbf{u} = \mathbf{b}.$$

Algorithm 3. Conjugate Gradient Algorithm

Set $\mathbf{u}^0 := 0; \mathbf{p}^0 := \mathbf{r}^0 := \mathbf{b}; \sigma_0 := \|\mathbf{r}^0\|^2$;
For $j = 0, 1, \ldots$ until convergence do
  $\mathbf{v}^j = A\mathbf{p}^j$; $\gamma_j = \sigma_j/(\mathbf{r}^j \cdot \mathbf{v}^j)$;
  $\mathbf{u}^{j+1} = \mathbf{u}^j + \gamma_j \mathbf{p}^j$; $\mathbf{r}^{j+1} = \mathbf{r}^j - \gamma_j \mathbf{u}^j$; $\sigma_{j+1} = \|\mathbf{r}^{j+1}\|^2$;
  $\chi_{j+1} = \sigma_{j+1}/\sigma_j$; $\mathbf{p}^{j+1} = \mathbf{r}^{j+1} + \chi_{j+1}\mathbf{p}^j$;
End

The above algorithm constructs implicitly a Lanczos tridiagonalisation which we write in the form

$$V_k^T A V_k = T_k,$$

where V_k is an orthogonal matrix and T_k is a symmetric and positive tridiagonal matrix with entries computable from the CG coefficients. The explicit form of $T_k \in \mathbb{R}^{k\times k}$ is given below

$$T_k = \begin{pmatrix} \alpha_1 & \beta_1 & & & 0 \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & & \alpha_{k-1} & \beta_{k-1} \\ 0 & & & \beta_{k-1} & \alpha_k \end{pmatrix}.$$

where for $j = 1, \ldots, k$,

$$\alpha_j = \frac{1}{\gamma_{j-1}} + \frac{\chi_{j-1}}{\gamma_{j-2}}, \quad \beta_j = \frac{\sqrt{\chi_j}}{|\gamma_{j-1}|},$$

with $\gamma_{-1} = 1, \chi_0 = 0$.

Several authors have proposed rules that compute error bounds for the Conjugate Gradient method [5, 6, 11, 12, 21, 22, 23, 29, 30, 31, 36]. Some of these rules compute estimates of the error in the Euclidean norm and others compute estimates related to the energy norm. We review below some of the existing results; we also introduce new estimates suitable to the adaptive finite element context.

4.1.1. The Hestenes and Stiefel estimate. In their seminal paper, Hestenes and Stiefel [24] propose a method to estimate the energy norm error that uses the values computed during the CG procedure. Strakoš and Tichý, [36], study the relation between the estimates proposed in [24, 21, 23, 29, 30, 31] and they prove that the Hestenes-Stiefel estimate [24] is numerically stable. The method uses the fact that the error can be written as a linear combination of residual norms:

$$\|\mathbf{u} - \mathbf{u}^k\|_A^2 = \sum_{k+1}^{N} \gamma_j \|\mathbf{r}^j\|^2.$$

Under the assumption that $e_A^{(k+d)} << e_A^{(k)}$, where the integer d denotes a suitable delay, the Hestenes and Stiefel estimate is given by the formula

$$\|\mathbf{u} - \mathbf{u}^k\|_A^2 \approx \sum_{j=k+1}^{k+d} \gamma_j \|\mathbf{r}^j\|^2. \tag{4.2}$$

In [21] $d = 10$ is indicated as a successful compromise, and numerical experiments support this conclusion [21, 22, 2]. In Section 5, we will indicate that the cheaper choice $d = 5$ can be reliable if the solution u of (2.1) is reasonably regular; in general, one can expect d to be required to be larger for ill-conditioned problems.

4.1.2. The Golub and Meurant bounds. The A-norm of the error at each CG step can be written in the following way, using the orthogonality $\mathbf{r}_k^T \mathbf{u}^k = 0$,

$$\|\mathbf{u} - \mathbf{u}^k\|_A^2 = \|\mathbf{r}_k\|_{A^{-1}}^2 = \mathbf{b}^T A^{-1} \mathbf{b} - \mathbf{b}^T \mathbf{u}^k. \tag{4.3}$$

Thus, the main difficulty in evaluating the above quantity is in the evaluation of the first term on the right-hand side of (4.3). This term can be written as

$$F(A) = \mathbf{b}^T A^{-1}\mathbf{b} = \int_{\lambda_{\min}(A)}^{\lambda_{\max}(A)} \lambda^{-1} \mathrm{d}\omega(\lambda),$$

where the measure ω is a non-decreasing step function with jump discontinuities depending on the Fourier coefficients of $\mathbf{b}$ at the eigenvalues of A. Golub and Meurant used this formulation to provide upper and lower bounds on the CG error, by employing Gauss-Radau and Gauss-Lobatto quadrature rules, respectively [21], [22]. The latter approach can be shown to be equivalent to the Hestenes and Stiefel estimate above.

The only guaranteed upper bound for the A-norm of the CG error uses a Gauss-Radau quadrature associated with the measure ω and with one node prescribed at $\lambda < \lambda_{\min}(A)$. Let

$$\hat{T}_{k+1} = \begin{pmatrix} \alpha_1 & \beta_1 & & & 0 \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & & \alpha_k & \beta_k \\ 0 & & & \beta_k & \hat{\alpha}_{k+1} \end{pmatrix}.$$

where

$$\hat{\alpha}_{k+1} = \lambda + \beta_k^2 \mathbf{e}_k^T (T_k - \lambda I_k)^{-1} \mathbf{e}_k$$

with $\mathbf{e}_k$ the k-th column of the $k \times k$ identity matrix. Assuming $0 < \lambda < \lambda_{\min}(A)$, the Cholesky decomposition $\hat{T}_{k+1} = \hat{R}_{k+1}^T \hat{R}_{k+1}$ can be shown to exist. Let now $\hat{\mathbf{y}}^{k+1}$ be the solution of

$$\hat{R}_{k+1}^T \hat{\mathbf{y}}^{k+1} = \|\mathbf{b}\| \hat{\mathbf{e}}_1,$$

where $\hat{\mathbf{e}}_1$ denotes the first column of the identity matrix of size $k+1$. Then an upper bound on the CG error is given by [21, 22]

$$\|\mathbf{u} - \mathbf{u}_k\|_A \leq \left| \hat{\mathbf{y}}_{k+1}^{k+1} \right|. \tag{4.4}$$

It is clear that in order to compute this bound, the lower bound λ is required. In fact, experiments show that a close lower bound on the smallest eigenvalue of A yields tight upper bounds for the CG error [31]. We consider this issue at the end of this section.

4.1.3. Other estimates. Another estimate for the CG error was derived in [11]. The derivation employs an anti-Gauss rule to evaluate the integral $F(A)$. We include a brief description here. Let

$$\tilde{T}_k = \begin{pmatrix} \alpha_1 & \beta_1 & & & & & 0 \\ \beta_1 & \alpha_2 & \beta_2 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & & \alpha_{k-2} & \beta_{k-2} & & \\ & & & \beta_{k-2} & \alpha_{k-1} & \sqrt{2}\beta_{k-1} & \\ 0 & & & & \sqrt{2}\beta_{k-1} & \alpha_k & \end{pmatrix}.$$

Assuming $\tilde{T}_k$ is positive definite, we consider as before the Cholesky factorisation $\tilde{T}_k = \tilde{R}_k^T \tilde{R}_k$ and similarly we denote by $\mathbf{y}^k$, $\tilde{\mathbf{y}}^k$ the solution of the linear systems

$$R^k \mathbf{y}^k = \|\mathbf{b}\| \mathbf{e}_1, \quad \tilde{R}_k^T \tilde{\mathbf{y}}^k = \|\mathbf{b}\| \mathbf{e}_1,$$

where $T_k = R_k^T R_k$. One can use these vectors to derive an upper bound on the error, under the additional assumption of rapid decay of the coefficients of $f(t) = 1/t$ in the expansion in terms of orthonormal polynomials with respect to the inner product $(g,h) := \int_0^\infty gh\mathrm{d}\omega$ (see [11] for details). The bound reads

$$\|\mathbf{u} - \mathbf{u}_k\|_A \leq \left[(\tilde{\mathbf{y}}_k^k + \mathbf{y}_k^k)(\tilde{\mathbf{y}}_k^k - \mathbf{y}_k^k)\right]^{1/2}. \tag{4.5}$$

We remark here that the above expression may not exist if $\tilde{T}_k$ is not positive definite; moreover, even if it exists, it is not guaranteed to be an upper bound.

4.2. Adaptive stopping criteria for CG. Criteria (3.10) (or (3.26)) cannot be employed in a practical context. Instead, the following generic criteria will be considered:

$$\mu_1 E(\tilde{U}_m)^2 + \mu_2 E(\tilde{U}_{m+1})^2 \leq \nu_1 \eta_m^2(\tilde{U}_m) + \nu_2 \eta_m^2(\tilde{U}_{m+1}), \tag{4.6}$$

and

$$E(\tilde{U}_m)^2 + \mu E(\tilde{U}_{m+1})^2 \leq \nu \eta_m^2(\tilde{U}_m), \tag{4.7}$$

where $E(\tilde{U}_m)$ denotes an estimate or bound for the error $\|U_m - \tilde{U}_m\|_a$. Note that if $E(\tilde{U}_m)$ is an upper bound, then the result of Theorem 3.3 (and of Corollary 3.4) hold and the inexact AFEM algorithm is guaranteed to converge. In general, estimates will not provide this guarantee, though a tight estimate or lower bound could also ensure the contraction result of Theorem 3.3, possibly at a different rate. For such cases, further analysis is required.

We discuss now the only *guaranteed* upper bound available – the Golub and Meurant bound (4.4) for the case where a in model problem (2.1) has minimum eigenvalue $a_{\min}$. As described above, this bound requires a lower bound on the smallest eigenvalue of the system matrix. This information is not readily available and, in general, it can be expensive to compute. We introduce below two bounds and an estimate for this quantity.

4.2.1. Eigenvalue bounds based on Poincaré inequalities.. To find a lower bound on the smallest eigenvalue of a discrete Laplacian, one could employ the Poincaré inequality

$$\|v\|_{L^2(\Omega)} \leq C_P |v|_{H^1(\Omega)}, \qquad \forall v \in \mathcal{V}_m \subset H^1_\Gamma(\Omega),$$

where $\Gamma \subseteq \partial\Omega$ is Lipschitz continuous. The constant $C_P = C_P(\Omega)$ depends on the domain Ω only and can be estimated for polygonal domains. Since

$$\|v\|_a^2 \geq a_{\min} |v|^2_{H^1(\Omega)}$$

the Poincaré inequality yields the following lower bound on Rayleigh quotients involving A_m:

$$\frac{a_{\min}}{C_P^2} \frac{\mathbf{v}^T M_m \mathbf{v}}{\mathbf{v}^T \mathbf{v}} \leq \frac{\mathbf{v}^T A_m \mathbf{v}}{\mathbf{v}^T \mathbf{v}}, \quad \forall \mathbf{v} \in \mathbb{R}^{N_m},$$

where M_m is the mass matrix corresponding to the m level of the AFEM algorithm. We conclude that

$$\frac{a_{\min}}{C_P^2} \lambda_{\min}(M_m) \leq \lambda_{\min}(A_m)$$

and the task is now to find a lower bound on the smallest eigenvalue of M_m. This is easily done in a finite element context as shown in [18], [19], [39] and we include the result here

$$\min_{\kappa \in \mathcal{T}_m} \lambda_{\min}(M_\kappa) \leq \lambda_{\min}(M_m), \tag{4.8}$$

where M_κ is the elemental mass matrix assembled on the element κ. The resulting bound on the smallest eigenvalue of A_m is

$$\lambda_{\min}(A_m) \geq \frac{a_{\min}}{C_P^2} \min_{\kappa \in \mathcal{T}_m} \lambda_{\min}(M_\kappa); \tag{4.9}$$

we call this bound *the global Poincaré bound*. Bound (4.8) is remarkably tight for matrices assembled on quasi-uniform or isotropic subdivisions. However, it can be fairly poor in the case of adaptive refinements. Improvements are described in [28]. We consider below an adaptive procedure which refines the Poincaré bound (see [28] for other possibilities).

Let $\tilde{\mathcal{M}}_m$ be the set of triangles marked for bisection at step m and let $\mathcal{D}_m \supseteq \tilde{\mathcal{M}}_m$ denote the region comprising all the refined elements. Let us further assume that following condition holds:

$$\partial\Omega \neq \mathcal{D}_m \cap \partial\Omega \neq \emptyset. \tag{4.10}$$

If $\mathcal{D}_m$ is a multiply-connected set, then we will assume that the above condition holds for each disjoint region it is comprised of. The following result can be found in [28].

PROPOSITION 4.1. *Let $\mathcal{D}_m$ satisfy* (4.10) *and let C_P^m denote its Poincaré constant. Let A_m, A_{m+1} denote the matrices assembled on consecutive subdivisions in the inexact AFEM algorithm 2. Then*

$$\lambda_{\min}(A_{m+1}) \geq \min\left\{\lambda_{\min}(A_m), \frac{a_{\min}}{(C_P^m)^2} \min_{\kappa \in \mathcal{D}_m} \lambda_{\min}(M_\kappa)\right\}. \tag{4.11}$$

The bound (4.11) is essentially an *adaptive Poincaré* lower bound for the smallest eigenvalue of the stiffness matrix. Given its form, the above bound is updated at each step, by taking the minimum of the current lower bound and the bound expressing the Poincaré inequality on the region $\mathcal{D}_m$.

REMARK 4.2. *The advantage of this approach is that the lower bound on $\lambda_{\min}(\tilde{A}_{m+1})$ is tight since it employs a smaller Poincaré constant, corresponding to the smaller region $\mathcal{D}_m$ which in addition has an isotropic subdivision.*

REMARK 4.3. *If the region $\mathcal{D}_m$ is multiply-connected, with each subregion satisfying condition* (4.10)*, the same approach applies, with the second term in* (4.11) *replaced by the set of bounds corresponding to each subregion.* If the region $\mathcal{D}_m$ does not satisfy (4.10), one can employ the following result, also taken from [28].

PROPOSITION 4.4. *Let A_m, A_{m+1} denote the stiffness matrices assembled on consecutive subdivisions in the inexact AFEM algorithm 2. Let $\tilde{A}_{m+1}, \tilde{M}_{m+1}$ denote the stiffness and mass matrices assembled on the subdivision corresponding to $\mathcal{D}_m$. Then for all $\chi > 0$ there holds*

$$\lambda_{\min}(A_{m+1}) \geq \min\left\{\lambda_{\min}(A_m), \frac{a_{\min}}{1 + \chi C_P^2} \lambda_{\min}(\tilde{A}_{m+1} + \chi \tilde{M}_{m+1})\right\}. \tag{4.12}$$

The above result requires that we update recursively the lower bound on the smallest eigenvalue of the system matrix by estimating the smallest eigenvalue of the smaller augmented matrix $\tilde{A}_{m+1} + \chi \tilde{M}_{m+1}$. This task may be achieved for larger sizes by using a domain decomposition approach: the region $\mathcal{D}_m$ is partitioned in several non-overlapping regions; each subdomain will provide a contribution for the assembly of $\tilde{A}_{m+1} + \chi \tilde{M}_{m+1}$ and it can be shown that the smallest minimum eigenvalue of all these smaller matrices is a lower bound on $\lambda_{\min}(\tilde{A}_{m+1} + \chi \tilde{M}_{m+1})$. For more details, see [28].

4.2.2. Estimates using the Lanczos algorithm. It is known that the underlying Lanczos tridiagonalisation constructed by the CG algorithm provides also estimates to the eigenvalues of the system matrix. In particular, there holds $\lambda_{\min}(A) \leq \lambda_{\min}(T_k)$ with the bound getting tighter with growing k. This suggests the following estimate at step $m+1$

$$\lambda := \lambda^{m+1} := c\lambda_{\min}(T_{k+1}^m), \tag{4.13}$$

with $c < 1$ a constant that will account for the reduction in $\lambda_{\min}(A)$ with refinement but also for the poor approximation of $\lambda_{\min}(A_m)$ by $\lambda_{\min}(T^m_{k+1})$. If corresponding to this choice of λ the Cholesky factorisation of $\hat{T}_{k+1}$ does not exist, we reduce c until the factorisation and hence (4.4) exist. Note that this is always possible.

We end this section with a remark regarding preconditioning. The implementation of the stopping criteria presented above can be extended to the case where a preconditioning routine is employed to accelerate convergence of the Conjugate Gradient method. Most criteria can be adapted directly to preconditioning contexts; however, the Golub and Meurant bounds require estimates of the smallest eigenvalue of the *preconditioned system*, which may not be straightforward to obtain. Some results regarding this situation can be found in [4]. We note here that the estimate (4.13) using the Lanczos algorithm generalises naturally in this case.

5. Numerical experiments. We investigate now the usefulness of criterion (3.10) (and, in particular, the case of (3.26)) and corresponding approximations of the form (4.7) as they are applied to the CG method to solve some standard adaptive finite element problems. We are interested in the following comparisons:

- 'inexact' vs. 'exact' meshes;
- comparison of energy errors for exact and inexact cases;
- comparison of order of convergence for exact and inexact cases;
- comparison with Euclidean stopping criteria;
- comparison between bounds and estimators of the CG error.

We used both the exact and inexact versions of the AFEM algorithm with various starting meshes. We initialised the algorithms with the exact solution on the coarsest mesh. We used for illustration purposes the ideal criterion (3.26); we also employed a number of practical approximations to the ideal criterion of the form (4.6) corresponding to various bounds or estimators for the CG error. We summarise these below.

1. DNR: the ideal bound (3.26) using the exact dual norm of the residual;
2. GM0: the ideal Golub-Meurant upper bound (4.4) using the exact $\lambda_{\min}(A_m)$;
3. GM1: the Golub-Meurant upper bound (4.4) with adaptive bounds (4.11), (4.12) for $\lambda_{\min}(A_m)$;
4. GM2: the Golub-Meurant upper bound (4.4) with global Poincaré bound (4.9) for $\lambda_{\min}(A_m)$;
5. GM3: the Golub-Meurant criterion (4.4) with the estimator (4.13) for $\lambda_{\min}(A_m)$ with $c = 1/2$;
6. HS: the Hestenes-Stiefel estimator (4.2) with a delay of $d = 5$ steps.
7. AG: the anti-Gauss estimator (4.5);
8. ER($|\log tol|$): the standard Euclidean residual with various stopping tolerances tol.

We note that only the first two criteria are of theoretical interest and are not available generally in a practical context. The following two criteria are the only guaranteed upper bounds for the CG error and thus the only bounds for which the convergence result of Thm 3.3 applies. Criteria 5-7 are estimators, while the criterion based on the Euclidean residual is also an empirical estimator. In order to compare in a fair and explicit fashion the performance of the CG algorithm equipped with the above stopping criteria we chose to transform the computational cost on each level into units corresponding to matrix-vector products on the last level. Since each CG iteration has a cost proportional to the number of nonzeros in the system matrix, the formula employed (and the tabulated variable) is

$$\text{mv} := \text{matvecs}(m) = \sum_{k=1}^{m} \frac{\text{nnz}(A_k)}{\text{nnz}(A_m)} \cdot \text{its}(k), \tag{5.1}$$

where $\text{nnz}(A_k)$ denotes the number of nonzero elements of A_k and $\text{its}(k)$ represents the number of CG iterations on level k. We employed the "practical" version of the criterion (3.10) from Corollary 3.4 with $C_{3.1} = 10$, $C_{2.2} = 40$, $\gamma = \tau\theta/2, \delta = \tau\theta/4, \epsilon = 10^{-3}\tau\theta/C_{3.1}$, $\beta = \frac{1}{2}(1 - 2\epsilon C_{3.1} - (1-\tau\theta)(1+\gamma)(1+\delta))$, yielding $\mu = 7.14 \cdot 10^4, \nu = 2.44$ in two dimensions and $\mu = 1.38 \cdot 10^5, \nu = 2.17$ in three dimensions in

(3.26). We remark here that while this choice is non-unique, the numerical results presented below appear to be qualitatively similar to the case where criteria corresponding to other constants are employed.

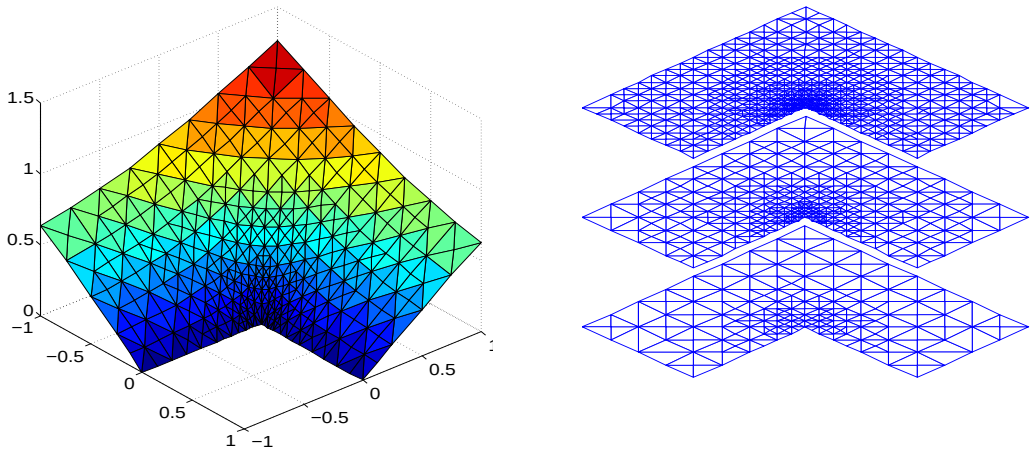
Finally, another important ingredient in our iterative solver is the starting guess. In all cases this choice was provided by the solution obtained at the previous refinement level interpolated onto the current (refined) level.

5.1. Experiments in two dimensions.

5.1.1. Test Problem 1. We solved problem (2.1) on an L-shaped domain with the right-hand side chosen so that the analytic solution is $u = r^{2/3} \sin \frac{2\theta}{3}$. The AFEM algorithms were implemented with a Dörfler marking parameter $\theta = 0.75$. The solution and typical meshes are shown in Figure 5.1. The results are displayed in Table 5.1. For each N_0 we give the final size of the problem N_m (where in fact $m = 10$), the energy norm of the exact error corresponding to various approximations $\tilde{U}_m$ and the number of matvecs (5.1) over the 10 AFEM iterations.

We notice first that the difference between exact and inexact meshes is negligible while the error level is similar in all cases. We also see that the ideal stopping criterion DNR achieves roughly the same level of accuracy with a very low number of matvecs. The ideal bound GM0 appears to exhibit the same behaviour with a few more matvecs and consequently a more accurate approximation. The practical bounds generally require more matvecs for the same level of accuracy. The guaranteed upper bounds GM1 and GM2 are the most accurate (excepting the ER bounds) and also require the largest number of matvecs among the computable bounds. The adaptive bound GM1 is indeed an improvement over GM2 at essentially no cost. Among the estimators, GM3 and the anti-Gauss (AG) criterion provided competitive choices. The Hestenes-Stiefel estimator appears to loose accuracy which indicates that the delay needs to be increased with the mesh size.

The results of Table 5.1 are also displayed in Figure 5.2. The first plot shows the energy norm of



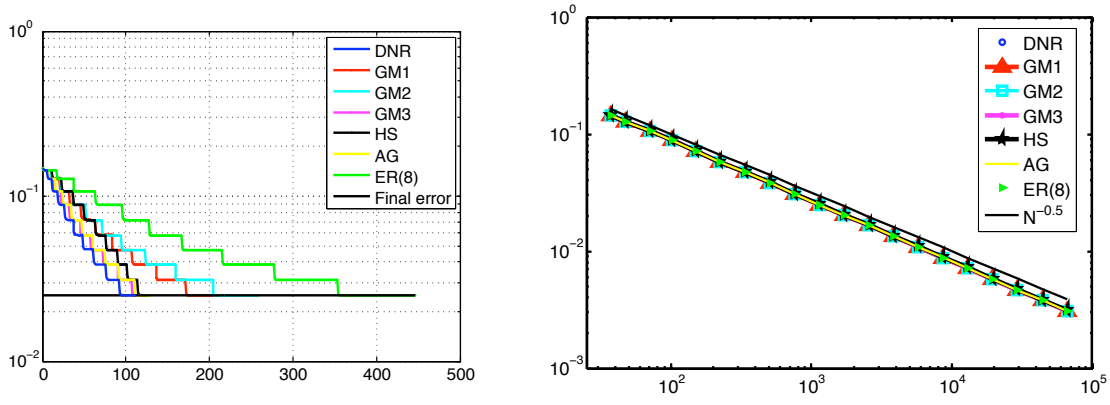
(a) Solution of test problem 1

(b) Adaptive meshes: $m = 4, 6, 8$; $N_0 = 83$.

FIG. 5.1. *Solution and adaptive meshes for test problem 1*

method	$N_0 = 28$			$N_0 = 83$			$N_0 = 262$		
	N_m	$\|u - \tilde{U}_m\|_a$	mv	N_m	$\|u - \tilde{U}_m\|_a$	mv	N_m	$\|u - \tilde{U}_m\|_a$	mv
exact	1,140	2.5028e–2	–	1,764	2.0155e–2	–	4,208	1.3033e–2	–
DNR	1,126	2.5086e–2	43	1,769	2.0176e–2	52	4,191	1.3058e–2	71
GM0	1,126	2.5107e–2	47	1,820	2.0033e–2	58	4,185	1.3053e–2	78
GM1	1,139	2.5035e–2	97	1,764	2.0155e–2	126	4,208	1.3033e–2	187
GM2	1,140	2.5028e–2	114	1,776	2.0124e–2	145	4,208	1.3033e–2	230
GM3	1,132	2.5052e–2	46	1,774	2.0138e–2	53	4,197	1.3057e–2	68
HS	1,136	2.5041e–2	33	1,782	2.0233e–2	37	4,194	1.3310e–2	31
AG	1,133	2.5031e–2	41	1,818	2.0032e–2	47	4,185	1.3061e–2	61
ER(6)	1,140	2.5028e–2	151	1,764	2.0155e–2	186	4,208	1.3033e–2	302
ER(8)	1,140	2.5028e–2	191	1,764	2.0155e–2	271	4,208	1.3033e–2	421
ER(10)	1,140	2.5028e–2	220	1,764	2.0155e–2	330	4,208	1.3033e–2	501

TABLE 5.1
Performance of stopping criteria: errors and matvecs (mv) for Test Problem 1 ($m = 10$).



(a) Errors $\|u - U_m^k\|_a$ for $m = 10$.

(b) Convergence rates for inexact AFEM for $m = 20$.

FIG. 5.2. *Convergence of inexact AFEM: CG with various stopping criteria.*

the error corresponding to CG iterates produced by the above stopping criteria throughout the adaptive process. We see that wasteful criteria such as ER exhibit long plateaux, spending a lot of iterations to achieve an insignificant reduction in the energy norm of the error. At the same time, in Figure 5.2(b) we see that the convergence rates of all the inexact algorithms remain virtually unchanged from the exact case, which for this test problem is $O(N_m^{-1/2})$.

The same problem was solved with 20 AFEM iterations, starting from the same three initial meshes. The results are shown in Table 5.2. The performance of the ideal estimator GM0 is not included. In this

	$N_0 = 28$			$N_0 = 83$			$N_0 = 262$		
method	N_m	$\|u - \tilde{U}_m\|_a$	mv	N_m	$\|u - \tilde{U}_m\|_a$	mv	N_m	$\|u - \tilde{U}_m\|_a$	mv
exact	66,115	3.0648e–3	–	101,292	2.4657e–3	–	235,574	1.6016e–3	–
DNR	66,004	3.0691e–3	170	101,290	2.4677e–3	200	235,087	1.6040e–3	285
GM1	66,102	3.0651e–3	358	101,285	2.4658e–3	423	235,519	1.6018e–3	509
GM2	66,115	3.0648e–3	882	101,444	2.4646e–3	1,097	235,574	1.6016e–3	1,665
GM3	65,994	3.0921e–3	99	101,438	2.4803e–3	145	235,275	1.6155e–3	193
HS	66,389	3.2464e–3	223	101,943	2.6552e–3	45	236,638	1.9113e–3	32
AG	66,048	3.0707e–3	166	102,023	2.4689e–3	84	235,198	1.6132e–3	90
ER(6)	66,115	3.0648e–3	962	101,292	2.4657e–3	1,168	235,574	1.6016e–3	1,702
ER(8)	66,115	3.0648e–3	1,346	101,292	2.4657e–3	1,643	235,574	1.6016e–3	2,457
ER(10)	66,115	3.0648e–3	1,677	101,292	2.4657e–3	2,004	235,574	1.6016e–3	3,025

TABLE 5.2
Performance of stopping criteria: errors and matvecs (mv) for Test Problem 1 ($m = 20$).

case also, the resulting meshes exhibit relatively small differences, with the level of error achieved similar to the exact case. There is one exception: estimator HS is very poor for large problems; this is due to the fact that this estimator is a lower bound which is known to be tight after a suitable delay and in our case, this delay is insufficient ($d = 5$). The guaranteed bounds GM1, GM2 provide again the most accurate approximations; they may appear expensive compared to the performance of the other estimators, but are certainly an improvement over the standard Euclidean criteria. In particular, the performance of GM1 is improving relative to GM2 and to the standard Euclidean criteria. Finally, the estimators GM3 and AG provide convenient alternatives which are cheap and, for many practical applications, sufficiently accurate.

5.1.2. Test Problem 2. We solved again problem (2.1) posed on $\Omega = (-1, 1) \times (-1, 1)$ and with the choice of diffusion coefficient

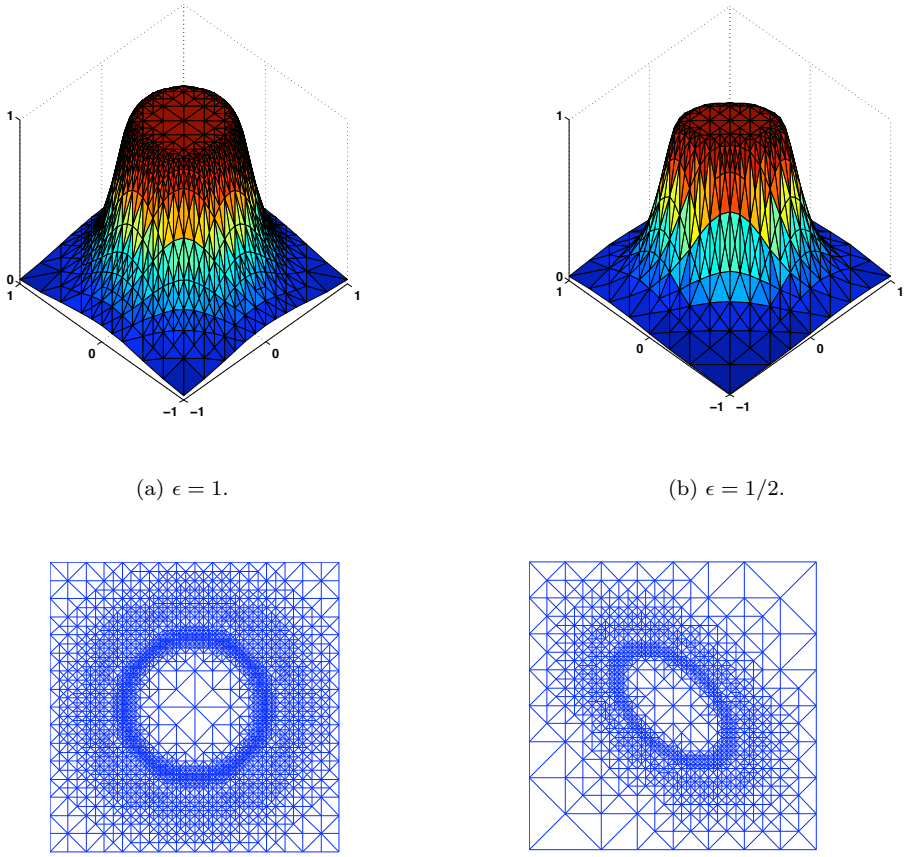
$$a = \frac{1}{\epsilon} \begin{pmatrix} 1 & \epsilon - 1 \\ \epsilon - 1 & 1 \end{pmatrix},$$

with $0 < \epsilon \leq 1$ and eigenvalues $\{1, 2/\epsilon - 1\}$ so that $a_{\min} = 1$. The right-hand side was chosen so that the exact solution is

$$u(x, y) = \tanh\left(\frac{0.1}{r^4 + 10^{-4}}\right), \quad r^2 = \frac{1}{\epsilon}\left(x^2 - 2(\epsilon - 1)xy + y^2\right).$$

The solution corresponding to $\epsilon = 1$ and $\epsilon = 1/2$ and the corresponding meshes are displayed in Figure 5.3.

The results are displayed in Table 5.3 for $\epsilon \in \{1, 1/2, 1/5\}$. We first note that (4.10) does not hold, as evident from the mesh plots in Figure 5.3. As a result, bound GM1 applies with lower bound (4.12) for the minimum eigenvalue of the system matrix. The choice of χ in (4.12) was of the order of $1/\min_{\kappa \in \mathcal{D}_m} \text{area}(\kappa)$; a bound for the smallest eigenvalue of the augmented matrix on the right in (4.12) was computed using a domain decomposition approach as described in [28]. The resulting performance is remarkably close to the ideal case DNR. The performance of GM2, the only other guaranteed bound, is an improvement over, but essentially of the order of the Euclidean criteria. The estimator HS with a delay $d = 5$ remains poor,



(a) $\epsilon = 1$.

(b) $\epsilon = 1/2$.

(c) $\epsilon = 1$.

(d) $\epsilon = 1/2$.

FIG. 5.3. *Solutions and adaptive meshes ($m = 10$) for test problem 2*

while the estimator AG performs poorly for $\epsilon = 1/5$. Estimator GM3 remains consistent, with the best performance overall.

5.2. Experiments in three dimensions.

5.2.1. Test Problem 3. For our final test, we solved problem (2.1) with $a = 1$ in $\Omega = (-1,\ 1)^3$ and the forcing function chosen so that the exact solution is $u = e^{-10r^2}$. We used the same Dörfler parameter $\theta = 0.75$ and started the adaptive algorithm from a range of initial regular meshes of tetrahedra and ran the procedure for $m = 10$ iterations. The refinement is concentrated near the origin, as illustrated in Figure 5.4, where the solution exhibits a sharp exponential decay.

The results are displayed in Table 5.4 and are similar to those obtained for the previous 2D experiment. As was the case for Test Problem 2, the criterion GM1 uses the bound (4.12) and is an improvement over GM2. Criterion AG turns out not to be defined for any level and this is indicated by an asterisk. The bound GM2 is cheaper than, but of the same order as, the ER criteria, while criteria GM3 and HS appear to yield consistently good quality approximations. In particular, criterion GM3 appears to yield marginally

	$\epsilon = 1$			$\epsilon = 1/2$			$\epsilon = 1/5$		
method	N_m	$\|u - \tilde{U}_m\|_a$	mv	N_m	$\|u - \tilde{U}_m\|_a$	mv	N_m	$\|u - \tilde{U}_m\|_a$	mv
exact	549,047	7.1750e–3	–	264,033	2.3346e–2	–	274,668	5.6532e–2	–
DNR	546,902	7.2040e–3	415	264,301	2.3416e–2	396	254,750	5.8860e–2	547
GM1	545,386	7.2009e–3	478	263,867	2.3359e–2	494	274,645	5.6488e–2	777
GM2	549,035	7.1750e–3	796	264,121	2.3343e–2	1,104	274,475	5.6535e–2	1,757
GM3	544,814	8.9772e–3	205	264,261	2.9244e–2	292	268,744	1.0138e–1	109
HS	539,606	1.4612e–2	11	264,814	1.1781e–1	13	275,467	3.3779e–1	15
AG	548,663	8.2482e–3	13	260,995	4.3329e–2	15	237,654	1.5098e+0	15
ER(6)	548,540	7.1782e–3	1,639	264,033	2.3346e–2	1,754	274,666	5.6532e–2	2,392
ER(8	549,047	7.1750e–3	2,075	264,033	2.3346e–2	2,356	274,668	5.6532e–2	3,095
ER(10)	549,047	7.1750e–3	3,003	264,033	2.3346e–2	3,007	274,668	5.6532e–2	4,592

TABLE 5.3
Performance of stopping criteria: errors and matvecs (mv) for Test Problem 2 ($m = 20, N_0 = 17$).

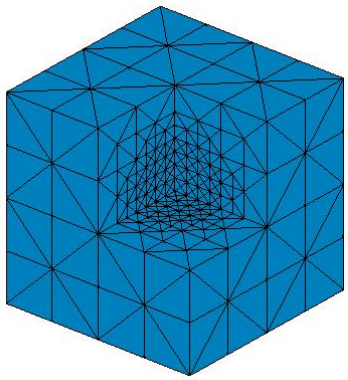


FIG. 5.4. *Adaptive meshing for test problem 3*

better solutions, though at a higher cost than criterion HS.

6. Concluding remarks. This work was motivated by the need to solve intractable large-scale problems arising in the context of adaptivity. For such problems, iterative methods are mandatory; however, classical stopping criteria, such as shareholding the Euclidean norm of the residual are known to be wasteful in practice. Our focus here has been to establish how accurate inexact linear solver approximations need to be in an adaptive context, while remaining provably convergent. In particular, we have showed that convergence holds provided the approximations replacing the exact solution satisfy a certain bound involving the energy norm of only algebraic errors on both current and previous level. This result yields guaranteed stopping criteria for iterative methods such as the conjugate gradient method using a number of guaranteed error bounds for the CG residual. In particular, Golub-Meurant type bounds are guaranteed to yield convergent and at times efficient results. Moreover, adaptive Poincaré inequalities, which allow

	$N_0 = 142$			$N_0 = 779$			$N_0 = 5,191$		
method	N_m	$\|u - \tilde{U}_m\|_a$	mv	N_m	$\|u - \tilde{U}_m\|_a$	mv	N_m	$\|u - \tilde{U}_m\|_a$	mv
exact	19,579	8.9670e-2	–	131,250	4.7497e-2	–	950,961	†	–
DNR	19,507	8.9617e-2	120	131,452	4.7744e-2	302	†	†	†
GM1	19,573	8.9665e-2	179	131,243	4.7498e-2	447	951,057	2.4695e-2	1,065
GM2	19,582	8.9672e-2	250	131,232	4.7497e-2	570	950,988	2.4696e-2	1,292
GM3	19,510	8.9682e-2	151	131,251	4.7484e-2	353	951,077	2.4695e-2	1,018
HS	19,648	9.0596e-2	120	131,606	4.9477e-2	291	958,982	2.6542e-2	676
AG	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$
ER(6)	19,587	8.9677e-2	238	131,246	4.7497e-2	465	951,239	2.4692e-2	958
ER(8	19,579	8.9670e-2	331	131,250	4.7497e-2	649	950,954	2.4697e-2	1,505
ER(10)	19,579	8.9670e-2	412	131,250	4.7497e-2	840	950,932	2.4697e-2	2,001

TABLE 5.4

Performance of stopping criteria: errors and matvecs (mv) for Test Problem 3 ($m = 10$) for various N_0. Legend: †: out of memory; −: does not apply; $*$: does not exist.

for better estimation of the energy norm of the algebraic errors, have employed resulting to good practical convergence results with competitive iteration count. Other estimates, such as the Hestenes and Stiefel criterion or Lanczos approximations have also been found to competitive in practice, without, of course, guaranteeing theoretical convergence of AFEM.

Finally, we note the forthcoming work [20] where the quasi-optimality of AFEM is considered for different stopping criterion. In the light of these developments, it would be interesting to investigate in the future whether quasi-optimality of AFEM can be also proven for the stopping criteria considered in this present work, taking into account the complexity considerations of the linear solvers, e.g., the use of optimal preconditioners for CG.

REFERENCES

[1] AINSWORTH, M., AND ODEN, J. T. *A posteriori error estimation in finite element analysis*. Pure and Applied Mathematics (New York). Wiley-Interscience [John Wiley & Sons], New York, 2000.

[2] ARIOLI, M. A stopping criterion for the conjugate gradient algorithms in a finite element method framework. *Numer. Math. 97*, 1 (2004), 1–24.

[3] ARIOLI, M., AND LOGHIN, D. Stopping criteria for mixed finite element problems. *Electron. Trans. Numer. Anal. 29* (2007/08), 178–192.

[4] ARIOLI, M., LOGHIN, D., AND WATHEN, A. J. Stopping criteria for iterations in finite element methods. *Numer. Math. 99*, 3 (2005), 381–410.

[5] ASHBY, S. F., HOLST, M. J., MANTEUFFEL, T. A., AND SAYLOR, P. E. The role of the inner product in stopping criteria for conjugate gradient iterations. *BIT 41*, 1 (2001), 26–52.

[6] AXELSSON, O., AND KAPORIN, I. Error norm estimation and stopping criteria in preconditioned conjugate gradient iterations. *Numer. Linear Algebra Appl. 8*, 4 (2001), 265–286.

[7] BARTELS, S., AND CARSTENSEN, C. A convergent adaptive finite element method for an optimal design problem. *Numer. Math. 108*, 3 (2008), 359–385.

[8] BINEV, P., DAHMEN, W., AND DEVORE, R. Adaptive finite element methods with convergence rates. *Numer. Math. 97*, 2 (2004), 219–268.

[9] BONITO, A., AND NOCHETTO, R. Quasi-optimal convergence rate of an interior penalty adaptive discontinuous Galerkin method. *Submitted for publication*.

[10] BRAESS, D., CARSTENSEN, C., AND HOPPE, R. H. W. Convergence analysis of a conforming adaptive finite element

method for an obstacle problem. *Numer. Math. 107*, 3 (2007), 455–471.

[11] Calvetti, D., Morigi, S., Reichel, L., and Sgallari, F. Computable error bounds and estimates for the conjugate gradient method. *Numer. Algorithms 25*, 1-4 (2000), 75–88. Mathematical journey through analysis, matrix theory and scientific computation (Kent, OH, 1999).

[12] Calvetti, D., Morigi, S., Reichel, L., and Sgallari, F. An iterative method with error estimators. *J. Comput. Appl. Math. 127*, 1-2 (2001), 93–119. Numerical analysis 2000, Vol. V, Quadrature and orthogonal polynomials.

[13] Carstensen, C., and Hoppe, R. H. W. Convergence analysis of an adaptive nonconforming finite element method. *Numer. Math. 103*, 2 (2006), 251–266.

[14] Cascon, J. M., Kreuzer, C., Nochetto, R. H., and Siebert, K. G. Quasi-optimal convergence rate for an adaptive finite element method. *SIAM J. Numer. Anal. 46*, 5 (2008), 2524–2550 (electronic).

[15] Ciarlet, P. G. *The finite element method for elliptic problems*, vol. 40 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. Reprint of the 1978 original.

[16] Deuflhard, P. Cascadic conjugate gradient methods for elliptic partial differential equations: algorithm and numerical results. In *Domain decomposition methods in scientific and engineering computing (University Park, PA, 1993)*, vol. 180 of *Contemp. Math.* Amer. Math. Soc., Providence, RI, 1994, pp. 29–42.

[17] Dörfler, W. A convergent adaptive algorithm for Poisson's equation. *SIAM J. Numer. Anal. 33*, 3 (1996), 1106–1124 (electronic).

[18] Fried, I. Bounds on the extremal eigenvalues of the finite element stiffness and mass matrices and their spectral condition number. *Journal of Sound and Vibration 22*, 4 (1972), 407 – 418.

[19] Fried, I. Bounds on the spectral and maximum norms of the finite element stiffness, flexibility and mass matrices. *Internat. J. Solids and Structures 9* (1973), 1013 – 1034.

[20] Garau, E., Morin, and Zuppa, C. Quasi-optimal convergence of an inexact adaptive finite element method. *in preparation*.

[21] Golub, G. H., and Meurant, G. Matrices, moments and quadrature. II. How to compute the norm of the error in iterative methods. *BIT 37*, 3 (1997), 687–705. Direct methods, linear algebra in optimization, iterative methods (Toulouse, 1995/1996).

[22] Golub, G. H., and Meurant, G. *Matrices, moments and quadrature with applications*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2010.

[23] Golub, G. H., and Strakoš, Z. Estimates in quadratic formulas. *Numer. Algorithms 8*, 2-4 (1994), 241–268.

[24] Hestenes, M. R., and Stiefel, E. Methods of conjugate gradients for solving linear systems. *J. Research Nat. Bur. Standards 49* (1952), 409–436 (1953).

[25] Hoppe, R. H. W., Kanschat, G., and Warburton, T. Convergence analysis of an adaptive interior penalty discontinuous Galerkin method. *SIAM J. Numer. Anal. 47*, 1 (2008/09), 534–550.

[26] Jiranek, P., Strakoš, Z., and Vohralik, M. A posteriori error estimates including algebraic error: Computable upper bounds and stopping criteria for iterative solvers. *Submitted for publication*.

[27] Karakashian, O. A., and Pascal, F. Convergence of adaptive discontinuous Galerkin approximations of second-order elliptic problems. *SIAM J. Numer. Anal. 45*, 2 (2007), 641–665 (electronic).

[28] Loghin, D. Realistic eigenvalue bounds for stiffness matrices. Tech. Rep. 9/2009, University of Birmingham, 2009.

[29] Meurant, G. The computation of bounds for the norm of the error in the conjugate gradient algorithm. *Numer. Algorithms 16*, 1 (1997), 77–87 (1998). Sparse matrices in industry (Lille, 1997).

[30] Meurant, G. *Computer solution of large linear systems*, vol. 28 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1999.

[31] Meurant, G. Numerical experiments in computing bounds for the norm of the error in the preconditioned conjugate gradient algorithm. *Numer. Algorithms 22*, 3-4 (1999), 353–365 (2000).

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

[33] Morin, P., Siebert, K. G., and Veeser, A. A basic convergence result for conforming adaptive finite elements. *Math. Models Methods Appl. Sci. 18*, 5 (2008), 707–737.

[34] Picasso, M. A stopping criterion for the conjugate gradient algorithm in the framework of anisotropic adaptive finite elements. *Comm. Numer. Methods Engrg. 25*, 4 (2009), 339 – 355.

[35] Stevenson, R. Optimality of a standard adaptive finite element method. *Found. Comput. Math. 7*, 2 (2007), 245–269.

[36] Strakoš, Z., and Tichý, P. On error estimation in the conjugate gradient method and why it works in finite precision computations. *Electron. Trans. Numer. Anal. 13* (2002), 56–80 (electronic).

[37] Veeser, A. Convergent adaptive finite elements for the nonlinear Laplacian. *Numer. Math. 92*, 4 (2002), 743–770.

[38] Verfürth, R. *A Review of* A Posteriori *Error Estimation and Adaptive Mesh-Refinement Techniques*. Wiley-Teubner, Chichester-Stuttgart, 1996.

[39] Wathen, A. J. Realistic eigenvalue bounds for the Galerkin mass matrix. *IMA J. Numer. Anal. 7*, 4 (1987), 449–457.