

## Stopping criteria for iterations in finite element methods

M. Arioli<sup>1</sup>, D. Loghin<sup>2</sup>, A. J. Wathen<sup>3</sup>

<sup>1</sup> Atlas Centre, Rutherford Appleton Laboratory, Oxon OX11 0QX, UK;  
e-mail: m.arioli@rl.ac.uk

<sup>2</sup> CERFACS, 42 ave G. Coriolis, Toulouse, 31057, France;  
e-mail: loghin@cerfas.fr

<sup>3</sup> Oxford University Computing Laboratory, Parks Road, Oxford, OX1 3QD, UK;  
e-mail: wathen@comlab.ox.ac.uk

Received March 21, 2003 / Revised version received February 25, 2004  
Published online November 26, 2004 – © Springer-Verlag 2004

**Summary.** This work extends the results of Arioli [1], [2] on stopping criteria for iterative solution methods for linear finite element problems to the case of nonsymmetric positive-definite problems. We show that the residual measured in the norm induced by the symmetric part of the inverse of the system matrix is relevant to convergence in a finite element context. We then use Krylov solvers to provide alternative ways of calculating or estimating this quantity and present numerical experiments which validate our criteria.

*Mathematics Subject Classification (2000):* 65N30, 65F10, 65F35

### 1 Introduction

Iterative methods of Krylov subspace type form a well-known and well-researched area in the context of solution methods for large sparse linear systems. In some cases, convergence can be described, in others not. Invariably however, the theoretical and practical convergence criterion is chosen to be the Euclidean norm of the residual, with the ubiquitous exception of the Conjugate Gradient method, where the ‘energy norm’ lends itself quite naturally to analysis. On the other hand, finite element methods which are an important source of large, sparse linear systems provide a natural norm for convergence. While this fact is well-known and has been noted particularly in the case of symmetric positive-definite problems [10], [11], [16], [19] only

---

*Correspondence to:* M. Arioli

recently have there been attempts to relate convergence in the ‘energy’ norm to the finite element context [7], [8], [9], [18], [1], [2].

In this work we consider the choice of stopping criteria for nonsymmetric positive-definite problems. The immediate difficulty encountered is that of defining a suitable norm in which to measure convergence. In the case of symmetric positive-definite problems, the energy or  $A$ -norm of the error is equal to the dual norm or  $A^{-1}$ -norm of the residual, which is the quantity that is estimated. In the nonsymmetric case, we show that a useful definition of dual norm is the norm induced by the symmetric part of  $A^{-1}$ . We show that one can also work with the norm induced by the inverse of the symmetric part of  $A$  for problems which are not too non-normal.

The paper is structured as follows. In Section 2 we describe the problem setting. In Section 3 we derive general stopping criteria while in Sections 4 and 5 we present ways of approximating the criteria introduced in the case of GMRES; we also consider the effect of preconditioning and derive the corresponding modified bounds. Finally, in Section 6 we investigate the stopping criteria by performing experiments on various discretizations of convection-diffusion problems.

## 2 Problem description

### 2.1 Abstract formulation

Consider the weak formulation

Find  $u \in \mathcal{H}$  such that for all  $v \in \mathcal{H}$

$$(1) \quad a(u, v) = f(v),$$

where  $\mathcal{H}$  is a Hilbert space of functions  $u$  defined on a closed subset  $\Omega$  of  $\mathbb{R}^d$ , with dual  $\mathcal{H}'$  and inner-product norm  $\|\cdot\|_{\mathcal{H}}$ , while  $a(\cdot, \cdot)$  is a nonsymmetric, positive-definite bilinear form on  $\mathcal{H} \times \mathcal{H}$  and  $f(\cdot) \in \mathcal{H}'$  is a continuous linear form on  $\mathcal{H}$ . Existence and uniqueness of solutions to problems of type (1) is guaranteed provided the following conditions hold for all  $u, v \in \mathcal{H}$

$$(2a) \quad a(w, v) \leq C_1 \|w\|_{\mathcal{H}} \|v\|_{\mathcal{H}}$$

$$(2b) \quad a(v, v) \geq C_2 \|v\|_{\mathcal{H}}^2,$$

with constants  $C_1, C_2$  independent of discretization. In the following, we take  $C_1, C_2$  to be the smallest and, respectively largest, such constants.

Condition (2b) is often used to replace – and implies – the weaker and sufficient conditions of Babuška ([3])

$$(3a) \quad \sup_{v \in \mathcal{H} \setminus \{0\}} \frac{a(w, v)}{\|v\|_{\mathcal{H}}} \geq C_2 \|w\|_{\mathcal{H}},$$

$$(3b) \quad \sup_{w \in \mathcal{H} \setminus \{0\}} \frac{a(w, v)}{\|w\|_{\mathcal{H}}} \geq C_2 \|v\|_{\mathcal{H}};$$

this is due to the fact that the weak formulation (1) with  $a(\cdot, \cdot)$  replaced by its symmetric part is often stable in the sense of Babuška (i.e., satisfies (3)), leading to (2b).

Finally, we note that the continuity condition (2a) implies that the bilinear form defines a continuous operator  $\mathcal{A} : \mathcal{H} \rightarrow \mathcal{H}'$  with norm  $\|\mathcal{A}\|_{\mathcal{H} \rightarrow \mathcal{H}'} \leq C_1$ .

## 2.2 Finite element approximation

An approximation to problem (1) is sought through projection onto a finite-dimensional space  $\mathcal{H}_h \subset \mathcal{H}$ ; the resulting formulation reads

Find  $u_h \in \mathcal{H}_h$  such that for all  $v_h \in \mathcal{H}_h$

$$(4) \quad a(u_h, v_h) = f(v_h).$$

Finite element methods choose  $\mathcal{H}_h$  to be a space of functions  $v_h$  defined on a subdivision  $\Omega_h$  of  $\Omega$  into simplices  $T$  of diameter  $h_T$ ;  $h$  denotes a piecewise constant function defined on  $\Omega_h$  via  $h|_T = h_T$ .

Since  $\mathcal{H}_h \subset \mathcal{H}$ , (2) are satisfied with constants independent of  $h$ ; thus, there exists a unique finite element approximation  $u_h$ . Moreover subtracting (4) from (1) yields the standard orthogonality condition for all  $v_h \in \mathcal{H}_h$

$$(5) \quad a(u - u_h, v_h) = 0,$$

which can be used (together with conditions (2)) to derive standard error estimates of the form

$$(6) \quad \|u - u_h\|_{\mathcal{H}} \leq \frac{C_1}{C_2} \inf_{v_h \in \mathcal{H}_h} \|u - v_h\|_{\mathcal{H}}.$$

*Remark 1* Replacing  $v_h$  with the interpolant of  $u$  and using interpolation error estimates leads to a priori bounds of the form

$$\|u - u_h\|_{\mathcal{H}} \leq C(h)C(u)$$

where  $C(u)$  is typically a constant depending only on  $u$  and its derivatives.

Conditions (2) can also be used in determining *a posteriori* error bounds. In particular, if we define the functional residual as a linear functional via

$$\langle \mathcal{R}(u_h), v \rangle := f(v) - a(u_h, v) = a(u - u_h, v) \quad \forall v \in \mathcal{H}$$

then dividing by  $\|v\|_{\mathcal{H}}$  and using (3a) and (2a) leads, respectively, to the following upper and lower bounds on the error

$$(7) \quad \frac{1}{C_1} \|\mathcal{R}(u_h)\|_{\mathcal{H}'} \leq \|u - u_h\|_{\mathcal{H}} \leq \frac{1}{C_2} \|\mathcal{R}(u_h)\|_{\mathcal{H}'}$$

where

$$\|\mathcal{R}(u_h)\|_{\mathcal{H}'} := \sup_{v \in \mathcal{H} \setminus \{0\}} \frac{|\langle \mathcal{R}(u_h), v \rangle|}{\|v\|_{\mathcal{H}}} = \sup_{v \in \mathcal{H} \setminus \{0\}} \frac{|a(u - u_h, v)|}{\|v\|_{\mathcal{H}}}$$

Alternatively, noting that  $\|\mathcal{A}\|_{\mathcal{H} \rightarrow \mathcal{H}'} \leq C_1$  (cf. (2a)) we can rewrite (7) as

$$(8) \quad \mathcal{B}\mathcal{E} \leq \frac{\|u - u_h\|_{\mathcal{H}}}{\|u_h\|_{\mathcal{H}}} =: \mathcal{F}\mathcal{E} \leq \frac{C_1}{C_2} \mathcal{B}\mathcal{E}$$

where

$$(9) \quad \mathcal{B}\mathcal{E} := \frac{\|\mathcal{R}(u_h)\|_{\mathcal{H}'}}{\|u_h\|_{\mathcal{H}} \|\mathcal{A}\|_{\mathcal{H} \rightarrow \mathcal{H}'}}$$

**Definition 1** *The quantities  $\mathcal{F}\mathcal{E}$ ,  $\mathcal{B}\mathcal{E}$  in (8),(9) are the functional forward and backward error respectively [2].*

*Remark 2* The dual norm of the functional residual,  $\|\mathcal{R}(u_h)\|_{\mathcal{H}'}$ , is not easy to compute and most *a posteriori* error bounds are derived as approximations of this quantity. However, in general it is known that  $\|\mathcal{R}(u_h)\|_{\mathcal{H}'}$  and thus  $\mathcal{B}\mathcal{E}$  are (polynomial) functions of the discretization parameter  $h$  and thus far from being close to machine precision. This is our main motivation for seeking new, improved stopping criteria. However, we will not be concerned here with the derivation of any error bounds but we will assume the following generic bound on the relative error

$$(10) \quad \frac{\|u - u_h\|_{\mathcal{H}}}{\|u_h\|_{\mathcal{H}}} \leq C(h)$$

where  $C(h)$  is available via an *a priori* or *a posteriori* error analysis.

It is evident from the above description that approximate solutions and errors on one hand and residuals and the right hand side data on the other belong to spaces with different topologies: the trial space  $\mathcal{H}$  and its dual. Moreover, the operator  $\mathcal{A}$  has a domain different from its range. We choose to preserve these essential features in our discrete formulation of the problem. To this aim, we introduce the following notation and results.

We first define a matrix norm  $\|\cdot\|_{H_1, H_2} : \mathbb{R}^{n \times n}$  via

$$\|M\|_{H_1, H_2} := \max_{\mathbf{x} \in \mathbb{R}^n \setminus \{0\}} \frac{\|M\mathbf{x}\|_{H_2}}{\|\mathbf{x}\|_{H_1}}$$

where  $M \in \mathbb{R}^{n \times n}$  and  $H_i \in \mathbb{R}^{n \times n}$ ,  $i = 1, 2, 3$  are symmetric and positive-definite matrices. This choice of norm will allow us to consider matrices as operators for which domain and range are equipped with different norms. We also note here that

$$(11) \quad \|M\|_{H_1, H_2^{-1}} = \|H_2^{-1/2} M H_1^{-1/2}\|,$$

where  $\|\cdot\|$  denotes the standard Euclidean norm.

We now state a result which can be found in [4].

**Theorem 1** *Let  $M \in \mathbb{R}^{n \times n}$  be nonsingular and let  $H \in \mathbb{R}^{n \times n}$  be a symmetric and positive-definite matrix. Then*

$$\begin{aligned} \|M\|_{H, H^{-1}} &= \max_{\mathbf{w} \in \mathbb{R}^n \setminus \{0\}} \max_{\mathbf{v} \in \mathbb{R}^n \setminus \{0\}} \frac{\mathbf{w}^T M \mathbf{v}}{\|\mathbf{w}\|_H \|\mathbf{v}\|_H}, \\ \|M^{-1}\|_{H^{-1}, H}^{-1} &= \min_{\mathbf{w} \in \mathbb{R}^n \setminus \{0\}} \max_{\mathbf{v} \in \mathbb{R}^n \setminus \{0\}} \frac{\mathbf{w}^T M \mathbf{v}}{\|\mathbf{w}\|_H \|\mathbf{v}\|_H}. \end{aligned}$$

The above result justifies the following definition.

**Definition 2** *The  $H$ -condition number of a matrix  $M$  is*

$$\kappa_H(M) := \|M\|_{H, H^{-1}} \|M^{-1}\|_{H^{-1}, H}.$$

We now turn to the discrete setting for the framework described above. Expanding  $u_h$  in a basis of  $\mathcal{H}_h$ , we can derive a linear system of equations involving the coefficients  $(\mathbf{u})_i$ ,  $i = 1, \dots, n$  of  $u_h$  in our choice of basis of  $\mathcal{H}_h$

$$(12) \quad A\mathbf{u} = \mathbf{f}$$

where  $n = \dim \mathcal{H}_h$  and  $A \in \mathbb{R}^{n \times n}$  is a non-singular, generally nonsymmetric, matrix. In fact there is an isomorphism  $\Pi_h$  between  $\mathbb{R}^n$  and  $\mathcal{H}_h$  which associates to every vector  $\mathbf{v} \in \mathbb{R}^n$  a function  $v_h \in \mathcal{H}_h$  via

$$\Pi_h \mathbf{v} = \sum_{i=1}^n v_i \phi_i,$$

where  $\{\phi_i, i = 1, \dots, n\}$  form a basis for  $\mathcal{H}_h$ . Henceforth, given a vector  $\mathbf{v} \in \mathbb{R}^n$  we will denote its functional counterpart  $\Pi_h \mathbf{v}$  by  $v_h$ . Note also that the above choice of basis defines a norm-matrix  $H$  via

$$H_{ij} = ((\phi_i, \phi_j))$$

where  $((\cdot, \cdot))$  is the  $\mathcal{H}$ -inner product. Hence

$$\|v_h\|_{\mathcal{H}} = \|\mathbf{v}\|_H = (\mathbf{v}^T H \mathbf{v})^{1/2}.$$

With this notation, the stability conditions (2) become

$$(13a) \quad \max_{\mathbf{w} \in \mathbb{R}^n \setminus \{0\}} \max_{\mathbf{v} \in \mathbb{R}^n \setminus \{0\}} \frac{\mathbf{w}^T A \mathbf{v}}{\|\mathbf{w}\|_H \|\mathbf{v}\|_H} \leq C_1$$

$$(13b) \quad \min_{\mathbf{v} \in \mathbb{R}^n \setminus \{0\}} \frac{\mathbf{v}^T A \mathbf{v}}{\|\mathbf{v}\|_H^2} \geq C_2$$

It is easy to see that there also exists a constant  $C_3 \leq C_1$  such that

$$(13c) \quad \max_{\mathbf{v} \in \mathbb{R}^n \setminus \{0\}} \frac{\mathbf{v}^T A \mathbf{v}}{\|\mathbf{v}\|_H^2} \leq C_3.$$

*Remark 3* In many situations of interest one can have  $C_3 = C_2$ . Moreover, if the symmetric part of  $A$  is  $H$ , then  $C_2 = C_3 = 1$ .

Finally, we derive the discrete versions of (8), (9) for the case where  $\mathcal{H}$  and its dual are replaced by  $\mathcal{H}_h$  and its dual; moreover, we assume that we seek an approximation  $\tilde{\mathbf{u}} \in \mathcal{H}_h$  to the solution  $\mathbf{u}$  of the linear system (12). Given the basis  $\{\phi_i, i = 1, \dots, n\}$  for  $\mathcal{H}_h$ , the discrete dual,  $\mathcal{H}'_h$ , is spanned by a dual basis  $\{\phi'_i, i = 1, \dots, n\}$ , defined via  $\langle \phi_i, \phi'_j \rangle = \delta_{ij}$ . As before, there exists an isomorphism  $\Pi'_h$  between  $\mathbb{R}^n$  and  $\mathcal{H}'_h$  defined similarly via  $\Pi'_h \mathbf{v}' = \sum_{i=1}^n v'_i \phi'_i$ ; moreover,  $\langle v, v' \rangle = \mathbf{v}^T \mathbf{v}'$ . Thus, one can define the residual  $R(\tilde{\mathbf{u}}) = \Pi'_h(\mathbf{b} - A\tilde{\mathbf{u}})$ , as a linear functional from  $\mathcal{H}'_h$  into  $\mathcal{H}_h$  with norm

$$\|R(\tilde{\mathbf{u}})\|_{\mathcal{H}'_h} := \sup_{v \in \mathcal{H}_h \setminus \{0\}} \frac{|\langle R(\tilde{\mathbf{u}}), v \rangle|}{\|v\|_{\mathcal{H}_h}} = \max_{\mathbf{v} \in \mathbb{R}^n \setminus \{0\}} \frac{|(\mathbf{b} - A\tilde{\mathbf{u}})^T \mathbf{v}|}{\|\mathbf{v}\|_H} = \|\mathbf{b} - A\tilde{\mathbf{u}}\|_{H^{-1}}.$$

The above formalism motivates the following definitions: given an approximation  $\tilde{\mathbf{u}}$  to the solution  $\mathbf{u}$  of the linear system  $A\mathbf{u} = \mathbf{f}$  we define discrete forward and backward errors via (cf. (8), (9))

$$(14) \quad FE := \frac{\|\mathbf{u} - \tilde{\mathbf{u}}\|_H}{\|\tilde{\mathbf{u}}\|_H}, \quad BE := \frac{\|\mathbf{f} - A\tilde{\mathbf{u}}\|_{H^{-1}}}{\|\tilde{\mathbf{u}}\|_H \|A\|_{H, H^{-1}}}.$$

In the following section (cf. Thm 2), we show that, as in the continuous case, a similar upper bound holds on the forward error  $FE \leq \frac{C_1}{C_2} BE$ . This is an expected result since (cf. (13), Thm 1)

$$(15) \quad \|A\|_{H, H^{-1}} \leq C_1, \quad \|A^{-1}\|_{H^{-1}, H} \geq C_2,$$

and hence for all  $n$

$$(16) \quad \kappa_H(A) \leq \frac{C_1}{C_2}.$$

Stopping criteria for iterations in finite element methods

In other words, we recover the well-known result of linear algebra [12]

$$\text{forward error} \leq \text{condition number} \times \text{backward error}$$

but with respect to the norms inherited from the problem formulation.

### 3 Stopping criteria

In many large-scale computations the exact solution  $\mathbf{u}$  of the linear system (12) is out of reach and an iterate  $\tilde{\mathbf{u}}$  is used to approximate the solution. Since we identify  $\tilde{\mathbf{u}}$  with a function  $\tilde{u}_h \in \mathcal{H}_h$ , we naturally expect a useful iterate  $\tilde{\mathbf{u}}$  to satisfy an error estimate similar to (10)

$$\frac{\|u - \tilde{u}_h\|_{\mathcal{H}}}{\|\tilde{u}_h\|_{\mathcal{H}}} \leq \tilde{C}(h),$$

where  $\tilde{C}(h)$  is of the same order as  $C(h)$  in (10). Our aim is to derive a sufficient and computable criterion for the above error bound to hold. First, we introduce some notation and useful results. Let  $M \in \mathbb{R}^{n \times n}$ . We denote by  $H_M = (M + M^T)/2$ ,  $S_M = (M - M^T)/2$  the symmetric and skew-symmetric parts of  $M$ , respectively. Moreover, if  $H_M$  is positive-definite, it induces a norm which we denote by

$$\|\cdot\|_M := \|\cdot\|_{H_M}.$$

We first prove the following results.

**Lemma 1** *Let conditions (13) hold. Then*

$$\frac{1}{\sqrt{C_3}} \|\mathbf{r}\|_A \leq \|\mathbf{r}\|_H \leq \frac{1}{\sqrt{C_2}} \|\mathbf{r}\|_A$$

and

$$\frac{\sqrt{C_2}}{C_1 C_3} \|\mathbf{r}\|_{H^{-1}} \leq \|\mathbf{r}\|_{A^{-1}} \leq \frac{1}{\sqrt{C_2}} \|\mathbf{r}\|_{H^{-1}}.$$

*Proof* See Appendix. □

**Theorem 2** *Let  $u$  be the solution of the weak formulation (1) and let  $\mathbf{u}$ ,  $u_h = \Pi_h \mathbf{u}$  satisfy*

$$\mathbf{A}\mathbf{u} = \mathbf{f}; \quad \frac{\|u - u_h\|_{\mathcal{H}}}{\|u_h\|_{\mathcal{H}}} \leq C(h).$$

*Then  $\tilde{u}_h = \Pi_h \tilde{\mathbf{u}}$  satisfies*

$$\frac{\|u - \tilde{u}_h\|_{\mathcal{H}}}{\|\tilde{u}_h\|_{\mathcal{H}}} \leq \tilde{C}(h) = O(C(h))$$

if

$$(17) \quad \frac{\|\mathbf{f} - A\tilde{\mathbf{u}}\|_{H^{-1}}}{\|\tilde{\mathbf{u}}\|_H} \leq \eta C(h) C_2,$$

for some  $\eta \in (0, 1)$ .

*Proof* Let  $\mathbf{r} = \mathbf{f} - A\tilde{\mathbf{u}}$ . We have

$$\begin{aligned} \frac{\|u - \tilde{u}_h\|_{\mathcal{H}}}{\|\tilde{u}_h\|_{\mathcal{H}}} &\leq \frac{\|u - u_h\|_{\mathcal{H}}}{\|u_h\|_{\mathcal{H}}} \frac{\|u_h\|_{\mathcal{H}}}{\|\tilde{u}_h\|_{\mathcal{H}}} + \frac{\|u_h - \tilde{u}_h\|_{\mathcal{H}}}{\|\tilde{u}_h\|_{\mathcal{H}}} \\ &\leq C(h) \left( 1 + \frac{\|u_h - \tilde{u}_h\|_{\mathcal{H}}}{\|\tilde{u}_h\|_{\mathcal{H}}} \right) + \frac{\|u_h - \tilde{u}_h\|_{\mathcal{H}}}{\|\tilde{u}_h\|_{\mathcal{H}}} \end{aligned}$$

and since

$$\begin{aligned} \frac{\|u_h - \tilde{u}_h\|_{\mathcal{H}}}{\|\tilde{u}_h\|_{\mathcal{H}}} &= \frac{\|A^{-1}\mathbf{r}\|_H}{\|\tilde{\mathbf{u}}\|_H} \\ &\leq \frac{\|A^{-1}\|_{H^{-1}, H} \|\mathbf{r}\|_{H^{-1}}}{\|\tilde{\mathbf{u}}\|_H} \\ &\leq \frac{1}{C_2} \frac{\|\mathbf{r}\|_{H^{-1}}}{\|\tilde{\mathbf{u}}\|_H} \quad (\text{using (15)}) \end{aligned}$$

we get

$$\frac{\|u - \tilde{u}_h\|_{\mathcal{H}}}{\|\tilde{u}_h\|_{\mathcal{H}}} \leq C(h)(1 + \eta C(h)) + \eta C(h) =: \tilde{C}(h).$$

□

*Remark 4* The stopping criterion (17) is equivalent to requiring the discrete backward error  $BE$  defined in (14) to be of the same order as the functional backward error  $BE = O(\mathcal{BE})$ . This is also a sufficient condition for the discrete forward error  $FE$  corresponding to our iterative solution to have the same order as the functional forward error  $\mathcal{FE}$ .

In fact, criterion (17) can be replaced with a tighter bound. By Lemma 1,

$$\|A^{-1}\mathbf{r}\|_H \leq \frac{1}{\sqrt{C_2}} \|A^{-1}\mathbf{r}\|_A \leq \frac{1}{\sqrt{C_2}} \|A^{-1}\|_{A^{-1}, A} \|\mathbf{r}\|_{A^{-1}} = \frac{1}{\sqrt{C_2}} \|\mathbf{r}\|_{A^{-1}}$$

and thus, we can replace the bound (17) with

$$(18) \quad \frac{\|\mathbf{f} - A\tilde{\mathbf{u}}\|_{A^{-1}}}{\|\tilde{\mathbf{u}}\|_H} \leq \eta C(h) \sqrt{C_2}.$$

The difference between the stopping criteria (17), (18) is not significant if the  $H$ -condition number (16) is not too large. This can be seen from the equivalence between  $\|\cdot\|_{H^{-1}}$  and  $\|\cdot\|_{A^{-1}}$  provided by Lemma 1. In particular, if

the symmetric part of  $A$  is  $H$ , then  $C_2 = C_3 = 1$  and the effective condition number is  $C_1$ . A large value of  $C_1$  corresponds to a ‘highly nonsymmetric’ problem for which the use of criterion (18) rather than (17) may be preferable. We return to this issue in the numerics section.

### 3.1 One more crime

In practice, the discretization of the weak formulation (1) is generally done in an approximate fashion, very often due to the computational costs involved. This approximation has been qualified as a *variational crime* [20], as it leads to a perturbed system

$$(A + \Delta A)\tilde{\mathbf{u}} = \mathbf{f}.$$

However, it is known that if the perturbation  $\Delta A$  is suitably small (usually within the finite element error), then the approximate solution  $\tilde{\mathbf{u}}$  satisfies the same error estimates as the exact solution  $\mathbf{u}$  [20]. In this context, the proposed stopping criteria represent but another variational crime as the following standard result shows (see also [2] or [17] for the case when  $l_p$  norms are employed.)

**Theorem 3** *Let  $\tilde{\mathbf{u}}$  satisfy*

$$\frac{\|\mathbf{f} - A\tilde{\mathbf{u}}\|_{H^{-1}}}{\|\tilde{\mathbf{u}}\|_H} \leq \eta C(h)C_2.$$

*Then there exists  $\Delta A$  such that*

$$(A + \Delta A)\tilde{\mathbf{u}} = \mathbf{f}$$

*and*

$$\|\Delta A\|_{H, H^{-1}} \leq \eta C(h)C_2$$

*Proof* See [2, Thm 1]. □

*Remark 5* The more general case where the right-hand side  $\mathbf{f}$  is perturbed is treated in [2]. We do not include the results here since in most engineering applications bounds of type (10) are preferred.

The stopping criteria derived above pose the problem of estimating the residual in the  $H^{-1}$ - or  $A^{-1}$ -norms. While this was possible for the symmetric and positive-definite case in a natural way (see [10], [11]), the use of a non-symmetric iterative method does not allow for the same methodology to be applied.

In the remainder of the paper we show how this norm can be estimated using the information contained in the Krylov space  $\mathcal{K}_k$ . For simplicity, we

will consider only the case  $H = (A + A^T)/2$  (and thus  $C_2 = 1$ ), i.e., the case when  $H$  defines the so-called ‘energy norm’ for the problem. In the next section, we show how the norm estimation is achieved for GMRES and FOM. Finally, in the case of central preconditioning with  $H$ , these two algorithms reduce to a three-term recurrence which computes directly  $\|\mathbf{r}^k\|_{H^{-1}}$ . This will be the subject of Section 5.

#### 4 Stopping criteria for GMRES and FOM

We recall here some of the basic facts and standard notation for the GMRES and FOM algorithms. The methods compute an orthonormal basis of  $\mathcal{K}_k$ ; the basis elements are the columns of  $V_k \in \mathbb{R}^{n \times k}$ . This orthonormalization is achieved via an Arnoldi process which yields the factorizations

$$V_k^T A V_k = H_k, \quad A V_k = V_{k+1} \tilde{H}_k$$

where  $H_k \in \mathbb{R}^{k \times k}$ ,  $\tilde{H}_k \in \mathbb{R}^{(k+1) \times k}$  are upper Hessenberg matrices, with  $H_k$  being obtained from  $\tilde{H}_k$  by deleting its last row. In the case of GMRES, a  $QR$ -factorization of  $\tilde{H}_k$  is computed (updated at each step)

$$\tilde{H}_k = Q_k R_k.$$

##### 4.1 Estimation of $\|\mathbf{r}^k\|_{H^{-1}}$

This can be done simply via

$$\|\mathbf{r}^k\|_{H^{-1}} \leq \lambda_{\min}^{-1/2}(H) \|\mathbf{r}^k\|.$$

Depending on the application, the smallest eigenvalue of  $H$  may or may not be estimated with sufficient accuracy. If we do not have such an estimate, we must content ourselves with estimates provided by the iterative process. In the case of GMRES and FOM this can be achieved as follows. Assuming no early termination, the method computes the following factorization of  $A$  involving an orthonormal matrix  $V_n$

$$V_n^T A V_n = H_n.$$

Thus,  $V_n^T H V_n = (H_n + H_n^T)/2 =: H_n^s$  and therefore

$$\lambda_{\min}(H) = \lambda_{\min}(H_n^s).$$

Since in practice we wish to use the algorithm only for a small number of steps  $k$ , an estimate of  $\lambda_{\min}(H)$  can be taken to be  $\lambda_{\min}(H_k^s)$ . Unfortunately, this estimate is always an upper bound on  $\lambda_{\min}(H)$ . In fact, we have the following monotonicity result.

Stopping criteria for iterations in finite element methods

**Lemma 2** Let  $H_k^S = (H_k + H_k^T)/2$ . Then

$$\lambda_{\min}(H_{k+1}^S) \leq \lambda_{\min}(H_k^S).$$

*Proof*

$$\begin{aligned} \lambda_{\min}(H_k^S) &= \min_{\mathbf{q}_k \in \mathbb{R}^k} \frac{\mathbf{q}_k^T H_k \mathbf{q}_k}{\|\mathbf{q}_k\|^2} \\ &= \min_{\mathbf{q}_k \in \mathbb{R}^k} \frac{\mathbf{q}_k^T V_k^T A V_k \mathbf{q}_k}{\|\mathbf{q}_k\|^2} \\ &= \min_{\mathbf{r} \in \mathcal{K}_k} \frac{\mathbf{r}^T A \mathbf{r}}{\|V_k^T \mathbf{r}\|^2} \\ &\geq \min_{\mathbf{r} \in \mathcal{K}_{k+1}} \frac{\mathbf{r}^T A \mathbf{r}}{\|V_k^T \mathbf{r}\|^2} \end{aligned}$$

Now, any  $\mathbf{r} \in \mathcal{K}_{k+1}$  can be written as  $\mathbf{r} = V_{k+1} \mathbf{q}_{k+1}$  for some  $\mathbf{q}_{k+1} \in \mathbb{R}^{k+1}$  and hence

$$\begin{aligned} \lambda_{\min}(H_k^S) &\geq \min_{\mathbf{q}_{k+1} \in \mathbb{R}^{k+1}} \frac{\mathbf{q}_{k+1}^T V_{k+1}^T A V_{k+1} \mathbf{q}_{k+1}}{\|V_k^T V_{k+1} \mathbf{q}_{k+1}\|^2} \\ &\geq \min_{\mathbf{q}_{k+1} \in \mathbb{R}^{k+1}} \frac{\mathbf{q}_{k+1}^T H_{k+1} \mathbf{q}_{k+1}}{\|\mathbf{q}_{k+1}\|^2} \\ &= \lambda_{\min}(H_{k+1}^S). \end{aligned}$$

□

This result enables us to approximate the stopping criteria as follows. Since by the previous lemma  $\lambda_{\min}(H_k^S) \searrow \lambda_{\min}(H)$  monotonically, there exists a  $k^*$  and a constant  $C^* = C^*(k^*)$  such that  $\lambda_{\min}(H) \geq C^* \lambda_{\min}(H_k^S)$  for all  $k > k^*$ . Hence, our stopping criterion becomes

$$(19) \quad \|\mathbf{r}^k\| \leq C^* \lambda_{\min}^{1/2}(H_k^S) \eta C(h).$$

Thus, we only have to compute  $\lambda_{\min}(H_k^S)$  and estimate  $C^*$ . In practice, the constant  $C^*$  is of order one for small values of  $k^*$ . We investigate this issue in the next section.

*Remark 6* Estimating  $\lambda_{\min}(H_k^S)$  can be done easily in the case of the FOM algorithm. However, in the case of GMRES this is not necessarily straightforward, since we do not store  $H_k$  but the  $R_k$  factor of the  $QR$ -factorization of  $\tilde{H}_k$ . In this case, a further approximation could be introduced

$$\lambda_{\min}(H_k^S) \leq \sigma_{\min}(H_k) \leq \sigma_{\min}(\tilde{H}_k) = \sigma_{\min}(R_k)$$

leading to the bound

$$\|\mathbf{r}^k\| \leq C^* \sigma_{\min}^{1/2}(R_k) \eta C(h),$$

where  $C^*$  is a constant which accounts for both convergence to  $\lambda_{\min}(H_k^s)$  and the difference between  $\lambda_{\min}(H_k^s)$  and  $\sigma_{\min}(R_k)$ , which cannot be guaranteed to be small and is not known a priori. However, this latter bound is useful in estimating  $\|\mathbf{r}^k\|_{A^{-1}}$ .

#### 4.2 Estimation of $\|\mathbf{r}^k\|_{A^{-1}}$

In this case we proceed similarly

$$\|\mathbf{r}^k\| \leq \|\mathbf{r}^k\| \sigma_{\min}^{-1/2}(A).$$

A similar monotonicity result holds for the singular values of  $\tilde{H}_k$  (cf. [14, Cor. 3.1.3])

$$\sigma_{\min}(\tilde{H}_k) \geq \sigma_{\min}(\tilde{H}_{k+1})$$

and thus there exists a  $k^*$  and a constant  $c^* = c^*(k^*)$  such that  $\sigma_{\min}(A) \leq c^* \sigma_{\min}(R_k)$  for all  $k > k^*$ . Thus the stopping criterion (18) can be replaced with

$$(20) \quad \|\mathbf{r}^k\| \leq c^* \sigma_{\min}^{1/2}(R_k) \eta C(h).$$

where, as before,  $c^*$  is a constant (of order one) which we need to estimate.

*Remark 7* We note that this criterion can be used both in the case of GMRES and FOM, since in the first case the matrix  $R_k$  is available and in the second case  $\tilde{H}_k$  is available (with  $\sigma_{\min}(\tilde{H}_k) = \sigma_{\min}(R_k)$ ).

#### 4.3 Restarted GMRES/FOM

There are many situations where the construction of an orthonormal basis for the Krylov subspace is limited to a small number of vectors. This leads to the restarted versions of GMRES or FOM. From the point of view of the above stopping criteria, this does not pose any major problems – we still need to estimate either  $\lambda_{\min}(H)$  or  $\sigma_{\min}(A)$  and this is done in a similar fashion. Thus, assuming we run the algorithms for  $m$  iterations of  $k$  steps each, we use the following approximations

$$(21) \quad \lambda_{\min}(H) \sim \min_{1 \leq i \leq m} \lambda_{\min}^{(i)}(H_k^s), \quad \sigma_{\min}(H) \sim \min_{1 \leq i \leq m} \sigma_{\min}^{(i)}(R_k)$$

where we denote by  $\lambda^{(i)}(H_k^s)$ ,  $\sigma^{(i)}(R_k)$ , the eigenvalues and singular values of the indicated matrices constructed at the  $i$ th iteration.

#### 4.4 The effect of preconditioning

In the case where a preconditioner is used, the Arnoldi algorithm constructs a similar factorization of the preconditioned matrix. We consider here only the case of right preconditioning for which the GMRES/FOM residual remains unchanged. The factorization is

$$AP^{-1}V_k = V_{k+1}\tilde{H}_k$$

and since

$$\|\mathbf{r}\|_{A^{-1}} \leq \sigma_{\min}^{-1/2}(A)\|\mathbf{r}\| \leq \sigma_{\min}^{-1/2}(AP^{-1})\sigma_{\min}^{-1/2}(P)\|\mathbf{r}\|$$

we can derive a stopping criterion similar to (20) using the approximation  $\sigma_{\min}(AP^{-1}) \sim \sigma_{\min}(R_k)$

$$(22) \quad \|\mathbf{r}^k\| \leq c^* \sigma_{\min}^{1/2}(R_k) \sigma_{\min}^{1/2}(P) \eta C(h).$$

However, this requires the estimation of the smallest singular value of  $P$  which may not be easy to achieve. We address this issue in Section 6.

### 5 A minimum residual algorithm

We have seen that in the case of GMRES estimation of  $\|\mathbf{r}^k\|_{H^{-1}}$  can be done provided the Hessenberg matrix is stored. On the other hand, the more relevant quantity  $\|\mathbf{r}^k\|_{A^{-1}}$  can be estimated quite naturally during the GMRES process. However, there is one situation where working with  $\|\mathbf{r}^k\|_{H^{-1}}$  leads to a three-term recurrence algorithm as well as useful preconditioning. The algorithm solves  $A\mathbf{u} = \mathbf{f}$  by minimizing  $\|\mathbf{f} - A\mathbf{u}\|_{H^{-1}}$  over the Krylov space. This is by no means a novel result and has been previously considered by Concus and Golub [6] and Widlund [21] in the context of preconditioning nonsymmetric matrices by their symmetric part. We consider below the version of this algorithm which minimizes the  $H^{-1}$ -norm of the residual, where  $H$  is the symmetric and positive-definite part of  $A$ .

Consider the modified problem

$$(23) \quad \tilde{A}\tilde{\mathbf{u}} = \tilde{\mathbf{f}}$$

where  $\tilde{A} = H^{-1/2}AH^{-1/2}$ ,  $\tilde{\mathbf{f}} = H^{-1/2}\mathbf{f}$ . Let us consider first the FOM algorithm applied to this system. As before, the residual is orthogonal to the Krylov space

$$\langle \tilde{\mathbf{r}}^k, \mathbf{q} \rangle = \langle H^{-1/2}(\mathbf{f} - A\mathbf{u}^k), \mathbf{q} \rangle = 0, \quad \forall \mathbf{q} \in \tilde{\mathcal{K}}_k$$

where

$$\tilde{\mathcal{K}}_k = \text{span} \left\{ \tilde{\mathbf{r}}^0, \tilde{A}\tilde{\mathbf{r}}^0, \dots, \tilde{A}^{k-1}\tilde{\mathbf{r}}^0 \right\} = H^{1/2}\mathcal{K}_k(H^{-1}A, H^{-1}\mathbf{r}^0).$$

Thus,  $\forall \mathbf{p} \in \mathcal{K}_k(H^{-1}A, H^{-1}\mathbf{r}^0)$

$$\langle H^{-1/2}\mathbf{r}^k, H^{1/2}\mathbf{p} \rangle = \langle \mathbf{r}^k, \mathbf{p} \rangle = \langle H^{-1}\mathbf{r}^k, \mathbf{p} \rangle_H = 0.$$

In other words, the standard FOM algorithm for (23) is also an orthogonal projection method with respect to the  $H$ -inner-product onto  $\mathcal{K}_k(H^{-1}A, H^{-1}\mathbf{r}^0)$ . Moreover, the advantage of this formulation is that there exists a three-term recurrence which solves this problem. We summarize this below.

**Lemma 3** *Let  $A$  have symmetric and positive-definite part  $H$ . The FOM algorithm applied to*

$$(H^{-1/2}AH^{-1/2})(H^{1/2}\mathbf{u}) = H^{-1/2}\mathbf{f}.$$

*in the Euclidean inner-product is equivalent to the FOM algorithm in the  $H$ -inner product applied to*

$$H^{-1}\mathbf{A}\mathbf{u} = H^{-1}\mathbf{f}.$$

*Moreover, the Arnoldi orthogonalization process applied to the normal matrix  $H^{-1/2}AH^{-1/2}$  yields a factorization*

$$V_k^T AV_k = H_k,$$

*where  $(H_k)_{ij} = 0$  for all  $|i - j| > 1$ .*

*Proof* See Appendix. □

This idea is contained in [21], although the author constructs a different tri-diagonalization than that constructed by FOM (Arnoldi). Similarly, using the above result one can modify the standard GMRES algorithm into a three-term recurrence which constructs the solution with the smallest residual over  $\mathcal{K}_k$  as measured in the  $H^{-1}$ -norm. We do not include the details here, but only present in the next section numerical results obtained with this modified version of GMRES.

*Remark 8* The action of the inverse of  $H$  as a preconditioner can be relaxed in practice. Indeed, solving to an accuracy of order  $o(C(h))$  is sufficient for convergence of the algorithm. We explore this issue in the next section.

## 6 Examples

In this section we are interested in establishing explicit stopping criteria for the generic example of finite element approximation of the solution of scalar elliptic equations.

Let  $\Omega \subset \mathbb{R}^d$  with boundary  $\Gamma$ . We will be using the following norms:

$$\begin{aligned}\|v(\mathbf{x})\|_{L^2(\Omega)} &= \|v(\mathbf{x})\|_0 = \left( \int_{\Omega} v(\mathbf{x})^2 d\mathbf{x} \right)^{1/2} \\ \|v(\mathbf{x})\|_{L^\infty(\Omega)} &= \|v(\mathbf{x})\|_\infty = \operatorname{ess\,sup}_{\mathbf{x} \in \Omega} |v(\mathbf{x})| \\ \|v(\mathbf{x})\|_{H^m(\Omega)} &= \|v(\mathbf{x})\|_m = \left( \sum_{|\alpha| \leq m} \int_{\Omega} |D^\alpha v(\mathbf{x})|^2 d\mathbf{x} \right)^{1/2}\end{aligned}$$

where

$$D^\alpha v(\mathbf{x}) = \frac{\partial^{|\alpha|} v(\mathbf{x})}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} = \partial_{x_1}^{\alpha_1} \dots \partial_{x_d}^{\alpha_d}$$

and  $\alpha = (\alpha_1, \dots, \alpha_d)$  is an index of order  $|\alpha| = \alpha_1 + \dots + \alpha_d$ . We also need to define the space  $H_0^1(\Omega)$

$$H_0^1(\Omega) = \{v(\mathbf{x}) \in H^1(\Omega) : v(\mathbf{x})|_\Gamma = 0\}$$

with norm

$$|v(\mathbf{x})|_{H_0^1(\Omega)} = |v(\mathbf{x})|_1 = \left( \sum_{|\alpha|=1} \int_{\Omega} |D^\alpha v(\mathbf{x})|^2 d\mathbf{x} \right)^{1/2}.$$

### 6.1 Elliptic problems

Consider the general second-order elliptic problem

$$(24a) \quad -\nabla \cdot (\mathbf{a}(\mathbf{x}) \nabla u) + \mathbf{b}(\mathbf{x}) \cdot \nabla u + c(\mathbf{x})u = f \text{ in } \Omega \subset \mathbb{R}^d$$

$$(24b) \quad u = 0 \text{ on } \Gamma.$$

where the matrix  $\mathbf{a}(\mathbf{x})$  is symmetric and positive definite for all  $\mathbf{x} \in \Omega$ , i.e.,

$$k_2(\mathbf{x}) |\boldsymbol{\xi}|^2 \leq \boldsymbol{\xi}^T \mathbf{a}(\mathbf{x}) \boldsymbol{\xi} \leq k_1(\mathbf{x}) |\boldsymbol{\xi}|^2$$

for some functions  $k_1(\mathbf{x}), k_2(\mathbf{x})$ . We also assume that the coefficients are bounded, i.e.,  $(\mathbf{a})_{ij}, (\mathbf{b})_i, c \in \mathbf{L}^\infty(\Omega)$ ,  $i, j = 1, \dots, d$ , and that the following condition holds

$$c(\mathbf{x}) - \frac{1}{2} \nabla \cdot \mathbf{b}(\mathbf{x}) \geq 0 \quad \forall \mathbf{x} \in \Omega.$$

The weak formulation seeks a solution  $u \in \mathcal{H} \equiv H_0^1(\Omega)$  such that

$$(25) \quad a(u, v) = f(v) \quad \text{for all } v \in H_0^1(\Omega)$$

where

$$a(w, v) = (\mathbf{a} \cdot \nabla w, \nabla v) + (\mathbf{b} \cdot \nabla w, v) + (cw, v).$$

It is straightforward to show that  $a(\cdot, \cdot)$  satisfies the continuity and coercivity conditions (2) with respect to the  $H_0^1$ -norm  $|\cdot|_1$  with constants

$$C_1 = \|k_1\|_{L^\infty(\Omega)} + \|\mathbf{b}\|_{L^\infty(\Omega)} + C(\Omega)\|c\|_{L^\infty(\Omega)}, \quad C_2 = \min_{\mathbf{x} \in \Omega} k_2(\mathbf{x}),$$

where  $C(\Omega)$  is a constant of order one which depends only on the domain.

Let now  $\mathcal{H}_h \subset \mathcal{H}$  be a space of piecewise polynomials defined on a partition  $\mathcal{T}_h$  of  $\Omega$  into simplices  $T$  of diameter  $h_T$ . As described in Section (2.2) the inclusion  $\mathcal{H}_h \subset \mathcal{H}$  ensures that the stability conditions (13a) and (13b) are satisfied with the constants  $C_1, C_2$  defined above. Moreover, discretizing (25) as

$$\mathbf{A}u = \mathbf{f},$$

the constants  $C_2, C_3$  in (13) are given as follows. If we choose to monitor the error with respect to  $|\cdot|_1$  then

$$C_3 = \|k_1\|_{L^\infty(\Omega)} + C(\Omega)\|c\|_{L^\infty(\Omega)}, \quad C_2 = \min_{\mathbf{x} \in \Omega} k_2(\mathbf{x}).$$

However, if we work with the energy norm defined by

$$(26) \quad |||w||| = a(w, w),$$

then  $C_2 = C_3 = 1$ .

## 6.2 Numerical experiments

To illustrate the ideas presented above, we chose to perform experiments on a 2D advection-diffusion problem ( $c = 0$ ). In particular, we chose to study the robustness of our stopping criteria with respect to the nonsymmetry in the problem. Thus, we solved a test problem for constant diffusivity tensors

$$\mathbf{a}(\mathbf{x}) = \nu I,$$

where the diffusion parameter  $\nu$  toggles the degree of nonsymmetry of the matrices involved. The test problem is thus

$$(27a) \quad -\nu \nabla^2 u + \mathbf{b}(x, y) \cdot \nabla u = f \text{ in } \Omega \equiv (-1, 1)^2$$

$$(27b) \quad u = 0 \text{ on } \Gamma,$$

with

$$b(x, y) = \begin{pmatrix} 2y(1 - x^2) \\ -2x(1 - y^2) \end{pmatrix}$$

and right-hand side  $f$  such that the solution  $u$  is

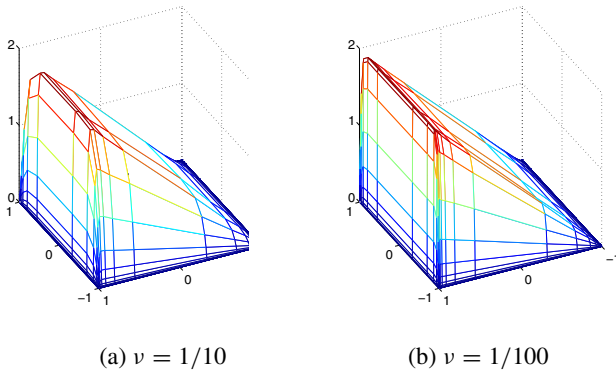
$$u(x, y) = \left( 1 - \frac{e^{(x-1)/\sqrt{v}} + e^{(-x-1)/\sqrt{v}}}{1 + e^{-2/\sqrt{v}}} \right) \cdot \left( 1 + y - 2 \frac{e^{(y-1)/v} + e^{(-2)/v}}{1 - e^{-2/v}} \right).$$

This choice of solution tries to mimick the behaviour of problems where boundary layers are present (see Fig. 1).

We consider here only the errors with respect to the  $H_0^1$ -norm, as the case where the energy norm (26) is employed yields similar results. We denote by  $u^I$  the linear interpolant of the solution at the mesh points. Our numerical results below will display the following estimators and errors:

- (i) FE: the exact relative (forward) errors  $|u - u_h^k|_1 / |u_h^k|_1$ ;
- (ii) FIE: the exact relative (forward) interpolation errors  $|u^I - u_h^k|_1 / |u_h^k|_1$ ;
- (iii) HINV: the exact  $H^{-1}$ -norm criterion (17)  $\eta C_2^{-1} \|\mathbf{r}^k\|_{H^{-1}} / \|\mathbf{u}^k\|_H$ ;
- (iv) AINV: the exact  $A^{-1}$ -norm criterion (18)  $\eta C_2^{-1/2} \|\mathbf{r}^k\|_{A^{-1}} / \|\mathbf{u}^k\|_H$ ;
- (v) HINV-est: the estimated  $H^{-1}$ -norm criterion (19)  $\eta C_2^{-1} \|\mathbf{r}^k\|_{\lambda_{\min}^{-1/2}(H_k^s)} / \|\mathbf{u}^k\|_H$ ;
- (vi) AINV-est: the estimated  $A^{-1}$ -norm criterion (20)  $\eta C_2^{-1/2} \|\mathbf{r}^k\|_{\sigma_{\min}^{-1/2}(H_k^s)} / \|\mathbf{u}^k\|_H$ ;
- (vii) the standard 2-norm stopping criterion  $\|\mathbf{r}^k\| / \|\mathbf{r}^0\|$ .

In all cases we chose the constants  $c^* = C^* = 1$  in (19), (20). The choice of  $\eta$  is related to the derivation of our criteria. Indeed, an informed choice would be a value of  $\eta$  which is  $o(h)$ . In our tests we used a value  $\eta = 0.15$  which is roughly the square root of the second coarsest mesh parameter. We found this choice to be robust for all parameter ranges.

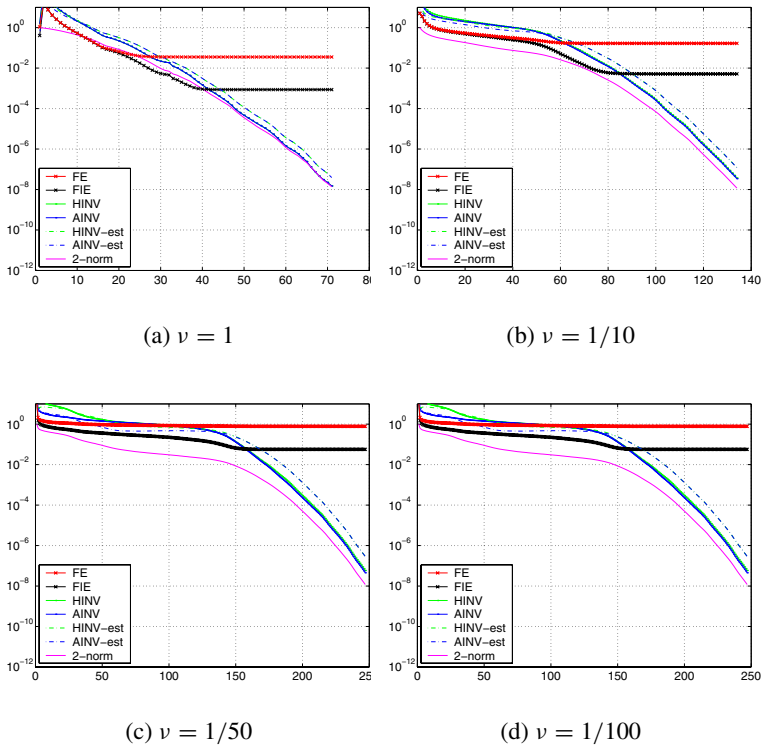


**Fig. 1.** Solution of advection-diffusion problem

### 6.3 GMRES without preconditioning

We begin with the case of a uniform partition of  $\Omega$  into squares of size  $h$  and bilinear basis functions. The GMRES convergence curves and derived bounds are displayed in Fig. 2.

As expected, the iterative process measured in the energy norm (the FE/FIE curve) exhibits a plateau which indicates that an approximate solution has been found with corresponding finite element error given by the level of the plateau. This level represents the functional backward error, which can only be reduced through a better choice of finite elements (such as that obtained through further refinement). From a practical point of view, we would ideally like to stop the iteration at the onset of this plateau, since no further improvement is obtained with respect to this norm. This requires information about the order of the finite element error. This is many times available either asymptotically through a priori studies or numerically through a posteriori error estimation on coarser meshes. We return to this issue at the end of this section.



**Fig. 2.** Comparison of stopping criteria for GMRES;  $h = 1/16$

We see that in all experiments the suggested criteria provide upper bounds for the quantities of interest – in this case, the  $H_0^1$ -norm of the error. Moreover, the achievable finite element error (the onset of plateau) require considerably fewer iterations than a stopping criterion such as the relative Euclidean norm of the residual being brought below  $10^{-8}$  (standard threshold).

Another remarkable fact is that the criterion (17) based on the dual norm of the residual is an upper bound for the interpolation error. The reason for this is not so surprising since in standard finite element calculations the interpolation error is usually smaller than the error in the energy or related norms (sometimes by a factor of  $h$ ). Thus, our stopping criterion gives an upper bound on both errors so that the iterates have either achieved the final error or have achieved an error bounded from above by our dual norm estimate. We also note here that the constants  $c^*$ ,  $C^*$  are indeed of order one for all values of  $\nu$ . This robustness also holds with respect to the mesh parameter.

The same convergence curves for the case of restarted GMRES are displayed in Fig. 3. They exhibit indeed the most dramatic difference between

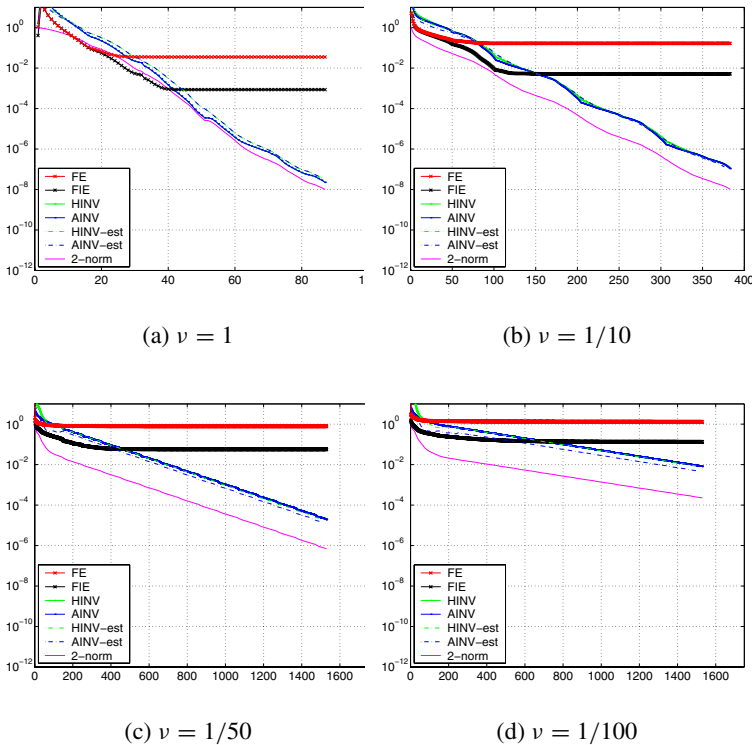


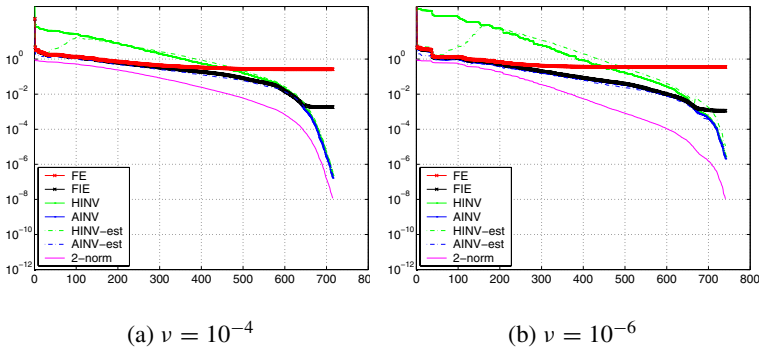
Fig. 3. Comparison of stopping criteria for GMRES(50);  $h = 1/16$

convergence in the  $H^{-1}$ - or  $A^{-1}$ -norm and the standard 2-norm criterion. Again, the estimation of the relevant convergence curves based on the approximation (21) works extremely well. Moreover, the difference between the two stopping criteria (17), (18) is negligible. However, this may not always be the case. Indeed, the equivalence between the two norms described by Lemma 1 deteriorates if the  $H$ -condition number of the problem deteriorates. For our test problem this happens when  $\nu$  is small *and* the discretization is nonuniform. We consider this case below.

For small values of  $\nu$ , the problem becomes more nonsymmetric, with the matrices more nonnormal. At the same time, the finite element error on uniform meshes of squares deteriorates and even becomes of order one. One way to avoid this is to refine the mesh suitably. Given the boundary layers in the solution, we chose an exponential refinement of the meshes. In this case, the parameter  $h$  is not defined in (19), (20) – we chose  $h := \|\mathbf{f}\|_M / \|\mathbf{f}\|$ , where  $M = (\phi_i, \phi_j)$  is the Gramian (mass) matrix with respect to the  $L^2(\Omega)$ -inner product  $(\cdot, \cdot)$ . The convergence curves are displayed in Fig. 4. Again, we see that the two norms of interest are approximated well; however, the exact convergence curves in the  $H^{-1}$ - and  $A^{-1}$ -norms are not close in the initial phase of the iterative process, but become almost identical close to the convergence stage.

We end this section with an illustration on the use of a priori error estimates to guide the stopping of the iterative process in the case of uniform meshes. In particular, following ([5]) we used in (19)  $C(h) = h/\sqrt{\nu}$  as stopping criteria for the energy error (FE). Similarly, for the interpolation error (FIE) we chose  $C(h) = h^2/\sqrt{\nu}$ . We chose to describe the resulting criteria in terms of the achieved ratio of the final error after  $k$  iterations

$$\rho^{(k)} = \frac{\|u - u_h\|_1}{\|u - u_h^k\|_1}$$



**Fig. 4.** Comparison of stopping criteria for GMRES – exponentially-stretched mesh

and the savings in terms of iterations using as reference the number of iterations  $K$  required by the standard criterion that the relative Euclidean norm of the residual be brought below a tolerance of  $10^{-8}$

$$\sigma^{(k)} = \frac{K - k}{K}.$$

The results are presented in Table 1 for a range of meshes and values of  $\nu$ . We chose to compare our stopping criterion with what we call the ‘ $\alpha$ 100%-converged case’, i.e., the onset of the plateau for the FE curve with a guaranteed fraction  $\alpha$  of accuracy. In particular we identified this nearly-converged case as the iteration  $k$  for which the exact error (FE curve) satisfies  $\rho^{(k)} \geq \alpha$ . In our experiments we chose  $\alpha = 0.98$ .

We see that in all cases the potential savings are between 40–60%, with greater reductions possible on coarser meshes. This is, of course, natural: as the mesh is refined, the error is reduced and the plateau occurs at levels closer and closer to that of the Euclidean residual criterion. However, in practice, for many problems of interest, the finite element error seldom reaches these high levels of accuracy.

Table 1 also reveals that using a simple a priori criterion to estimate the finite element error can work rather well for discretizations on quasi-uniform meshes. This, again, is the case in particular on coarser meshes. With the exception  $\nu = 1/100$ , for all values of  $\nu$  our criterion came within 7% of the potential savings.

Similar behaviour is noticed in the case where the interpolation error is taken as a reference level – in this case we use the notation  $\rho_I^{(k)}, \sigma_I^{(k)}$  to denote the corresponding savings indicators where

**Table 1.** Full GMRES iterations ( $k$ ), energy error indices ( $\rho^{(k)}$ ) and savings ( $\sigma^{(k)}$ ) for (i) the 98%-converged case and (ii) using stopping criterion (19) with  $C(h) = h/\sqrt{\nu}$

$\nu$	$h$	Exact			Bound			$K$
		$k$	$\rho^{(k)}$	$\sigma^{(k)}$	$k$	$\rho^{(k)}$	$\sigma^{(k)}$	
1	1/16	28	0.984	0.61	29	0.988	0.59	72
	1/32	64	0.983	0.55	66	0.989	0.54	144
	1/64	139	0.981	0.51	146	0.994	0.48	286
1/10	1/16	64	0.981	0.52	69	0.995	0.49	135
	1/32	135	0.981	0.49	149	0.997	0.44	269
	1/64	284	0.981	0.47	316	0.998	0.40	534
1/50	1/16	138	0.981	0.44	144	0.990	0.42	247
	1/32	283	0.980	0.43	316	0.997	0.36	498
	1/64	572	0.980	0.40	646	0.999	0.30	961
1/100	1/16	191	0.981	0.40	192	0.981	0.39	318
	1/32	371	0.980	0.43	420	0.995	0.35	648
	1/64	786	0.980	0.40	905	0.998	0.31	1313

$$\rho_I^{(k)} = \frac{\|u^I - u_h\|_1}{\|u^I - u_h^k\|_1}$$

with  $u^I$  the linear interpolant of the exact solution  $u$ . Since for our application the interpolation error tends to be one factor of  $h$  smaller, the number of iterations required to satisfy our criterion will be greater than in the case where the energy was the relevant quantity. The results are displayed in Table 2. While smaller, the potential savings remain important (between 25–45%) and are approximated somewhat less accurately with our criterion based on a priori error estimates (within 10% of potential savings). Moreover, we note a rather negligible (up to 0.025%) loss in robustness for larger values of  $\nu$ , corresponding to the cases where the termination occurs up to 4 iterations less than the in the ‘optimal case’.

#### 6.4 Preconditioned GMRES

We turn now to the case where preconditioning is employed to speed up the iteration process. As specified in Section 4, we consider only the case of right preconditioning, which has the advantage of preserving the residual, a property which enabled us to derive the stopping criterion (22). However, the use of this stopping criterion requires the estimation of the smallest singular value of our preconditioner  $P$ . In some cases this estimation can be performed cheaply, but in general it may be quite difficult to provide this information. The approximation we use is described below.

**Table 2.** Full GMRES iterations ( $k$ ), interpolation error indices ( $\rho_I^{(k)}$ ) and savings ( $\sigma_I^{(k)}$ ) for (i) the 98%-converged case and (ii) using stopping criterion (19) with  $C(h) = h^2/\sqrt{\nu}$

$\nu$	$h$	$k$	Exact		$k$	Bound		$K$
			$\rho_I^{(k)}$	$\sigma_I^{(k)}$		$\rho_I^{(k)}$	$\sigma_I^{(k)}$	
1	1/16	41	0.988	0.43	40	0.975	0.44	72
	1/32	91	0.984	0.37	91	0.984	0.37	144
	1/64	195	0.981	0.32	198	0.990	0.31	286
1/10	1/16	89	0.980	0.34	88	0.979	0.35	135
	1/32	194	0.980	0.28	190	0.969	0.29	269
	1/64	410	0.980	0.23	406	0.975	0.24	534
1/50	1/16	156	0.984	0.37	178	1.001	0.28	247
	1/32	35	0.980	0.32	385	1.000	0.23	498
	1/64	704	0.982	0.26	788	0.999	0.18	961
1/100	1/16	207	0.981	0.35	237	0.999	0.26	318
	1/32	445	0.980	0.31	510	0.999	0.21	648
	1/64	972	0.980	0.26	1093	1.000	0.17	1313

All preconditioning techniques require the solution of a linear system involving the preconditioner matrix  $P$ :

$$P\mathbf{z} = \mathbf{v}.$$

Since

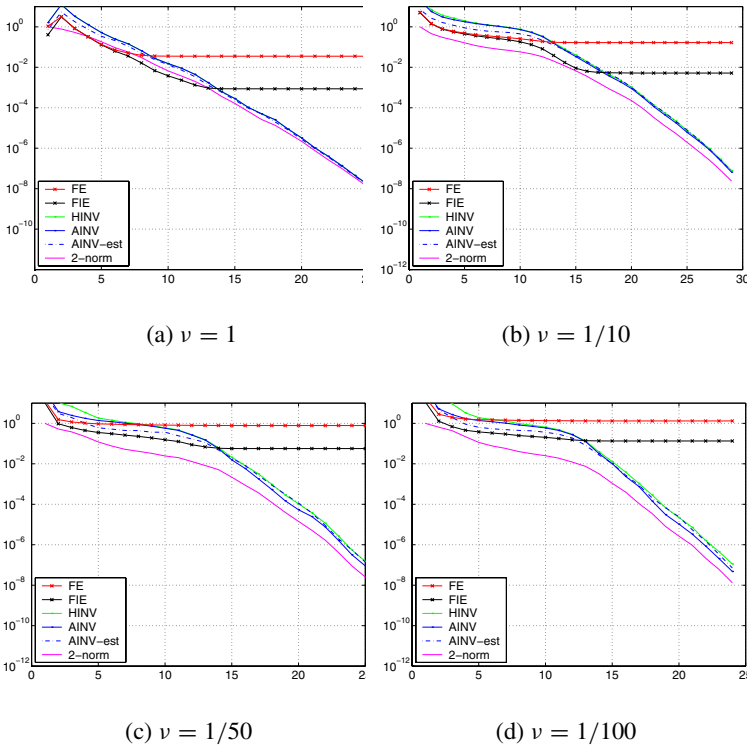
$$\sigma_{\min}(P) = \sigma_{\max}^{-1}(P^{-1}) = \left( \max_{\mathbf{v} \in \mathbb{R}^n \setminus \{0\}} \frac{\|P^{-1}\mathbf{v}\|}{\|\mathbf{v}\|} \right)^{-1}$$

we choose to approximate  $\sigma_{\min}(P)$  via

$$\sigma_{\min}(P) \sim \left( \max_k \frac{\|\mathbf{z}^k\|}{\|\mathbf{v}^k\|} \right)^{-1} = \min_k \frac{\|\mathbf{v}^k\|}{\|\mathbf{z}^k\|}$$

where  $\mathbf{z}^k = P^{-1}\mathbf{v}^k$ . In the case of GMRES, the vector  $\mathbf{v}^k$  is the vector generated by the Arnoldi process, so that  $\|\mathbf{v}^k\| = 1$ .

The performance of GMRES with ILU preconditioning is displayed in Fig. 5. While the number of iterations is greatly reduced, the convergence



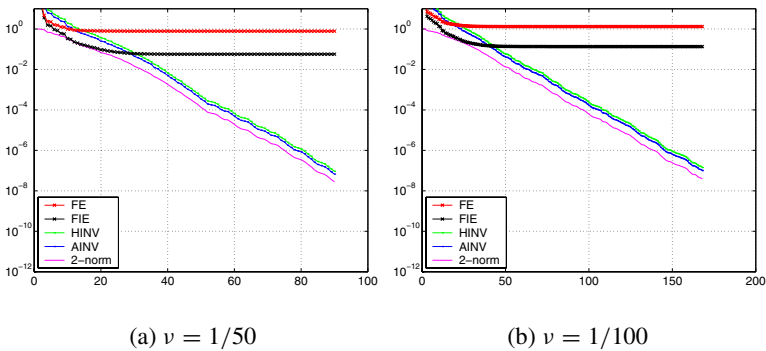
**Fig. 5.** Comparison of stopping criteria for GMRES with ILU(0) preconditioning;  $h = 1/16$

behaviour is similar to the unpreconditioned case. Moreover, the approximation of the residual  $A^{-1}$ -norm described above appears to work extremely well. However, in general we expect over- or under-estimation to occur, in which case alternative methods for the estimation of the smallest singular value of  $P$  may have to be employed.

### 6.5 Three-term GMRES

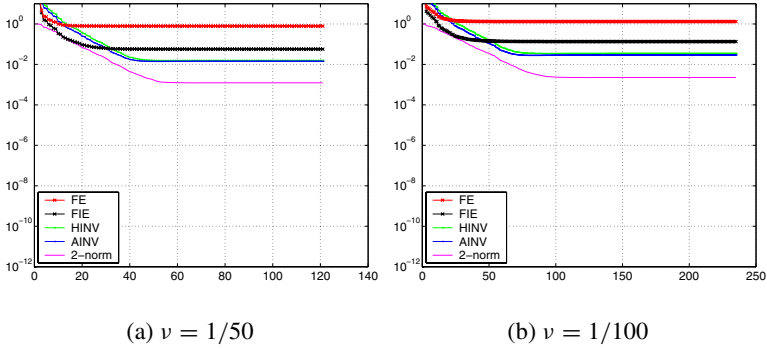
We end this section with numerical results obtained with the minimum residual algorithm based on a three-term recurrence described in Section 4. We recall here that this is essentially the GMRES algorithm implemented in the  $H$ -norm with left-preconditioner  $H$ . The norm of the modified residual in this method is the quantity we seek,  $\|\mathbf{r}^k\|_{H^{-1}}$ . The results are displayed in Fig. 6. As before, our stopping criterion (17) provides an upper bound for the convergence of quantities of interest, such as  $H_0^1$ -norm of the error or the interpolation error. More remarkable, though, is the fact that in this case the solver yields iterates whose 2-norm residual traces closely the convergence curves of interest. This is a phenomenon also noticed in the case of a similar GMRES implementation used for the solution of flow problems [15].

The same experiments were run with inexact implementation of the preconditioner  $H$ . More precisely, we solved systems with  $H$  using CG with an incomplete Cholesky preconditioner and a stopping criterion as described in Arioli [1]; the tolerance was chosen to be of order  $h^{5/2}$ , which for this problem is  $h^{1/2}$  less than the order of the interpolation error. The results are displayed in Fig. 7. We see indeed that our criterion is an upper bound for both the finite element error  $|u - u_h^k|_1$  and the interpolation error  $|u^I - u_h^k|_1$ . Moreover, the



**Fig. 6.** Comparison of stopping criteria for  $H$ -norm minimum residual algorithm: exact preconditioning

## Stopping criteria for iterations in finite element methods

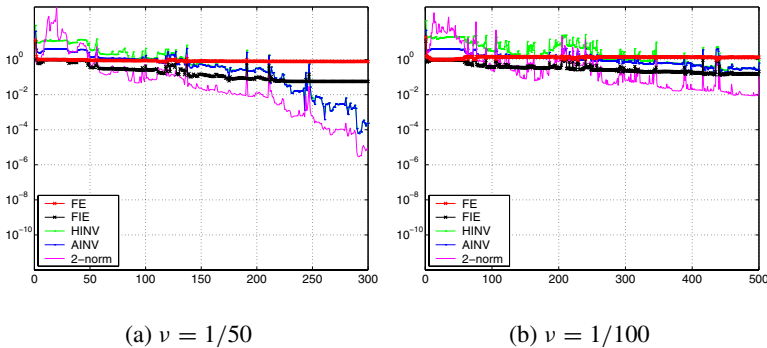


**Fig. 7.** Comparison of stopping criteria for  $H$ -norm minimum residual algorithm: inexact preconditioning

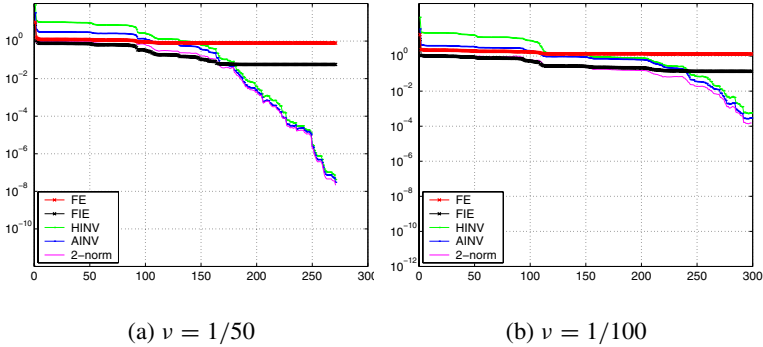
inexact solves do not affect the convergence curve in the regime where it is relevant.

### 6.6 Other iterative methods

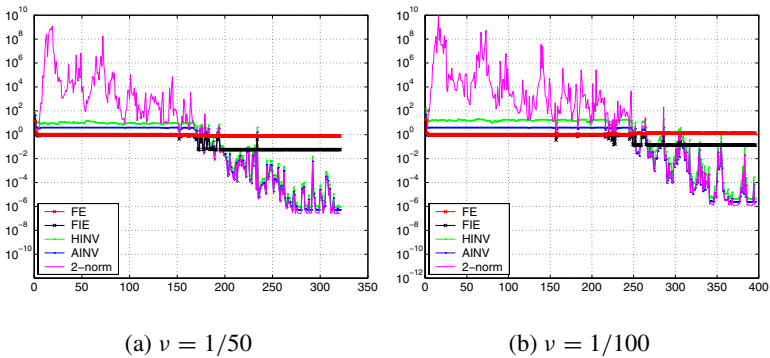
In order to test further the relevance of the  $A^{-1}$ - and  $H^{-1}$ -norms of the residual, we ran experiments with BICGSTAB, QMR and CGS. The results for the case of discretization on uniform meshes are displayed in Figs 8, 9, 10. We again see that in all cases the two residuals provide upper bounds for the energy norm of the error and interpolation error. In particular, the  $A^{-1}$ -norm provides again the closest approximation to these quantities, with the  $H^{-1}$ -norm bound possibly deteriorating for more nonsymmetric problems,



**Fig. 8.** Comparison of stopping criteria for BICGSTAB:  $h=1/16$



**Fig. 9.** Comparison of stopping criteria for QMR:  $h=1/16$



**Fig. 10.** Comparison of stopping criteria for CGS:  $h=1/16$

as Fig. 9 shows. As for the 2-norm of the residual, the behaviour oscillates between the smooth, relevant convergence curve of QMR to the oscillating, large residuals exhibited by CGS. However, the issue of dynamic estimation of the  $A^{-1}$ - and  $H^{-1}$ -norms is not as straightforward as in the case of GMRES.

## 7 Conclusion

The message of this paper is simple: do not accurately compute the solution of an inaccurate problem. This was highlighted already in [1] for the case of symmetric and positive-definite problems – our contribution here was the generalization to the case of nonsymmetric problems. The proposed stopping criteria require the calculation of the residual in a norm related to the problem formulation. It also requires information about the discretization error which

may not always be available. In practice, a priori error estimates on quasi-uniform meshes or a posteriori error calculations may provide sufficient for this purpose. The former estimator was successfully employed in our test case. The latter appears also to be promising [9].

Overall, we demonstrated that the suggested criteria are relevant to convergence in the energy-norm (or equivalent norms) while at the same time highlighting the fact that the standard criterion based on the Euclidean norm of the residual has no relevance to the quantities of interest and is in general wasteful. Further generalizations of these ideas include the case of indefinite problems and mixed finite element discretizations of systems of partial differential equations, where the use of mixed norms in which to measure convergence arises quite naturally. We hope to address some of these issues in a future paper.

*Acknowledgements.* We thank Serge Gratton for useful discussions and comments.

## 8 Appendix

*Proof of Lemma 1* We need to show

$$\frac{1}{\sqrt{C_3}} \|\mathbf{r}\|_A \leq \|\mathbf{r}\|_H \leq \frac{1}{\sqrt{C_2}} \|\mathbf{r}\|_A$$

and

$$\frac{\sqrt{C_2}}{C_1 C_3} \|\mathbf{r}\|_{H^{-1}} \leq \|\mathbf{r}\|_{A^{-1}} \leq \frac{1}{\sqrt{C_2}} \|\mathbf{r}\|_{H^{-1}}.$$

The first equivalence is just a restating of the discrete stability conditions (13b), (13c). For the second we have

$$\begin{aligned} \frac{\mathbf{r}^T A^{-1} \mathbf{r}}{\mathbf{r}^T H^{-1} \mathbf{r}} &\leq \sigma_1(H^{1/2} A^{-1} H^{1/2}) \\ &= \sigma_n^{-1}(H^{-1/2} A H^{-1/2}) \\ &\leq \left( \min_{\mathbf{x} \in \mathbb{R}^n \setminus \{0\}} \frac{\mathbf{x}^T H^{-1/2} A H^{-1/2} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} \right)^{-1} \\ &= \left( \min_{\mathbf{y} \in \mathbb{R}^n \setminus \{0\}} \frac{\mathbf{y}^T A \mathbf{y}}{\mathbf{y}^T H \mathbf{y}} \right)^{-1} \\ &\leq C_2^{-1} \end{aligned}$$

by using (13b). Finally, since

$$C_2 \leq \frac{\mathbf{r}^T A \mathbf{r}}{\mathbf{r}^T H \mathbf{r}} = \frac{\mathbf{r}^T H_A \mathbf{r}}{\mathbf{r}^T H \mathbf{r}} \leq C_3,$$

we have

$$\frac{\mathbf{r}^T A^{-1} \mathbf{r}}{\mathbf{r}^T H^{-1} \mathbf{r}} = \frac{\mathbf{r}^T A^{-1} \mathbf{r}}{\mathbf{r}^T H_A^{-1} \mathbf{r}} \cdot \frac{\mathbf{r}^T H_A^{-1} \mathbf{r}}{\mathbf{r}^T H^{-1} \mathbf{r}} \geq C_3^{-1} \min_{\mathbf{r} \in \mathbb{R}^n \setminus \{0\}} \frac{\mathbf{r}^T A^{-1} \mathbf{r}}{\mathbf{r}^T H_A^{-1} \mathbf{r}} \geq C_3^{-1} \min_{\mathbf{y} \in \mathbb{R}^n \setminus \{0\}} \frac{\mathbf{y}^T \tilde{A}^{-1} \mathbf{y}}{\mathbf{y}^T \mathbf{y}},$$

where  $\tilde{A} = I + \tilde{N}$ ,  $\tilde{N} = H_A^{-1/2} S_A H_A^{-1/2}$ . Since  $\tilde{A}$  (and thus  $\tilde{A}^{-1}$ ) is a normal matrix, its field of values is the convex hull of its eigenvalues ([13, p.11]). Hence,

$$\min_{\mathbf{y} \in \mathbb{R}^n \setminus \{0\}} \frac{\mathbf{y}^T \tilde{A}^{-1} \mathbf{y}}{\mathbf{y}^T \mathbf{y}} = \min_k \operatorname{Re} \frac{1}{\lambda_k(\tilde{A})} = \min_k \operatorname{Re} \frac{1}{1 + \lambda_k(\tilde{N})} = \frac{1}{\max_k \left| \lambda_k(\tilde{A}) \right|^2}$$

and since  $\|H^{-1/2} A H^{-1/2}\| = \|A\|_{H, H^{-1}} = C_1$  (cf. (11), (15)) we get

$$\max_k \left| \lambda_k(\tilde{A}) \right| \leq \|H_A^{-1/2} A H_A^{-1/2}\| \leq \|H^{-1/2} A H^{-1/2}\| \kappa_2(H_A^{-1/2} H^{1/2})$$

and the result follows.  $\square$

*Proof of Lemma 3* Consider the two equivalent linear systems

$$\tilde{A} \tilde{\mathbf{u}} = \tilde{\mathbf{f}}, \quad \hat{A} \mathbf{u} = \hat{\mathbf{f}}$$

where

$$\tilde{A} = H^{-1/2} A H^{-1/2}, \quad \tilde{\mathbf{u}} = H^{1/2} \mathbf{u}, \quad \tilde{\mathbf{f}} = H^{-1/2} \mathbf{f}, \quad \hat{A} = H^{-1} A, \quad \hat{\mathbf{f}} = H^{-1} \mathbf{f}.$$

The first part of the Lemma follows from the equivalence of the Arnoldi algorithms below.

<b>Arnoldi in</b> $(\cdot, \cdot)$	<b>Arnoldi in</b> $(\cdot, \cdot)_H$
$\tilde{\mathbf{v}}_1 := \tilde{\mathbf{r}}_0 / \ \tilde{\mathbf{r}}_0\ $	$\hat{\mathbf{v}}_1 := \hat{\mathbf{r}}_0 / \ \hat{\mathbf{r}}_0\ _H$
<b>do</b> $j = 1, 2, \dots, m$	<b>do</b> $j = 1, 2, \dots, m$
$\tilde{h}_{ij} = (\tilde{A} \tilde{\mathbf{v}}_j, \tilde{\mathbf{v}}_i), 1 \leq i \leq j$	$\hat{h}_{ij} = (\hat{A} \hat{\mathbf{v}}_j, \hat{\mathbf{v}}_i)_H, 1 \leq i \leq j$
$\tilde{\mathbf{w}}_j = \tilde{A} \tilde{\mathbf{v}}_j - \sum_i^j \tilde{h}_{ij} \tilde{\mathbf{v}}_i$	$\hat{\mathbf{w}}_j = \hat{A} \hat{\mathbf{v}}_j - \sum_i^j \hat{h}_{ij} \hat{\mathbf{v}}_i$
<b>if</b> $\tilde{h}_{j+1,j} = \ \tilde{\mathbf{w}}_j\  = 0$ <b>stop</b>	<b>if</b> $\hat{h}_{j+1,j} = \ \hat{\mathbf{w}}_j\ _H = 0$ <b>stop</b>
$\tilde{\mathbf{v}}_{j+1} = \tilde{\mathbf{w}}_j / \tilde{h}_{j+1,j}$	$\hat{\mathbf{v}}_{j+1} = \hat{\mathbf{w}}_j / \hat{h}_{j+1,j}$
<b>end do</b>	<b>end do</b>

where  $\tilde{\mathbf{v}}_i = H^{1/2} \hat{\mathbf{v}}_i$ ,  $\tilde{\mathbf{w}}_i = H^{1/2} \hat{\mathbf{w}}_i$ ,  $\tilde{\mathbf{r}}_0 = H^{-1/2} \mathbf{r}^0$ ,  $\hat{\mathbf{r}}_0 = H^{-1} \mathbf{r}^0$ ,  $\mathbf{r}^0 = \mathbf{f} - A \mathbf{u}^0$  for some initial guess  $\mathbf{u}^0$ . In particular, they yield the same Hessenberg matrices since

$$\tilde{h}_{ij} = (\tilde{A} \tilde{\mathbf{v}}_j, \tilde{\mathbf{v}}_i) = (H^{-1/2} A \hat{\mathbf{v}}_j, H^{1/2} \hat{\mathbf{v}}_i) = (H^{-1} A \hat{\mathbf{v}}_j, \hat{\mathbf{v}}_i)_H = \hat{h}_{ij}.$$

For the second part, we work with the Arnoldi algorithm in the Euclidean inner-product and system matrix  $\tilde{A} = I + N$ , where  $N = -N^T$  is skew-symmetric. For ease of presentation we drop the  $\sim$ 's. We now prove by induction on  $j$  that for all  $i \leq j - 2$

$$h_{ij} = \mathbf{v}_i^T (I + N) \mathbf{v}_j = 0.$$

We first note that (i)  $h_{ij} = 1$  if  $i = j$ , (ii)  $h_{ij} = \mathbf{v}_i^T N \mathbf{v}_j$  if  $i < j$  and (iii)  $\mathbf{v}_i^T N^3 \mathbf{v}_i = 0$ , since  $N$  is skew-symmetric. Since  $\mathbf{w}_1 = N \mathbf{v}_1$  and  $\mathbf{w}_2 = N \mathbf{v}_2 - h_{12} \mathbf{v}_1$  we have using (i)-(iii)

$$h_{13} = \mathbf{v}_1^T N \mathbf{v}_3 = \frac{\mathbf{v}_1^T N \mathbf{w}_2}{h_{32}} = \frac{\mathbf{v}_1^T N (N \mathbf{v}_2 - h_{12} \mathbf{v}_1)}{h_{32}} = \frac{\mathbf{v}_1^T N^2 \mathbf{v}_2}{h_{32}} = \frac{\mathbf{v}_1^T N^3 \mathbf{v}_1}{h_{32} h_{21}} = 0$$

and the first inductive step holds. Assume now that for all  $i \leq j - 2$ ,  $h_{ij} = \mathbf{v}_i^T N \mathbf{v}_j = 0$ . Then (iv)  $\mathbf{w}_j = N \mathbf{v}_j - h_{j-1,j} \mathbf{v}_{j-1}$ . Hence, for all  $i \leq j - 2$ ,

$$0 = h_{i,i-1} \mathbf{v}_i^T N \mathbf{v}_j = \mathbf{v}_j^T N^T \mathbf{w}_{i-1} = \mathbf{v}_j^T N^T N \mathbf{v}_{i-1} = -\mathbf{v}_j^T N^2 \mathbf{v}_{i-1}$$

i.e., we have (v)  $\mathbf{v}_j^T N^2 \mathbf{v}_i = 0$  for all  $i \leq j - 2$ . We now prove that  $h_{i,j+1} = 0$  for all  $i \leq j - 1$ . We have using (iv)

$$h_{i,j+1} = \frac{\mathbf{v}_i^T N \mathbf{w}_j}{h_{j+1,j}} = \frac{\mathbf{v}_i^T N (N \mathbf{v}_j - h_{j-1,j} \mathbf{v}_{j-1})}{h_{j+1,j}} = \frac{\mathbf{v}_i^T N^2 \mathbf{v}_j - h_{j-1,j} \mathbf{v}_i^T N \mathbf{v}_{j-1}}{h_{j+1,j}}.$$

If  $i \leq j - 3$ , by the inductive hypothesis  $\mathbf{v}_i^T N \mathbf{v}_{j-1} = 0$  and by (v)  $\mathbf{v}_i^T N^2 \mathbf{v}_j = 0$  and hence  $h_{i,j+1} = 0$ . If  $i = j - 2$  then  $h_{j-2,j+1} = 0$  also since  $\mathbf{v}_{j-2}^T N^2 \mathbf{v}_j - h_{j-1,j} \mathbf{v}_{j-2}^T N \mathbf{v}_{j-1} = \mathbf{v}_{j-2}^T N^2 \mathbf{v}_j + h_{j-1,j} h_{j-1,j-2} = 0$  because

$$h_{j-1,j} = \mathbf{v}_{j-1}^T N \mathbf{v}_j = \frac{\mathbf{v}_j^T N^T \mathbf{w}_{j-2}}{h_{j-1,j-2}} = -\frac{\mathbf{v}_j^T N^2 \mathbf{v}_{j-2}}{h_{j-1,j-2}}.$$

Finally, if  $i = j - 1$ ,  $h_{j-1,j+1} = \mathbf{v}_{j-1}^T N^2 \mathbf{v}_j / h_{j+1,j} = 0$  since  $\mathbf{v}_{j-1}^T N^2 \mathbf{v}_j = 0$  for all  $j \geq 2$ . This we prove again by induction. Assuming  $\mathbf{v}_{j-1}^T N^2 \mathbf{v}_j = 0$ , we have using (iv), (iii)

$$\mathbf{v}_j N^2 \mathbf{v}_{j+1} = \frac{\mathbf{v}_j N^2 \mathbf{w}_j}{h_{j+1,j}} = \frac{\mathbf{v}_j N^2 (N \mathbf{v}_j - h_{j-1,j} \mathbf{v}_{j-1})}{h_{j+1,j}} = 0.$$

The result follows by noting that

$$\mathbf{v}_1 N^2 \mathbf{v}_2 = \mathbf{v}_1 N^2 \mathbf{w}_1 / h_{21} = \mathbf{v}_1 N^3 \mathbf{v}_1 / h_{21} = 0.$$

□

## References

1. Arioli, M.: A stopping criterion for the conjugate gradient algorithm in a finite element method framework. *Numer. Math.* **97**(1), 1–24 (2004)
2. Arioli, M., Noulard, E., Russo, A.: Stopping criteria for iterative methods: applications to PDEs. *Calcolo* **38**, 97–112 (2001)
3. Babuška, I.: Error bounds for finite element methods. *Numer. Math.* **16**, 322–333 (1971)
4. Brezzi, F., Bathe, K.J.: A discourse on the stability conditions for mixed finite element formulations. *Comp. Meth. Appl. Mech. Engrg.* **82**, 27–57 (1990)
5. Ciarlet, P.G.: *The finite element method for elliptic problems*. North Holland, Amsterdam, 1978
6. Concus, P., Golub, G.H.: A generalized conjugate gradient method for nonsymmetric systems of linear equations. In: R. Glowinski, J.L. Lions, (eds.), *Proc. Second Internat. Symp. on Computing Methods in Applied Sciences and Engineering*, volume **134** of *Lecture Notes in Economics and Mathematical Systems*, Berlin, 1976. Springer Verlag
7. Deuffhard, P.: Cascadic conjugate gradient methods for elliptic partial differential equations. Algorithm and numerical results. In: J. Xu, D. Keyes, (eds.), *Proceedings of the 7th International Conference on Domain Decomposition Methods*, AMS Providence, 1993, pp. 29–42
8. Deuffhard, P.: Cascadic conjugate gradient methods for elliptic partial differential equations: algorithm and numerical results. *Contemporary Mathematics* **180**, 29–42 (1994)
9. Deuffhard, P., Bornemann, F.A.: The cascadic multigrid method for elliptic problems. *Numerische Mathematik* **75**, 135–152 (1996)
10. Golub, G.H., Meurant, G.: Matrices, moments and quadrature II; how to compute the norm of the error in iterative methods. *BIT* **37**(3), 687–705 (1997)
11. Golub, G.H., Strakoš, Z.: Estimates in quadratic formulas. *Numer. Algorithms* **8**, 2–4 (1994)
12. Higham, N.J.: *Accuracy and stability of numerical algorithms*. SIAM, 2002
13. Horn, R.A., Johnson, C.R.: *Matrix analysis*. Cambridge University Press, 1985
14. Horn, R.A., Johnson, C.R.: *Topics in matrix analysis*. Cambridge University Press, 1991
15. Loghin, D., Wathen, A.J.: *Analysis of block preconditioners for saddle-point problems*. Technical Report 13, Oxford University Computing Laboratory, 2002, Submitted to SISC
16. Meurant, G.: Numerical experiments in computing bounds for the norm of the error in the preconditioned conjugate gradient algorithm. *Numer. Algorithms* **22**, 353–365 (1999)
17. Rigal, J.L., Gaches, J.: On the compatibility of a given solution with the data of a linear system. *J. Assoc. Comput. Mach.* **14**(3), 543–548 (1967)
18. Starke, G.: Field-of-values analysis of preconditioned iterative methods for non-symmetric elliptic problems. *Numer. Math.* **78**, 103–117 (1997)
19. Strakoš, Z., Tichý, P.: On error estimation by conjugate gradient method and why it works in finite precision computations. *Electronic Transactions on Numerical Analysis* **13**, 56–80 (2002)
20. Strang, W.G., Fix, G.J.: *An analysis of the finite element method*. Prentice-Hall, 1973
21. Widlund, O.: A Lanczos method for a class of nonsymmetric systems of linear equations. *SIAM J. Numer. Anal.* **15**(4), 1978