STOPPING CRITERIA FOR ITERATIONS IN FINITE-ELEMENT METHODS

Mario Arioli[1], Daniel Loghin[2], and Andy J. Wathen[3]

ABSTRACT

This work extends the results of Arioli (2002), Arioli, Noulard and Russo (2001) on stopping criteria for iterative solution methods for 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 provide alternative ways of calculating or estimating this quantity and present numerical experiments which validate our criteria.

Keywords: Conjugate gradient method, finite-element method.

AMS(MOS) subject classifications: 65F10, 65N30.

Current reports available by anonymous ftp to ftp.numerical.rl.ac.uk in directory pub/reports.

[1] M.Arioli@rl.ac.uk, Rutherford Appleton Laboratory,

[2] Daniel.Loghin@cerfacs.fr, CERFACS, 42 ave G. Coriolis, Toulouse, 31057, France

[3] Oxford University Computing Laboratory, Parks Road, Oxford, OX1 3QD, UK.

Computational Science and Engineering Department
Atlas Centre
Rutherford Appleton Laboratory
Oxon OX11 0QX

March 31st, 2003

Contents

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 (Golub and Meurant 1997), (Golub and Strakos 1994), (Meurant 1999), (Strakoš and Tichý 2002), only recently have there been attempts to relate convergence in the 'energy' norm to the finite element context (Arioli 2002), (Arioli et al. 2001), (Starke 1997). In particular, Arioli Arioli (2002) was the first to consider the original finite element setting to provide convergence criteria for the Conjugate Gradient method.

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

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

$$a(u, v) = f(v), \tag{1}$$

where $\mathcal{H}$ is a Hilbert space of functions u defined on a closed subset Ω 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 $u, v \in \mathcal{H}$

$$a(w,v) \leq C_1 \|w\|_{\mathcal{H}} \|v\|_{\mathcal{H}} \tag{2a}$$
$$a(v,v) \geq C_2 \|v\|_{\mathcal{H}}^2, \tag{2b}$$

with constants C_1, C_2 independent of discretization.

Condition (2b) is often used to replace the weaker (and sufficient) conditions of Babuška (1971)

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

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).

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

$$\text{Find} \quad u_h \in \mathcal{H}_h \quad \text{such that for all } v_h \in \mathcal{H}_h$$
$$a(u_h, v_h) = f(v_h). \tag{4}$$

Finite element methods choose $\mathcal{H}_h$ to be a space of functions v_h defined on a subdivision Ω_h of Ω into simplices T of diameter h_T; h denotes a piecewise constant function defined on Ω_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 $v_h \in \mathcal{H}_h$

$$a(u - u_h, v_h) = 0, \tag{5}$$

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

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

Remark 2.1 *Replacing v_h with the interpolant of u, $\mathcal{I}_h u$, 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

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

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

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

where

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

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

$$\mathcal{BE} \leq \frac{\|u - u_h\|_{\mathcal{H}}}{\|u_h\|_{\mathcal{H}}} =: \mathcal{FE} \leq \frac{C_1}{C_2}\mathcal{BE} \tag{8}$$

where

$$\mathcal{BE} := \frac{\|R(u_h)\|_{\mathcal{H}'}}{\|u_h\|_{\mathcal{H}}\|a\|_{\mathcal{H}\to\mathcal{H}'}} \tag{9}$$

Definition 2.1 *The quantities $\mathcal{FE}, \mathcal{BE}$ in (8),(9) are the functional forward and backward error respectively (Arioli et al. 2001).*

Remark 2.2 *The dual norm of the functional residual, $\|R(u_h)\|_{\mathcal{H}'}$, 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 $\|R(u_h)\|_{\mathcal{H}'}$ and thus $\mathcal{BE}$ are (polynomial) functions of the discretization parameter h and thus far from being close to machine precision. This we will use to advantage in the derivation of stopping criteria. However, we will not be concerned here with the derivation of any bounds but we will assume the following generic bound on the relative error*

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

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

We end this subsection with some standard notation. Expanding u_h in a basis of $\mathcal{H}_h$, we can derive a linear system of equations involving the coefficients $(\mathbf{u})_i, i = 1, \ldots, n$ of u_h in our choice of basis of $\mathcal{H}_h$

$$A\mathbf{u} = \mathbf{f} \tag{11}$$

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 Π_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

$$\Pi_h\mathbf{v} = \sum_{i=1} \mathbf{v}_i\phi_i,$$

where $\{\phi_i, i = 1, \ldots, n\}$ 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

$$\|v_h\|_{\mathcal{H}} = \|\mathbf{v}\|_H.$$

We will also use the matrix norm $\|\cdot\|_{H_1,H_2} : R^{n\times n}$ defined via

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

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. We also note here that

$$\|M\|_{H_1,H_2^{-1}} = \|H_2^{-1/2} M H_1^{-1}\|. \tag{12}$$

Finally, the stability conditions (2) become

$$\max_{\mathbf{w}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \max_{\mathbf{v}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \mathbf{w}^t A\mathbf{v} \leq C_1\|\mathbf{w}\|_H\|\mathbf{v}\|_H \tag{13a}$$

$$\min_{\mathbf{v}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \mathbf{v}^t A\mathbf{v} \geq C_2\|\mathbf{v}\|_H^2 \tag{13b}$$

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

$$\max_{\mathbf{v}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \mathbf{v}^t A\mathbf{v} \leq C_3\|\mathbf{v}\|_H^2. \tag{13c}$$

Remark 2.3 *In many situations of interest one can have $C_3 = C_2$. Moreover, if the symmetric part of A is H, then $C_2 = C_3 = 1$.*

We now state a result which can be found in Brezzi and Bathe (1990).

Theorem 2.2 *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*

$$\|M\|_{H,H^{-1}} = \max_{\mathbf{w}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \max_{\mathbf{v}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \frac{\mathbf{w}^t M\mathbf{v}}{\|\mathbf{w}\|_H\|\mathbf{v}\|_H},$$

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

The above result justifies the following definition.

Definition 2.3 *The H-condition number of a matrix M is*

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

Thus the stability conditions (2) simply say that the discrete problem (4) is well-conditioned with respect to the H-norm:

$$\|A\|_{H,H^{-1}} = C_1, \quad \|A^{-1}\|^{-1}_{H^{-1},H} \geq C_2, \tag{14}$$

and hence for all n

$$\kappa_H(A) \leq \frac{C_1}{C_2}. \tag{15}$$

Finally, we note that when solving $A\mathbf{u} = \mathbf{f}$ the discrete versions of (8), (9) are

$$FE := \frac{\|\mathbf{u} - \tilde{\mathbf{u}}\|_H}{\|\tilde{\mathbf{u}}\|_H}, \quad BE := \frac{\|\mathbf{f} - A\tilde{\mathbf{u}}\|_{H^{-1}}}{\|\tilde{\mathbf{u}}\|_H \|A\|_{H,H^{-1}}}. \tag{16}$$

3 Stopping criteria

In many large-scale computations the exact solution $\mathbf{u}$ of the linear system (11) 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 3.1 *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 3.2 *Let u be the solution of the weak formulation (1) and let $\mathbf{u}, u_h = \Pi_h \mathbf{u}$ satisfy*

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

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

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

for some $t \geq 0$.

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

$$\begin{aligned}
\frac{\|u_h - \tilde{u}_h\|_{\mathcal{H}}}{\|\tilde{u}_h\|_{\mathcal{H}}} &= \frac{\|A^{-1}\mathbf{r}\|_H}{\|\tilde{\mathbf{u}}\|_H} \\
&\leq \frac{\|A^{-1}\|_{H^{-1},H}\|\mathbf{r}\|_{H^{-1}}}{\|\tilde{\mathbf{u}}\|_H} \\
&\leq \frac{1}{C_2}\frac{\|\mathbf{r}\|_{H^{-1}}}{\|\tilde{\mathbf{u}}\|_H} \qquad \text{(using (14))}
\end{aligned}$$

we get

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

□

Remark 3.1 *The stopping criterion (17) is equivalent to requiring the discrete backward error* BE *defined in (16) 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 3.1,

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

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

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

The difference between the stopping criteria (17), (18) is not significant if the H-condition number (15) is not too large. This can be seen from the equivalence between $\|\cdot\|_{H^{-1}}, \|\cdot\|_{A^{-1}}$ provided by Lemma 3.1. 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* (Strang and Fix 1973), as it leads to a perturbed system

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

However, it is known that if the perturbation ΔA 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 $\mathbf{u}$ (Strang and Fix 1973). In this context, the proposed stopping criteria represent but another variational crime as the following standard result shows (see Arioli et al., 2001; see also Rigal and Gaches, 1967 for the case when l_p norms are employed.)

Theorem 3.3 *Let* $\tilde{\mathbf{u}}$ *satisfy*

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

Then there exists ΔA *such that*

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

and

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

Proof See (Arioli et al. 2001, Thm 1). □

Remark 3.2 *The more general case where the right-hand side* $\mathbf{f}$ *is perturbed is treated by Arioli et al. (2001). 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 Golub and Meurant, 1997, Golub and Strakos, 1994), the use of a nonsymmetric 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 basis $\mathcal{K}_k$. 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

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

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 contend ourselves with estimates provided by the iterative process. In the case of GMRES and FOM this can be achieved as follows. Assuming no breakdown, 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.

Lemma 4.1 *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 $\lambda_{min}(H) \leq C^* \lambda_{min}(H_k^s)$ for all $k > k^*$. Hence, our stopping criterion becomes

$$\|\mathbf{r}^k\| \leq C^* \lambda_{min}^{1/2}(H_k^s) h^t C(h). \tag{19}$$

We recall here that the value of $t \geq 0$ can be chosen according to application (a larger value, for a more pessimistic convergence criterion). 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 4.1 *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

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

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

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

A similar monotonicity result holds for the singular values of $\tilde{H}_k$ (cf. Horn and Johnson, 1991, 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

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

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

Remark 4.2 *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

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

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 ith 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 $\sigma_{min}(AP^{-1}) \sim \sigma_{min}(R_k)$

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

However, this requires the estimation of the smallest singular value of P which may not be easy to achieve. We address this issue in the 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 Glowinski and Lions (1976) and Widlund (1978) 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

$$\tilde{A}\tilde{\mathbf{u}} = \tilde{\mathbf{f}} \tag{23}$$

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

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

where

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

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

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

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 5.1 *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

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

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

$$V_k^t A V_k = H_k,$$

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

Proof See Appendix. □

This idea is contained in the work of Widlund (1978), although the author constructs a different tridiagonalization 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 5.1 *The action of the inverse of H as a preconditioner can be relaxed in practice. Indeed, solving to an accuracy of order $o(C(h))$ (say, $h^t C(h)$) 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 Γ. We will be using the following norms:

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

where

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

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

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

with norm

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

6.1 Elliptic problems

Consider the general second-order elliptic problem

$$\begin{aligned} -\nabla \cdot (\mathbf{a}(\mathbf{x})\nabla u) + \mathbf{b}(\mathbf{x}) \cdot \nabla u + c(\mathbf{x})u &= f \qquad \text{in } \Omega \subset \mathbb{R}^d \qquad (24a) \\ u &= 0 \qquad \text{on } \Gamma. \qquad (24b) \end{aligned}$$

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

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

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

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

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

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

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 Ω 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

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

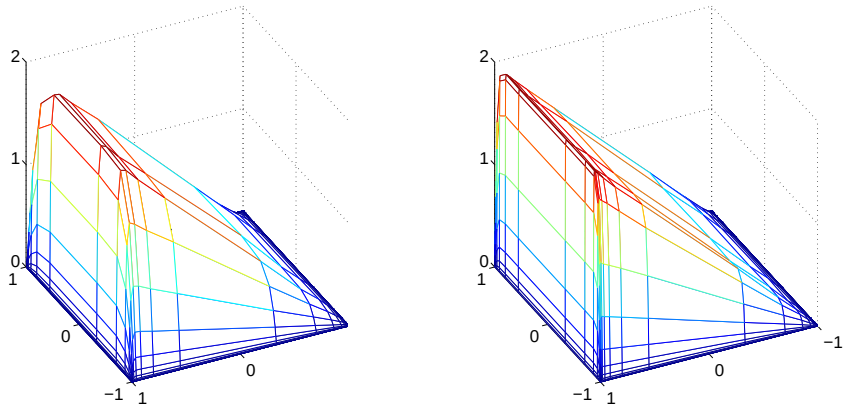
the constants C_2, C_3 in (13) are given as follows. If we choose to monitor the error with respect to $|\cdot|_1$ then

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

However, if we work with the energy norm defined by

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

then $C_2 = C_3 = 1$. We consider both cases below.



(a) $\nu = 1/10$ (b) $\nu = 1/100$

Figure 1: *Solution of advection-diffusion problem.*

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 ν toggles the degree of nonsymmetry of the matrices involved. The test problem is thus

$$\begin{aligned} -\nu\nabla^2 u + \mathbf{b}(x,y)\cdot\nabla u &= f \qquad \text{in } \Omega \equiv (-1,1)^2 && \text{(27a)} \\ u &= 0 \qquad \text{on } \Gamma, && \text{(27b)} \end{aligned}$$

with

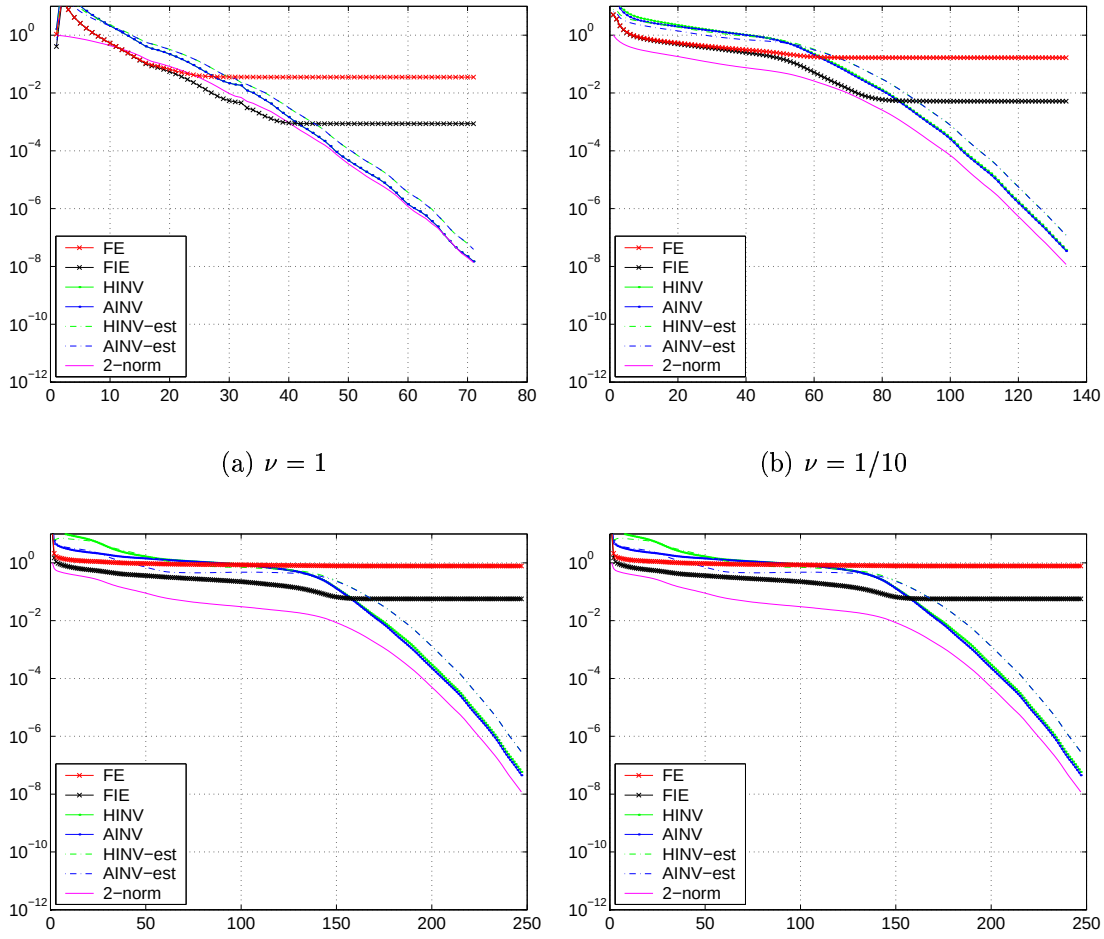
$$b(x,y) = \begin{pmatrix} 2y(1-x^2) \\ -2x(1-y^2) \end{pmatrix}$$

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

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

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

We first consider the errors with respect to the H_0^1-norm. 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 $|u - u_h^k|_1/|u_h^k|_1$;



(a) $\nu = 1$ (b) $\nu = 1/10$

(c) $\nu = 1/50$ (d) $\nu = 1/100$

Figure 2: *Comparison of stopping criteria for GMRES;* $h = 1/32$.

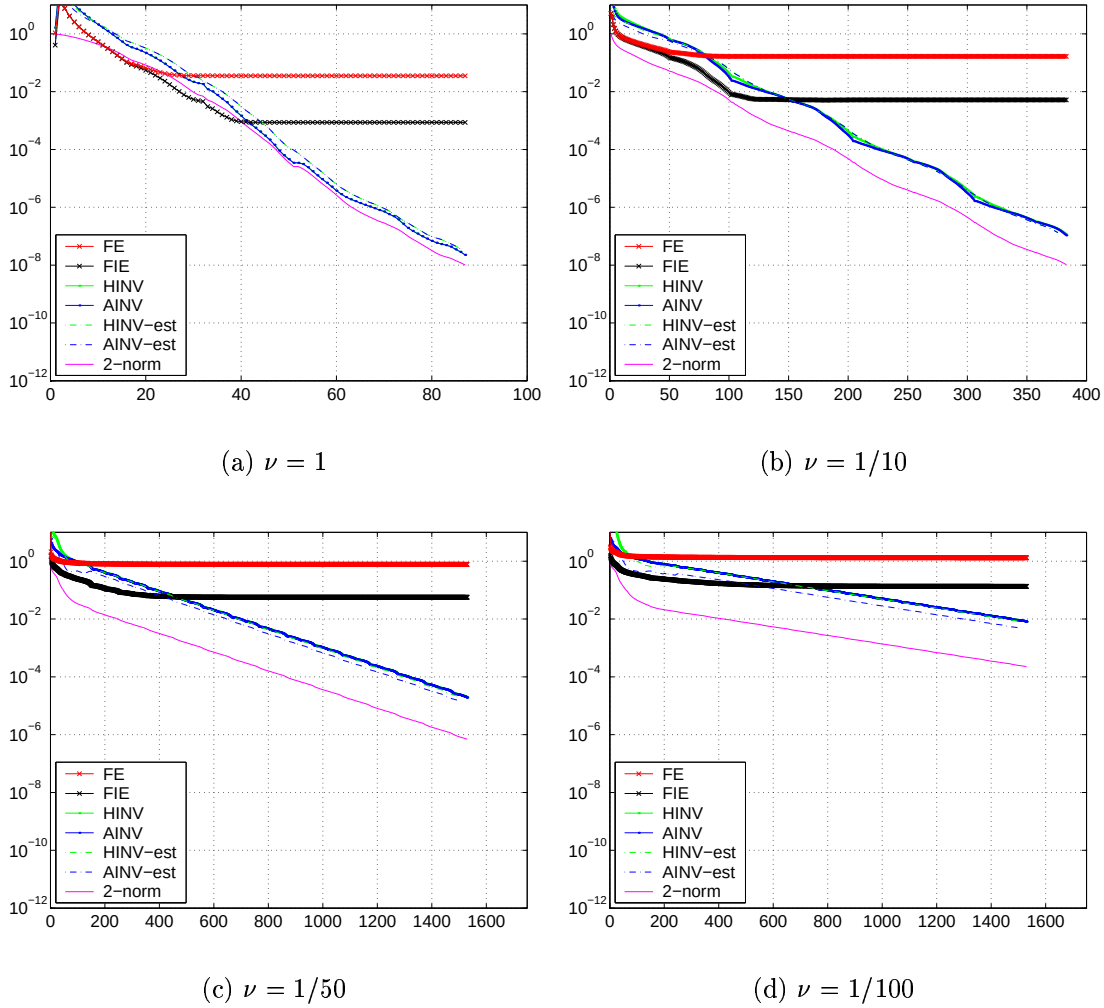
(ii) FIE: the exact relative (forward) interpolation errors $|u^I - u_h^k|_1/|u_h^k|_1$;
(iii) HINV: the exact H^{-1}-norm criterion (17) $h^{-t}C_2^{-1}\|\mathbf{r}^k\|_{H^{-1}}/\|\mathbf{u}^k\|_H$;
(iv) AINV: the exact A^{-1}-norm criterion (18) $h^{-t}C_2{}^{-1/2}\|\mathbf{r}^k\|_{A^{-1}}/\|\mathbf{u}^k\|_H$;
(v) HINV-est: the estimated H^{-1}-norm criterion (19) $h^{-t}C_2^{-1}\|\mathbf{r}^k\|\lambda_{min}^{-1/2}(H_k^s)/\|\mathbf{u}^k\|_H$;
(vi) AINV-est: the estimated A^{-1}-norm criterion (20) $h^{-t}C_2^{-1/2}\|\mathbf{r}^k\|\sigma_{min}^{-1/2}(H_k^s)/\|\mathbf{u}^k\|_H$;
(vii) the standard 2-norm stopping criterion $\|\mathbf{r}^k\|$.
We chose the exponent t to be in all cases $t = 1/2$ and the constants $c^* = C^* = 1$ in (19), (20).

6.3 GMRES without preconditioning

We begin with the case of a uniform partition of Ω into squares of size h and linear basis functions. The GMRES convergence curves are displayed in Fig. 2.

We see that if we are interested in satisfying a tolerance with respect to the H_0^1-norm, then in all cases we have to perform far fewer iterations than with a standard 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



(a) $\nu = 1$ (b) $\nu = 1/10$

(c) $\nu = 1/50$ (d) $\nu = 1/100$

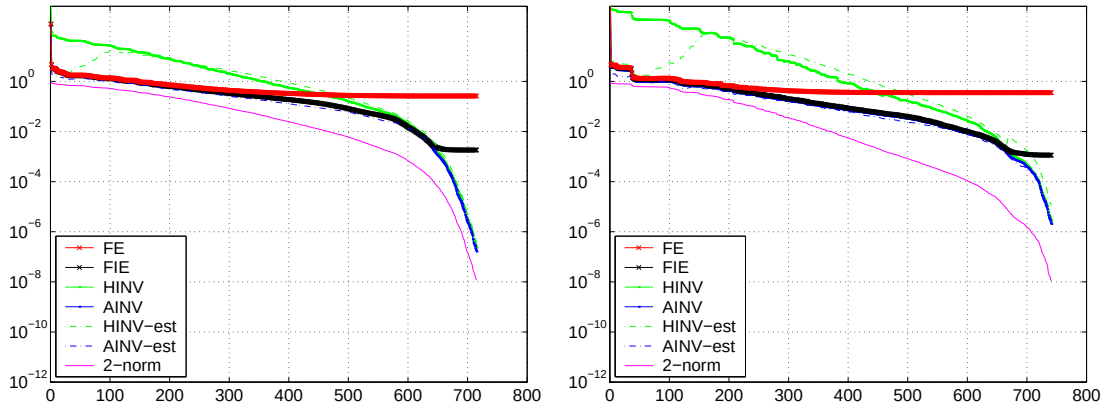
Figure 3: *Comparison of stopping criteria for GMRES(50);* $h = 1/32$.

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 ν. This robustness also holds with respect to the mesh parameter, although we do not include here experiments to demonstrate this fact.

The same convergence curves for the case of restarted GMRES are displayed in Fig. 3. They exhibit indeed the most dramatic difference between 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 3.1 deteriorates if the H-condition number of the problem deteriorates. For our test problem this happens when ν is small *and* the discretization is nonuniform. We consider this case below.

For small values of ν, 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 Grammian (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 norm 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.

Finally, the case when the H_0^1-norm is replaced with the energy norm (26) simplifies considerably for the advection-diffusion problem under consideration. Indeed, the energy norm is nothing but a scaling by $\sqrt{\nu}$ of the H_0^1-norm. Hence, all convergence curves



(a) $\nu = 10^{-4}$ (b) $\nu = 10^{-6}$

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

displayed in the previous figures are scaled accordingly except for the relative Euclidean norm which stays the same. Thus, by varying ν, we can obtain GMRES convergence curves based on the Euclidean residual which can be either above or below the relevant convergence curve for the finite element error. For this reason we do not display all the performance again, but emphasise the fact that the standard stopping criterion used in iterative solvers has no relation to the actual convergence of quantities of interest in the original problem.

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.

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

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

Since

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

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

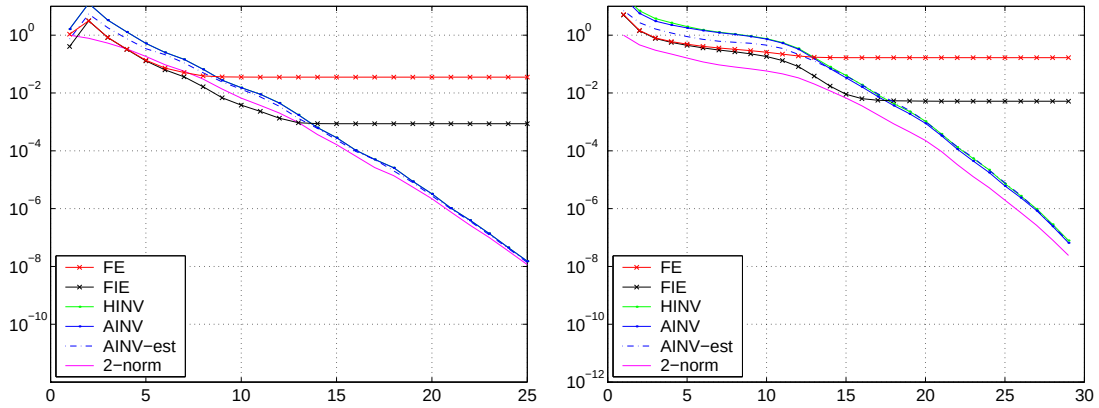
$$\sigma_{min}(P) \sim \left(\max_k \frac{\|\mathbf{z}^k\|}{\|\mathbf{v}^k\|} \right)^{-1} = \min_k \frac{\|\mathbf{v}^k\|}{\|\mathbf{z}^k\|}$$

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 $\|\mathbf{v}^k\| = 1$.

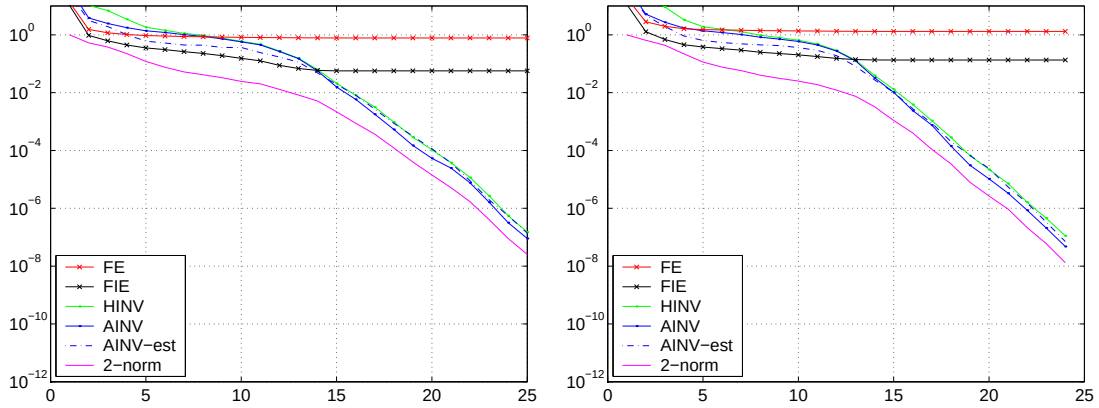
The performance of GMRES with ILU preconditioning is displayed in Fig. 5. While the number of iterations is greatly reduced, the convergence behaviour is similar to the unpreconditioned case. Moreover, the approximation of the residual A^{-1}-norm described above appears to work extremely well. However, 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, $\|\mathbf{r}^k\|_{H^{-1}}$.



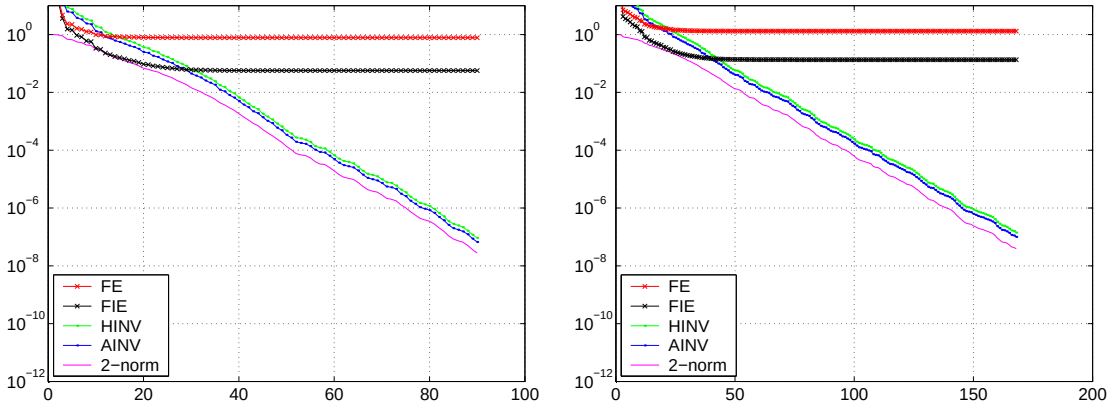
(a) $\nu = 1$ (b) $\nu = 1/10$



(c) $\nu = 1/50$ (d) $\nu = 1/100$

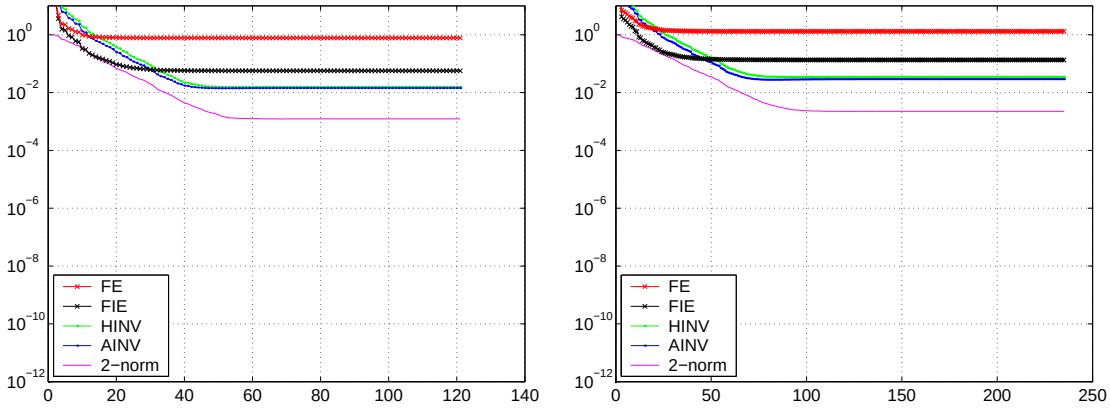
Figure 5: *Comparison of stopping criteria for GMRES with ILU(0) preconditioning;* $h = 1/32$.

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 (Loghin and Wathen 2002). 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 by Arioli (2002); 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 inexact



(a) $\nu = 1/50$ (b) $\nu = 1/100$

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



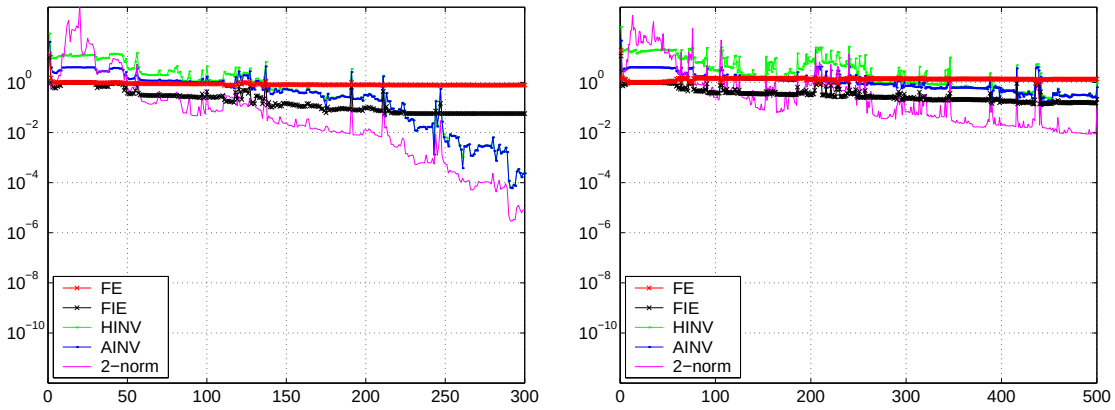
(a) $\nu = 1/50$ (b) $\nu = 1/100$

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

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

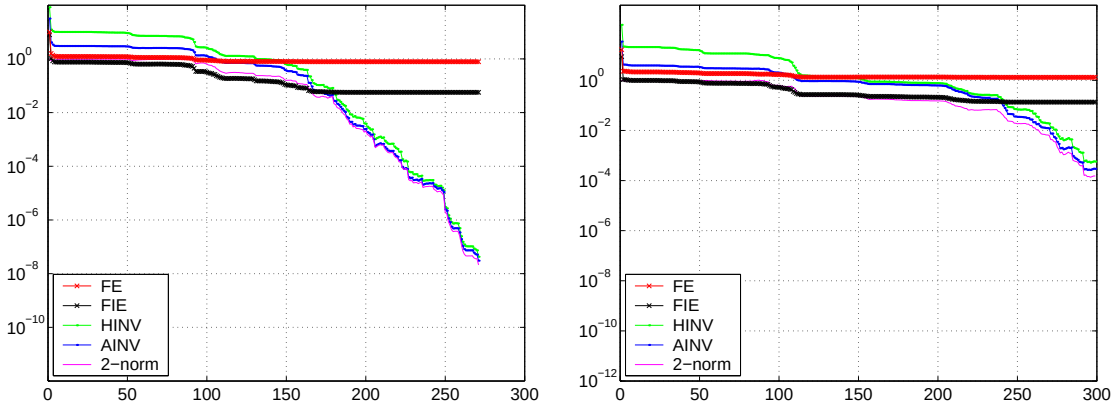
6.6 Other iterative methods

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.



(a) $\nu = 1/50$ (b) $\nu = 1/100$

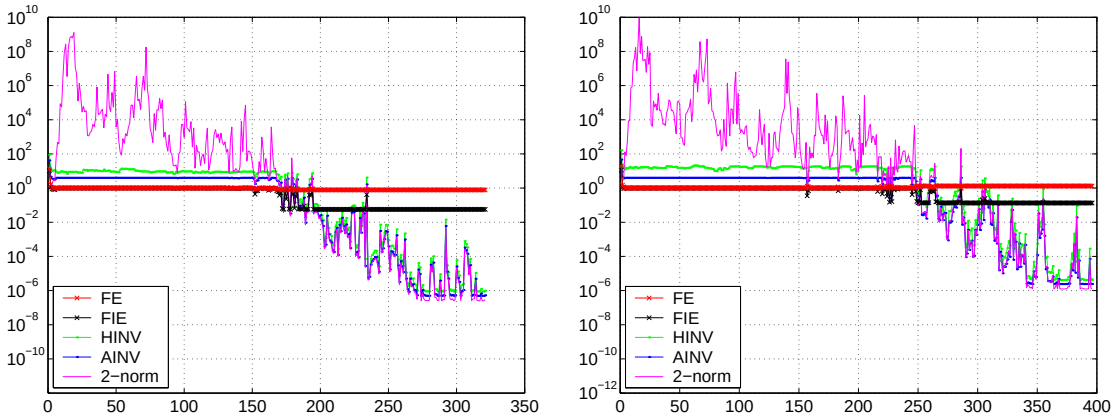
Figure 8: *Comparison of stopping criteria for BICGSTAB: h=1/32.*



(a) $\nu = 1/50$ (b) $\nu = 1/100$

Figure 9: *Comparison of stopping criteria for QMR: h=1/32.*

In particular, the A^{-1}-norm provides again the closest approximation to the these quantities, with the H^{-1}-norm bound possibly deteriorating for more nonsymmetric problems, 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.



(a) $\nu = 1/50$ (b) $\nu = 1/100$

Figure 10: *Comparison of stopping criteria for CGS: h=1/32.*

7 Conclusion

The message of this paper is simple: do not accurately compute the solution of an inaccurate problem. This was highlighted already by Arioli (2002) 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. 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 3.1.
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

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

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

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

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 (Horn and Johnson, 1985, p. 11). Hence,

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

and since $\|H^{-1/2}AH^{-1/2}\| = \|A\|_{H,H^{-1}} = C_1$ (cf. (12), (14)) we get

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

and the result follows. □

Proof of Lemma 5.1. Consider the two equivalent linear systems

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

where

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

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

Arnoldi in $(\cdot,\cdot)$
$\tilde{\mathbf{v}}_1 := \tilde{\mathbf{r}}_0/\|\tilde{\mathbf{r}}_0\|$
do $j = 1, 2, \ldots, m$
$\tilde{h}_{ij} = (\tilde{A}\tilde{\mathbf{v}}_j, \tilde{\mathbf{v}}_i)$, $1 \le i \le j$
$\tilde{\mathbf{w}}_j = \tilde{A}\tilde{\mathbf{v}}_j - \sum_i^j \tilde{h}_{ij}\tilde{\mathbf{v}}_i$
if $\tilde{h}_{j+1,j} = \|\tilde{\mathbf{w}}_j\| = 0$ **stop**
$\tilde{\mathbf{v}}_{j+1} = \tilde{\mathbf{w}}_j/\tilde{h}_{j+1,j}$
end do

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

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 $\tilde{\ }$ s. We now prove by induction on j that for all $i \le j-2$

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

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 $i \le j-2$, $h_{ij} = \mathbf{v}_i^t N\mathbf{v}_j = 0$. Then (iv) $\mathbf{w}_j = N\mathbf{v}_j - h_{j-1,j}\mathbf{v}_{j-1}$. Hence, for all $i \le j-2$,

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

i.e., we have (v) $\mathbf{v}_j^t N^2\mathbf{v}_i = 0$ for all $i \le j-2$. We now prove that $h_{i,j+1} = 0$ for all $i \le j-1$. 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 $i \le j-3$, 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 $i = j-1$, $h_{j-1,j+1} = \mathbf{v}_{j-1}N^2\mathbf{v}_j/h_{j+1,j} = 0$ since $\mathbf{v}_{j-1}N^2\mathbf{v}_j = 0$ for all $j \ge 2$. This use prove again by induction. Assuming $\mathbf{v}_{j-1}N^2\mathbf{v}_j = 0$, 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

Arioli, M. (2002), A stopping criterion for the conjugate gradient algorithm in a finite element method framework, Technical Report RAL-TR-2002-034, RAL.

Arioli, M., Noulard, E. and Russo, A. (2001), 'Stopping criteria for iterative methods: applications to PDEs', *Calcolo* **38**, 97–112.

Babuska, I. (1971), 'Error-bounds for finite element method', *Numer. Math.* **16**, 322–333.

Brezzi, F. and Bathe, K. J. (1990), 'A discourse on the stability conditions for mixed finite element formulations', *Comp. Meth. Appl. Mech. Engrg.* **82**, 27–57.

Glowinski, R. and Lions, J. L., eds (1976), *A generalized conjugate gradient method for nonsymmetric systems of linear equations*, Vol. 134 of *Lecture Notes in Economics and Mathematical Systems*, Springer Verlag, Berlin.

Golub, G. and Meurant, G. (1997), 'Matrices, moments and quadrature II; how to compute the norm of the error in iterative methods', *BIT* **37**, 687–705.

Golub, G. and Strakos, Z. (1994), 'Estimates in quadratic formulas', *Numer. Algorithms* **8**, 241–268.

Horn, R. A. and Johnson, C. R. (1985), *Matrix Analysis*, Cambridge University Press.

Horn, R. A. and Johnson, C. R. (1991), *Topics in Matrix Analysis*, Cambridge University Press.

Loghin, D. and Wathen, A. J. (2002), Analysis of block preconditioners for saddle-point problems, Technical Report Technical Report 13, Oxford University Computing Laboratory.

Meurant, G. (1999), 'Numerical experiments in computing bounds for the norm of the error in the preconditioned conjugate gradient algorithm', *Numerical Algorithms* **22**, 353–365.

Rigal, J. L. and Gaches, J. (1967), 'On the compatibility of a given solution with the data of a linear system', *J. Assoc. Comput. Mach.* **14**(3), 543–548.

Starke, G. (1997), 'Field-of-values analysis of preconditioned iterative methods for nonsymmetric elliptic problems', *Numer. Math.* **78**, 103–117.

Strakoš, Z. and Tichý, P. (2002), 'On error estimation by conjugate gradient method and why it works in finite precision computations', *Electronic Transactions on Numerical Analysis* **13**, 56–80.

Strang, W. G. and Fix, G. J. (1973), *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ.

Widlund, O. (1978), 'A Lanczos method for a class of nonsymmetric systems of equations', *SIAM J. Numer. Anal.* **15**(4), 801–812.