ESTIMATES OF THE NORM OF THE ERROR IN SOLVING LINEAR SYSTEMS WITH FOM AND GMRES

GÉRARD MEURANT*

Abstract. We provide formulas for the norm of the error when solving non symmetric linear systems with the full orthogonalization method (FOM) and the generalized minimum residual method (GMRES) as well as relations between the error norm and the residual norm. From these formulas we are able to compute estimates of the norm of the error during the iterations. Since stopping criteria based on the norm of the residual may sometimes be misleading, such estimates could lead to a more robust way to stop the iterations. Numerical experiments show that the proposed norm estimates work nicely on difficult linear systems.

1. Introduction.

We consider solving a linear system

$$Ax = b$$

where A is a non singular real matrix of order n with the full orthogonalization method (FOM) and the generalized minimum residual method (GMRES) which are Krylov methods based on the Arnoldi orthogonalization process; see Saad [24], [25] and Saad and Schultz [26]. The initial residual is denoted as $r^0 = b - Ax^0$ where x^0 is the starting vector. The Krylov subspace of order k based on A and r^0 which is denoted as $\mathcal{K}_k(r^0, A)$ is $span\{r^0, Ar^0, \ldots, A^{k-1}r^0\}$. The approximate solution x^k at iteration k is sought as $x^k \in x^0 + \mathcal{K}_k(r^0, A)$ such that the residual vector $r^k = b - Ax^k$ satisfies an orthogonality condition with the Krylov subspace. The orthogonality condition is different in FOM and GMRES.

There are many papers and books which give expressions or bounds for the norms of the residual r^k; see for instance [26], [6], [8], [23], [14], [28]. In most implementations of FOM or GMRES the iterations are stopped using the l_2 norm of the residual. We will see that for some numerical examples this can be misleading. The iterations may be stopped too soon or too late giving an approximate solution far from the exact one or being more costly to compute than needed. Another stopping criterion is based on the backward error; see [1]. Curiously enough there are not many papers considering the error norm with the exception of [17]. The fact that the residual norm can be misleading for stopping the iterations was already pointed out when A is symmetric by Hestenes and Stiefel in their seminal paper on the conjugate gradient method [15], p. 416.

The aim of the present paper is to derive formulas for the l_2 norm of the error vectors $\epsilon^k = x - x^k$ in FOM and GMRES. Of course, these expressions involve some terms which are not directly computable at iteration k. Nevertheless we will use these exact formulas to compute estimates of the norms of the errors during the FOM or GMRES iterations, a few iterations before the current one. This program has already been achieved for the conjugate gradient (CG) algorithm and the A-norm of the error when the matrix A is symmetric and positive definite in [12], [18] and [19] and for the l_2 norm of the error in [20]; for a summary see also [21] and [11]. In [27] it was shown that these techniques work also for CG in finite precision arithmetic. This is an important point for their reliable use in practical computations. Even though we do not consider rounding errors, the present paper can be seen as an extension of these ideas to the nonsymmetric case since when A is symmetric FOM reduces to CG and GMRES to MINRES.

*(gerard.meurant@gmail.com) revised version, Dec 2010

The contents of the paper are as follows. Section 2 recalls the definitions of FOM and GMRES proposed by Saad [25] and Saad and Schultz [26]. Section 3 gives exact expressions for the l_2 norm of the error in FOM. These formulas use entries of the inverses of the Hessenberg matrices which are progressively constructed in the Arnoldi process [2]. Throughout the paper we will assume that all these square Hessenberg matrices are non singular. This corresponds to the fact that all the FOM iterates exist and GMRES does not stagnate. In section 4 we give alternate expressions for the norms of the errors to simplify them and above all to obtain expressions which are more amenable to computations in finite precision arithmetic. In section 5 we also relate the formulas for the norm of the error to the norm of the residual. Section 6 uses the formulas of section 4 to estimate the error norm during the FOM iterates and gives numerical examples which show that estimating the norm of the error can give a more reliable stopping criterion than using the residual norm. Relations between the norms of the residuals in FOM and GMRES have been studied previously in [6] and [8]; see also [9], [10]. Section 7 studies the relations between the error norms in FOM and GMRES. From these results section 8 provides formulas for the error norm in GMRES and relations between the error and residual norms. From these formulas we show in section 9 how to compute estimates of the error norm in GMRES. Numerical experiments on the same examples as in section 6 show that the proposed approach can give good approximations of the error norm. We compare our estimates with the backward error [1] and also with an error estimate proposed by Brezinski; see [4] and [5]. Finally section 10 provides some conclusions and perspectives.

Throughout this paper e^k denotes the kth column of the identity matrix of different orders.

2. FOM and GMRES. Let V_k be a matrix whose columns are orthonormal basis vectors $v^j, j = 1, \ldots, k$ of the Krylov subspace $\mathcal{K}_k(r^0, A)$. The iterates of FOM or GMRES are defined as

$$x^k = x^0 + V_k z^k.$$

The basis vectors v^j are usually computed recursively using the modified Gram–Schmidt (MGS) algorithm in the Arnoldi process [2]. The matrix relation for the matrix V_k is the following:

$$AV_k = V_k H_k + h_{k+1,k} v^{k+1} (e^k)^T, \tag{2.1}$$

where H_k is an upper Hessenberg matrix of order k with elements $h_{i,j}$. Therefore,

$$H_k = \begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & \cdots & h_{1,k} \\ h_{2,1} & h_{2,2} & & & \vdots \\ & h_{3,2} & \ddots & & \vdots \\ & & \ddots & \ddots & \vdots \\ & & & h_{k,k-1} & h_{k,k} \end{pmatrix}. \tag{2.2}$$

We also have $H_k = V_k^T A V_k$ and $AV_n = V_n H_n$ if we assume that the Arnoldi process does not terminate early; that is, if $h_{k+1,k} \neq 0$ for $k = 1, \ldots, n-1$.

In FOM, which is an orthogonal residual (OR) method, we ask for the residual $r^k = b - Ax^k$ to be orthogonal to the Krylov subspace. It gives

$$(r^k)^T V_k = 0.$$

The vector of coordinates z^k is obtained by writing

$$V_k^T r^k = V_k^T(b - Ax^k) = V_k^T(b - Ax^0 - AV_k z^k) = V_k^T r^0 - V_k^T A V_k z^k = 0.$$

This gives

$$H_k z^k = V_k^T r^0 = \|r^0\| V_k^T v^1 = \|r^0\| e^1, \tag{2.3}$$

since in the Arnoldi process $v^1 = r^0/\|r^0\|$. The FOM kth iterate x^k exist only if H_k is non singular. The linear system (2.3) is solved using an LU or preferably a QR factorization of the Hessenberg matrix H_k.

In GMRES the l_2 norm of the residual is minimized at each iteration. This gives the orthogonality condition $(r^k)^T A V_k = 0$. The matrix relation of equation (2.1) can also be written as

$$AV_k = V_{k+1} H_k^{(e)}, \tag{2.4}$$

where $H_k^{(e)}$ is the $(k+1) \times k$ extended matrix

$$H_k^{(e)} = \begin{pmatrix} H_k \\ h_{k+1,k}(e^k)^T \end{pmatrix}.$$

The norm of the residual can be written as

$$\begin{aligned} \|b - Ax^k\| &= \|r^0 - AV_k z^k\|, \\ &= \| \, \|r^0\| V_{k+1} e^1 - V_{k+1} H_k^{(e)} z^k\|, \\ &= \| \, \|r^0\| e^1 - H_k^{(e)} z^k\|. \end{aligned}$$

The coordinates z^k are computed by solving the least squares problem

$$\min_z \| \, \|r^0\| e^1 - H_k^{(e)} z\|. \tag{2.5}$$

The theoretical solution is given by the pseudo inverse of $H_k^{(e)}$,

$$z^k = \|r^0\| \, ([H_k^{(e)}]^T H_k^{(e)})^{-1} [H_k^{(e)}]^T e^1.$$

In practical computations the solution of the least squares problem is obtained by using a QR factorization of $H_k^{(e)}$ using rotations. Contrary to FOM, GMRES cannot break down as long as $h_{k+1,k} \neq 0$ since the least squares problem can always be solved. If there exists an index m for which $h_{m+1,m} = 0$ we have found the solution of the linear system. Note that the backward stability of MGS-GMRES has been proved in [22].

3. Formulas for the error norm in FOM. Our first goal in this paper is to give formulas for the l_2 norm of the error $\epsilon^k = x - x^k$. It is well known that the relation between the error and the residual is $A\epsilon^k = r^k$.

As stated before we assume that the matrices H_k, $1 \leq k \leq n$ are nonsingular and therefore all the FOM iterates exist. Whatever is the algorithm, FOM or GMRES, $r^k = r^0 - AV_k z^k$ and we have

$$\|\epsilon^k\|^2 = (A^{-1}r^0, A^{-1}r^0) - 2(A^{-1}r^0, V_k z^k) + (V_k z^k, V_k z^k). \tag{3.1}$$

When we consider FOM we have the following result.

THEOREM 3.1. *In the FOM method the square of the l_2 norm of the error $\|\epsilon^k\|^2$ is given by*

$$\|r^0\|^2[(H_n^{-1}e^1, H_n^{-1}e^1) - (H_k^{-1}e^1, H_k^{-1}e^1) + 2h_{k+1,k}(H_k^{-1}e^1, e^k)((H_n^{-1}e^{k+1})^k, H_k^{-1}e^1)],$$

where $(H_n^{-1}e^{k+1})^k$ denotes the k first components of the $k+1$ st column of the inverse of H_n.

Proof. In FOM the vector of coefficients z^k in the orthogonal basis is given by $H_k z^k = \|r^0\|e^1$ and then the kth iterate is $x^k = x^0 + V_k z^k$. The first term on the right hand side of equation (3.1) is $(A^{-1}r^0, A^{-1}r^0)$. We write $r^0 = \|r^0\|v^1 = \|r^0\|V_n e^1$. Since with our hypothesis $AV_n = V_n H_n$ and $H_n = V_n^T A V_n$ is assumed to be non singular, we have $V_n H_n^{-1} = A^{-1}V_n$. Therefore,

$$(A^{-1}r^0, A^{-1}r^0) = \|r^0\|^2(H_n^{-1}e^1, H_n^{-1}e^1).$$

By orthogonality of the matrices V_k the third term in equation (3.1) is (z^k, z^k). Hence,

$$(V_k z^k, V_k z^k) = \|r^0\|^2(H_k^{-1}e^1, H_k^{-1}e^1).$$

It is more difficult to deal with the middle term $(A^{-1}r^0, V_k z^k)$. It seems that the easiest way is the following. We write

$$(A^{-1}r^0, V_k z^k) = \|r^0\|(e^1, V_k^T A^{-T} V_k z^k),$$

and we remark that from equation (2.1)

$$V_k^T A^{-T} V_k = H_k^{-T} - h_{k+1,k} H_k^{-T} e^k (v^{k+1})^T A^{-T} V_k.$$

Therefore,

$$(A^{-1}r^0, V_k z^k) = \|r^0\|[(H_k^{-1}e^1, z^k) - h_{k+1,k}(H_k^{-1}e^1, e^k)(A^{-1}v^{k+1}, V_k z^k)].$$

We have

$$(A^{-1}v^{k+1}, V_k z^k) = \|r^0\|(A^{-1}V_n e^{k+1}, V_k H_k^{-1}e^1) = \|r^0\|(V_n H_n^{-1}e^{k+1}, V_k H_k^{-1}e^1).$$

But,

$$(V_n H_n^{-1}e^{k+1}, V_k H_k^{-1}e^1) = ((H_n^{-1}e^{k+1})^k, H_k^{-1}e^1),$$

where $(H_n^{-1}e^{k+1})^k$ denotes the vector of the k first components of $H_n^{-1}e^{k+1}$. Hence,

$$(A^{-1}r^0, V_k z^k) = \|r^0\|^2[(H_k^{-1}e^1, H_k^{-1}e^1) - h_{k+1,k}(H_k^{-1}e^1, e^k)((H_n^{-1}e^{k+1})^k, H_k^{-1}e^1)].$$

We have a factor -2 in front of this term. The first term on the right hand side can be regrouped with (z^k, z^k) to obtain the result. □

Note that if the Arnoldi process stops at iteration m with $h_{m+1,m} = 0$ we can replace n by m in the previous theorem. We also remark that at iteration k we do not know H_n, so the norm of the error cannot be directly computed by the formula of Theorem 3.1. This formula is somehow similar to what was obtained for CG in the symmetric case in [20], Theorem 2.1.

Finally to close this section we remark that if the matrix A has a positive definite symmetric part, then $(A\epsilon^k, \epsilon^k)$ defines the square of a norm of the error. It turns out that, as it is the case for CG (see [21]), one can also obtain formulas for this norm. They are in fact simpler than the formulas for the l_2 norm of the error but we cannot give them here due to the lack of space.

4. Simplifications of the formula for the error norms in FOM. We do not know the signs of all the terms in the formula of Theorem 3.1 and it has been shown for the conjugate gradient case when A is symmetric and positive definite that it may not be appropriate to compute differences similar to $(H_n^{-1}e^1, H_n^{-1}e^1) - (H_k^{-1}e^1, H_k^{-1}e^1)$ in finite precision arithmetic; see [12], [18], [27] and [21] for a summary. It is thus interesting to try to express the first term of the right hand side of the formula of Theorem 3.1 as a function of the second one to avoid computing the difference. To this end we write H_n block-wise as

$$H_n = \begin{pmatrix} H_k & W_k \\ Y_k^T & \tilde{H}_k \end{pmatrix}, \tag{4.1}$$

where $k < n$. Since H_n is upper Hessenberg and H_k is square we note that Y_k^T has just one nonzero element $h_{k+1,k}$ in the top right corner. Thus $Y_k^T = h_{k+1,k}e^1(e^k)^T$. The matrices H_k of order k and $\tilde{H}_k$ of order $n-k$ are square upper Hessenberg and W_k is generally a full rectangular matrix. Let $S_k = H_k - W_k\tilde{H}_k^{-1}Y_k^T$ be the Schur complement; the inverse of H_n can be written as

$$H_n^{-1} = \begin{pmatrix} S_k^{-1} & -S_k^{-1}W_k\tilde{H}_k^{-1} \\ -\tilde{H}_k^{-1}Y_k^T S_k^{-1} & \tilde{H}_k^{-1} + \tilde{H}_k^{-1}Y_k^T S_k^{-1}W_k\tilde{H}_k^{-1} \end{pmatrix}.$$

Because of the special structure of Y_k^T we have

$$S_k = H_k - h_{k+1,k}(W_k\tilde{H}_k^{-1}e^1)(e^k)^T,$$

so, S_k is a rank-one modification of H_k. By using the Sherman-Morrison formula (see [13]) we obtain

$$S_k^{-1} = H_k^{-1} + \frac{h_{k+1,k}}{1 - h_{k+1,k}(e^k, H_k^{-1}W_k\tilde{H}_k^{-1}e^1)}H_k^{-1}W_k\tilde{H}_k^{-1}e^1(e^k)^T H_k^{-1}.$$

Note that if $h_{k+1,k}(e^k, H_k^{-1}W_k\tilde{H}_k^{-1}e^1) = 1$ then S_k is singular. Since we have assumed that H_k is non singular, this would imply that H_n is singular contrary to our hypothesis. The difference $(H_n^{-1}e^1, H_n^{-1}e^1) - (H_k^{-1}e^1, H_k^{-1}e^1)$ can now be expressed in the following way.

LEMMA 4.1.
Let $w^k = W_k\tilde{H}_k^{-1}e^1$ *and*

$$\gamma_k = \frac{h_{k+1,k}(e^k, H_k^{-1}e^1)}{1 - h_{k+1,k}(e^k, H_k^{-1}w^k)}.$$

Then,

$$\begin{aligned}(H_n^{-1}e^1, H_n^{-1}e^1) - (H_k^{-1}e^1, H_k^{-1}e^1) &= (h_{k+1,k}(e^k, S_k^{-1}e^1))^2(\tilde{H}_k^{-1}e^1, \tilde{H}_k^{-1}e^1) \\ &+ 2\gamma_k(H_k^{-1}e^1, H_k^{-1}w^k) + \gamma_k^2(H_k^{-1}w^k, H_k^{-1}w^k).\end{aligned}$$

Proof. We are interested in the first column of H_n^{-1} and, in particular, in $S_k^{-1}e^1$ for which we have

$$S_k^{-1}e_1 = H_k^{-1}e^1 + \gamma_k H_k^{-1}W_k\tilde{H}_k^{-1}e^1 = H_k^{-1}e^1 + \gamma_k H_k^{-1}w^k.$$

The remaining part of the first column of H_n^{-1} which is $-\tilde{H}_k^{-1}Y_k^T S_k^{-1}e^1$ can be written as

$$-h_{k+1,k}(e^k, S_k^{-1}e^1)\tilde{H}_k^{-1}e^1.$$

Hence,

$$(H_n^{-1}e^1, H_n^{-1}e^1) = (S_k^{-1}e^1, S_k^{-1}e^1) + (h_{k+1,k}(e^k, S_k^{-1}e^1))^2(\tilde{H}_k^{-1}e^1, \tilde{H}_k^{-1}e^1),$$

and

$$(S_k^{-1}e^1, S_k^{-1}e^1) = (H_k^{-1}e^1, H_k^{-1}e^1) + 2\gamma_k(H_k^{-1}e^1, H_k^{-1}w^k) + \gamma_k^2(H_k^{-1}w^k, H_k^{-1}w^k),$$

with $w^k = W_k\tilde{H}_k^{-1}e^1$. □

Then, for the other part of the formula of Theorem 3.1, we are interested in computing the k first elements of $H_n^{-1}e^{k+1}$ (which are denoted as $(H_n^{-1}e^{k+1})^k$) where e^{k+1} is here the $k+1$st column of the identity matrix of order n.

LEMMA 4.2. *Using the notations of Lemma 4.1, we have*

$$(H_n^{-1}e^{k+1})^k = -\frac{\gamma_k}{h_{k+1,k}(e^k, H_k^{-1}e^1)}H_k^{-1}w^k.$$

Proof. By construction the vector $(H_n^{-1}e^{k+1})^k$ is the first column of the top right block of the inverse that is, $-S_k^{-1}W_k\tilde{H}_k^{-1}e^1$. Using the results for S_k^{-1}, we obtain

$$\begin{aligned}-S_k^{-1} \quad & W_k\tilde{H}_k^{-1}e^1 = \\ & -\left[H_k^{-1} + \frac{h_{k+1,k}}{1 - h_{k+1,k}(e^k, H_k^{-1}W_k\tilde{H}_k^{-1}e^1)}H_k^{-1}W_k\tilde{H}_k^{-1}e^1(e^k)^T H_k^{-1}\right] W_k\tilde{H}_k^{-1}e^1.\end{aligned}$$

Hence, since $(e^k)^T H_k^{-1}W_k\tilde{H}_k^{-1}e^1$ is a scalar, this is

$$-\left[1 + \frac{h_{k+1,k}(e^k, H_k^{-1}W_k\tilde{H}_k^{-1}e^1)}{1 - h_{k+1,k}(e^k, H_k^{-1}W_k\tilde{H}_k^{-1}e^1)}\right] H_k^{-1}W_k\tilde{H}_k^{-1}e^1 =$$

$$-\frac{1}{1 - h_{k+1,k}(e^k, H_k^{-1}w^k)}H_k^{-1}w^k.$$

This last expression can be rewritten using γ_k defined in Lemma 4.1. □

Finally we have the following expression for the square of the norm of the error in FOM.

THEOREM 4.3. *Assume the block partitioning of H_n as in equation (4.1) and denote $w^k = W_k\tilde{H}_k^{-1}e^1$ and*

$$\gamma_k = \frac{h_{k+1,k}(e^k, H_k^{-1}e^1)}{1 - h_{k+1,k}(e^k, H_k^{-1}w^k)}.$$

Then

$$\|\epsilon^k\|^2/\|r^0\|^2 = \{h_{k+1,k}\,[\,(e^k, H_k^{-1}e^1) + \gamma_k(e^k, H_k^{-1}w^k)]\,\}^2\|\tilde{H}_k^{-1}e^1\|^2 + \gamma_k^2\|H_k^{-1}w^k\|^2.$$

Proof. We use the results of Lemmas 4.1 and 4.2. The third term in the right hand side of the formula of Theorem 3.1 is given by

$$2h_{k+1,k}(H_k^{-1}e^1, e^k)((H_n^{-1}e^{k+1})^k, H_k^{-1}e^1) = -2\gamma_k(H_k^{-1}w^k, H_k^{-1}e^1).$$

We see that this term cancels with the same one but of opposite sign in the formula of Lemma 4.1. ∎

Note that for $\|\epsilon^k\|^2$ we now have the sum of two positive quantities. The quantities which are involved are $\tilde{H}_k^{-1}e^1$, w^k, $H_k^{-1}w^k$ and $H_k^{-1}e^1$. Of course, at FOM iteration k we do not know $\tilde{H}_k$ and w^k yet. Expressing everything in terms of γ_k the norm of the error squared can be written even more simply as

$$\|\epsilon^k\|^2/\|r^0\|^2 = \gamma_k^2 \left\{ \|\tilde{H}_k^{-1}e^1\|^2 + \|H_k^{-1}w^k\|^2 \right\}. \tag{4.2}$$

5. Relations with the residual norm. In this section we use an expression for the norm of the residual of the FOM method to relate it to the norm of the error. The following result was proved in [24].

LEMMA 5.1. *The FOM norm of the residual is*

$$\|r^k\| = \|r^0\| h_{k+1,k} |(H_k^{-1}e^1, e^k)|. \tag{5.1}$$

We see that the norm of the residual is small if $h_{k+1,k}$ or $|(H_k^{-1}e^1, e^k)|$ (or both) are small. In particular if $h_{k+1,k} = 0$ we have found an invariant subspace of A and the solution of the linear system. The norm of the residual can be used in the expressions for the l_2 norm of the error. The next theorem shows that we obtain a simple and elegant formula for the norm of the error in terms of the norm of the residual.

THEOREM 5.2. *Assume the block partitioning of H_n as in equation (4.1) and denote $w^k = W_k\tilde{H}_k^{-1}e^1$. Then,*

$$\|\epsilon^k\|^2 = \|r^k\|^2 \frac{\|\tilde{H}_k^{-1}e^1\|^2 + \|H_k^{-1}w^k\|^2}{[1 - h_{k+1,k}(e^k, H_k^{-1}w^k)]^2}. \tag{5.2}$$

Proof. This result is obvious using the definition of γ_k, equation (4.2) and Lemma 5.1. ∎

From Theorem 5.2 we see that the (squares) of the norms of the error and the residual are close if and only if the multiplying factor in equation (5.2) is close to 1. Note that this factor depends on iterations $k+1$ to n through $\tilde{H}_k$ and w^k.

6. Estimates of the norm of the error in FOM. To approximate the norm of the error in FOM we use the same technique as in [12] and [18] introducing a delay d which is a strictly positive integer. At iteration k of FOM we approximate $\|\epsilon^{k-d}\|^2$ by replacing H_k by H_{k-d} and H_n by H_k in the formula of Theorem 4.3. We now use the partitioning

$$H_k = \begin{pmatrix} H_{k-d} & W_{k-d} \\ Y_{k-d}^T & \tilde{H}_{k-d} \end{pmatrix}.$$

Note that the notations are different from the previous ones since H_k is of order k and $\tilde{H}_{k-d}$ of order d. Then, at iteration k we approximate the square $\|\epsilon^{k-d}\|^2$ of the norm of the error at iteration $k-d$ using the result of Theorem 4.3 by

$$\begin{aligned}\chi_{k-d} = \|r^0\|^2 [\{ & h_{k-d+1,k-d} ((e^{k-d}, H_{k-d}^{-1}e^1) + \gamma_{k-d}(e^{k-d}, H_{k-d}^{-1}w^{k-d})) \}^2 \|\tilde{H}_{k-d}^{-1}e^1\|^2 \\ & + \gamma_{k-d}^2 \|H_{k-d}^{-1}w^{k-d}\|^2],\end{aligned}$$

with $w^{k-d} = W_{k-d}\tilde{H}_{k-d}^{-1}e^1$ and

$$\gamma_{k-d} = \frac{h_{k-d+1,k-d}(e^{k-d}, H_{k-d}^{-1}e^1)}{1 - h_{k-d+1,k-d}(e^{k-d}, H_{k-d}^{-1}w^{k-d})}.$$

Note that χ_{k-d} is the sum of two positive quantities. A rationale for the choice of this approximation is the following. Let us write the difference of the squares of the error norms at iterations $k-d$ and k using the formula of Theorem 3.1. Some terms cancel and we obtain

$$\begin{aligned}\|\epsilon^{k-d}\|^2 - \|\epsilon^k\|^2 = \|r^0\|^2 \{&(H_k^{-1}e^1, H_k^{-1}e^1) - (H_{k-d}^{-1}e^1, H_{k-d}^{-1}e^1)\\ &+ 2h_{k-d+1,k-d}(H_{k-d}^{-1}e^1, e^{k-d})((H_n^{-1}e^{k-d+1})^{k-d}, H_{k-d}^{-1}e^1)\\ &- 2h_{k+1,k}(H_k^{-1}e^1, e^k)((H_n^{-1}e^{k+1})^k, H_k^{-1}e^1)\,\}.\end{aligned}$$

Adding and subtracting the term

$$2\|r^0\|^2\, h_{k-d+1,k-d}(H_{k-d}^{-1}e^1, e^{k-d})((H_k^{-1}e^{k-d+1})^{k-d}, H_{k-d}^{-1}e^1)$$

and using manipulations similar to those in section 4, we obtain

$$\begin{aligned}\|\epsilon^{k-d}\|^2 - \|\epsilon^k\|^2 = \chi_{k-d} + \|r^0\|^2 \{&2h_{k-d+1,k-d}(H_{k-d}^{-1}e^1, e^{k-d})((H_n^{-1}e^{k-d+1})^{k-d}\\ &- (H_k^{-1}e^{k-d+1})^{k-d}, H_{k-d}^{-1}e^1) - 2h_{k+1,k}(H_k^{-1}e^1, e^k)((H_n^{-1}e^{k+1})^k, H_k^{-1}e^1)\,\}.\end{aligned}$$

It turns out that when FOM starts to converge we often have $\|\epsilon^k\|^2 \ll \|\epsilon^{k-d}\|^2$ and, additionally, the sum of the last two terms is small, meaning that χ_{k-d} gives a reasonable approximation of $\|\epsilon^{k-d}\|^2$.

Concerning the implementation, at iteration k we have to compute the last element of $H_{k-d}^{-1}e^1$, $\tilde{H}_{k-d}^{-1}e^1$, w^{k-d} and $H_{k-d}^{-1}w^{k-d}$. As for computing the iterate x^k we use rotations. The last element of $H_{k-d}^{-1}e^1$ is already known since this has been computed to obtain the iterate x^{k-d} or at least $\|r^{k-d}\|$ if the solution is computed only at the end of the iterations. Computing $\tilde{H}_{k-d}^{-1}e^1$ can be done with $d-1$ rotations. This is not expensive since usually d will be small. The most computationally expensive part is obtaining $H_{k-d}^{-1}w^{k-d}$. This again is done using the rotations computed in the Arnoldi process from step 1 to step $k-d-1$. Moreover we have to apply the rotations to the columns of W_{k-d}. But this has already been done during FOM iterations $k-d$ to k. The result is a part of the triangular matrix R_k that is used to compute the solution at iteration k. We then have to take a linear combination of the columns of W_{k-d} and to solve a triangular system to obtain $H_{k-d}^{-1}w^{k-d}$. Note that these computations are more expensive than for CG where the estimates of the norm of the error are essentially obtained for free; see [18], [21] and [27].

Let us consider a few numerical examples. In these experiments we use the FOM method with a QR factorization to solve the linear system $H_k z^k = \|r^0\|e^1$. Moreover, we use the implementation given in [28]. When computing the Arnoldi vector v^{k+1} one compares Av^k with the result of the modified Gram-Schmidt step before normalization. If the norm of the result is smaller than $\tau\|Av^k\|$ a reorthogonalization is performed. In some of the examples described below this was used for the very last iterations to avoid the near singularity of the matrices H_k. This must not be considered has a limitation for the estimates of the norm of the error. If one does not have to go up to the very last iteration, the estimates can be used with the backward stable MGS-GMRES without any reorthogonalization.

All the matrices which are considered are not normal. The first example E1 is the matrix e05r0500 from the Matrix Market, arising from a driven cavity fluid dynamics problem with a Reynolds number $Re = 500$. The order is $n = 236$, the condition number is $\kappa(A) = 1.16\ 10^6$, the extreme singular values are $\min(\sigma_i) = 4.9\ 10^{-5}$ and $\max(\sigma_i) = 57.20$. For this matrix there are complex eigenvalues with a negative real part and it is not positive real. The right hand side comes from the Matrix Market files. This computation was done without any reorthogonalization. The results are displayed in figure 6.1. The solid curve displays the norm of the error computed with the solution given by Gaussian elimination with pivoting. In the left part the dotted curve is the norm of the residual. We see that the residual norm is oscillating and increasing at the beginning. There is no decrease before iteration 150. The error norm is smoother than the residual norm which is much smaller than the error norm but they both start to decrease almost at the same time. The estimate of the error norm with $d = 1$ is not very good at the beginning since it is smaller than the error norm by several orders of magnitude but after 100 iterations it gives the right error level. Moreover it captures well the large peaks in the error norm curve. The oscillations of the estimate can be smoothed to some extent by increasing the value of d as we will see with the next example. One may ask why the estimate is wrong at the beginning of the computation. The main reason is that for d small $\|\tilde{H}_{k-d}^{-1}e^1\|$ is a bad approximation to the exact value. Moreover the vector $w^{k-d} = W_{k-d}\tilde{H}_{k-d}^{-1}e^1$ is not a good approximation of the exact unknown value when d is small and when we are at the beginning of the computation in the case where no convergence takes place even though this is not as critical as for $\|\tilde{H}_{k-d}^{-1}e^1\|$. To get good estimates for the first iterations we have to substantially increase the delay d. The right part of figure 6.1 compares the results with $d = 1$ and $d = 100$. Note that with $d = 100$ we can only compute an estimate up to iteration 136. We see that with a large value of d we capture quite well the level of stagnation of the error norm for the first 150 iterations. However, using such a large value of d is generally not practical and moreover it becomes useless when FOM starts to converge. Then, a small value of d gives a good approximation of the error norm. It would be nice if one can choose d adaptively, but this does not seem easy to achieve.

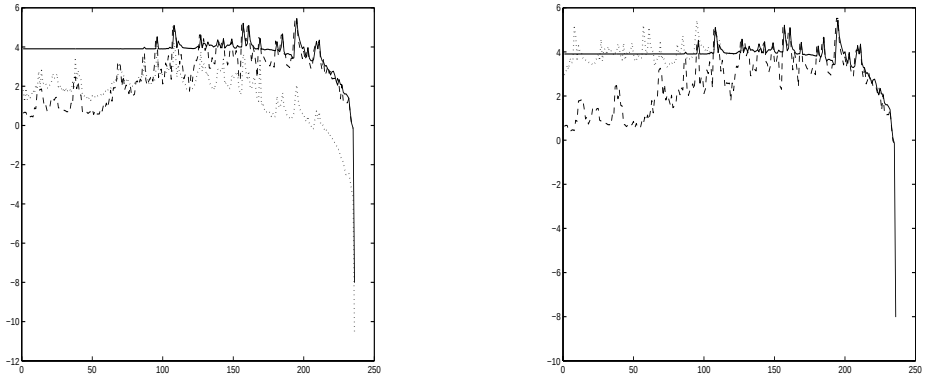

FIG. 6.1. *FOM E1:* $\log_{10}$ *of the error norm (solid), (left) estimate* $d = 1$ *(dashed), residual norm (dotted), (right) estimates* $d = 1$ *(dashed) and* $d = 100$ *(dotted)*

The second example E2 is the matrix steam1 from the Matrix Market from a 3D steam model of oil reservoir. The order is $n = 240$, the condition number is

$\kappa(A) = 2.82\ 10^7$, the extreme singular values are $\min(\sigma_i) = 0.767$, $\max(\sigma_i) = 2.17\ 10^7$. The eigenvalues of the matrix are real and negative. We use a random right hand side. For this example we use reorthogonalization for the very last iterations. The results are given in figure 6.2 for $d = 1$ in the left part and for $d = 10$ and $d = 20$ in the right part. Similarly as in the previous example, the residual norm is widely oscillating. However, the error norm is smooth but there is not much improvement for more than 100 iterations. The estimate for $d = 1$ is oscillating and it captures well the final decrease but we would like to obtain something better. Increasing the value of d improves the estimate of the error norm as it can be seen in figure 6.2 with $d = 10$ and 20. The level of the error norm before the decrease is well reproduced by these estimates.

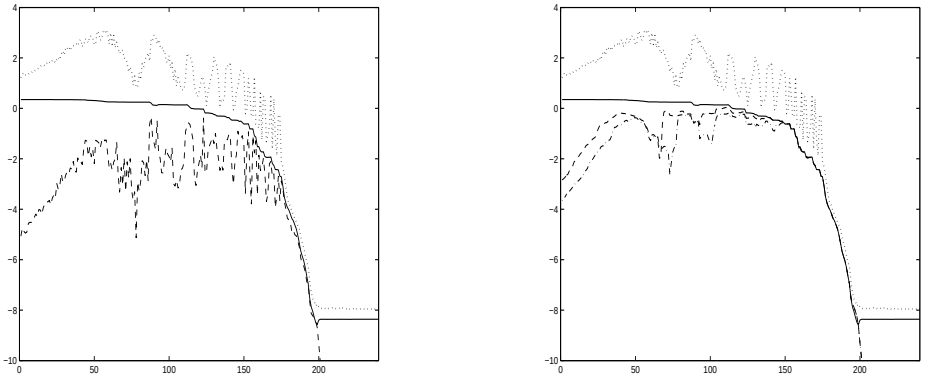

FIG. 6.2. *FOM E2:* $\log_{10}$ *of the error norm (solid) and residual norm (dotted), (left) estimate* $d = 1$ *(dashed), (right) estimates* $d = 10$ *(dash-dotted) and* $d = 20$ *(dashed)*

The third example E3 is the matrix steam2 from the Matrix Market arising from a 3D steam model of oil reservoir. The order is $n = 600$, the condition number is $\kappa(A) = 3.78\ 10^6$, the extreme singular values are $\min(\sigma_i) = 1238.55$, $\max(\sigma_i) = 4.68\ 10^9$. The eigenvalues of the matrix are real and negative. We use a random right hand side. The results are given in figure 6.3 in the left part for $d = 1$ and in the right part for $d = 10$. This example is interesting since even though the residual norm is oscillating we have successively plateaus where the error norm is almost stagnating and short sequences with a rapid decrease. This is an example for which it is misleading to use the norm of the residual to stop the iterations since the norm of the error is much smaller than the norm of the residual. With $d = 10$ the estimate is quite close to the exact norm of the error except at the beginning of the iterations.

The fourth example E4 arises from the discretization of a 2D convection–diffusion problem,

$$-\Delta u + 2e^{2(x^2+y^2)}\frac{\partial u}{\partial x} = f,$$

in the unit square with Dirichlet boundary conditions using finite differences and upwind differencing for the first order term. The cartesian regular mesh is 50×50, excluding boundaries. Therefore the order of the matrix A is $n = 2500$. The condition number is $\kappa(A) = 1359.18$ and the extreme singular values are $\min(\sigma_i) = 7.55\ 10^{-3}$, $\max(\sigma_i) = 10.26$. The real parts of the eigenvalues are positive. We use a random right hand side for the linear system. Figure 6.4 displays the results for the error and

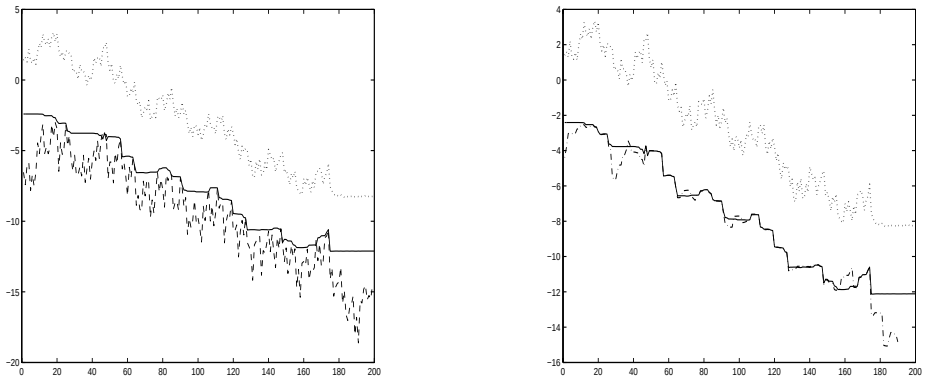

FIG. 6.3. *FOM E3:* $\log_{10}$ *of the error norm (solid) and residual norm (dotted), (left) estimate* $d = 1$ *(dashed), (right) estimate* $d = 10$ *(dash-dotted)*

residual norms and the estimate for $d = 1$ in the left part. The right part of the figure compares the estimates for $d = 1, 10$ and 20. There is not much improvement between $d = 10$ and $d = 20$ since the estimate for $d = 10$ is already quite close to the error norm. For this problem FOM converges quite well with a smooth residual norm.

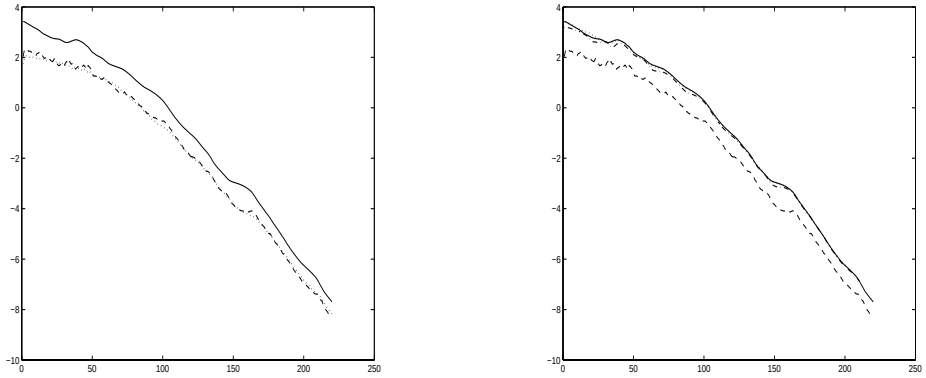

FIG. 6.4. *FOM E4:* $\log_{10}$ *of the error norm (solid), (left) estimate* $d = 1$ *(dashed) and residual norm (dotted), (right) estimates* $d = 1$ *(dashed),* $d = 10$ *(dash-dotted) and* $d = 20$ *(dotted)*

7. Relations between FOM and GMRES. Before considering computing the norm of the error in GMRES we look at the relations between the OR method (FOM) and the MR method (GMRES). It is well known that there are simple relations between the two, at least for the residual norms, see for instance [6], [8], [10] and [9]. We will give an expression which involves the matrices H_k. We denote with an index O (resp. M) the variables for the OR–FOM (resp. MR–GMRES) method. Let t^k be the last column of $(H_k^T H_k)^{-1}$ and t_k^k its last element, that is,

$$t_k^k = (e^k, (H_k^T H_k)^{-1} e^k) = \|H_k^{-T} e^k\|^2. \tag{7.1}$$

THEOREM 7.1. *Let*

$$\delta_{k+1} = \frac{h_{k+1,k}^2}{1 + h_{k+1,k}^2 t_k^k},$$

and $u^k = \delta_{k+1} t^k$. *Then,*

$$x_M^k = x_O^k - (z_O^k)_k V_k u^k,$$

$$\epsilon_M^k = \epsilon_O^k + (z_O^k)_k V_k u^k,$$

where z_O^k *is the coordinate vector of the FOM method,* $z_O^k = \|r^0\| H_k^{-1} e^1$.

Proof. Let $y = h_{k+1,k} e^k$. We have

$$H_k^{(e)} = \begin{pmatrix} H_k \\ y^T \end{pmatrix}$$

and

$$[H_k^{(e)}]^T H_k^{(e)} = H_k^T H_k + y y^T.$$

Hence, $[H_k^{(e)}]^T H_k^{(e)}$ is a rank–one modification of $H_k^T H_k$. This means that we can use the Sherman–Morrison formula (see [13]) to compute the inverse of $[H_k^{(e)}]^T H_k^{(e)}$. We have

$$([H_k^{(e)}]^T H_k^{(e)})^{-1} = (H_k^T H_k)^{-1} - \frac{((H_k^T H_k)^{-1} y)(y^T (H_k^T H_k)^{-1})}{1 + (y, (H_k^T H_k)^{-1} y)}.$$

Then, assuming that the initial residual is the same in both algorithms, the solution z_M^k is given by

$$z_M^k = \|r^0\| \left[(H_k^T H_k)^{-1} - \delta_{k+1} ((H_k^T H_k)^{-1} e^k)((e^k)^T (H_k^T H_k)^{-1}) \right] H_k^T e^1.$$

Noting that $z_O^k = \|r^0\| H_k^{-1} e^1$ this shows that

$$z_M^k = z_O^k - (z_O^k)_k u^k.$$

This proves the relation for x_M^k. Subtracting the solution x on both sides and changing signs we obtain the relation for ϵ_M^k. ∎

We now turn to the relation between the residual norms of FOM and GMRES.

THEOREM 7.2. *Using* t_k^k *defined in equation (7.1) we have*

$$\|r_M^k\|^2 = \frac{\|r_O^k\|^2}{1 + h_{k+1,k}^2 t_k^k}. \tag{7.2}$$

Proof. Denoting $\omega_k = (z_O^k)_k$ for simplicity, from Theorem 7.1 we have

$$r_M^k = b - A x_M^k = r_O^k + \omega_k A V_k u^k,$$

where $u^k = \delta_{k+1} t^k$. This gives

$$\|r_M^k\|^2 = \|r_O^k\|^2 + 2\omega_k (r_O^k, A V_k u^k) + \omega_k^2 (A V_k u^k, A V_k u^k). \tag{7.3}$$

Let us first consider $(r_O^k, A V_k u^k)$ in the second term. Using equation (2.1), the orthogonality of the basis vectors and the relation of r_O^k and v^{k+1} we have

$$(r_O^k, A V_k u^k) = -h_{k+1,k}^2 \|r^0\| (H_k^{-1} e^1, e^k)(e^k, u^k).$$

Multiplying by $2\omega_k$ we obtain

$$\begin{aligned} 2\omega_k(r_O^k, AV_k u^k) &= -2\omega_k h_{k+1,k}^2 \|r^0\|(H_k^{-1}e^1, e^k)\delta_{k+1}t_k^k, \\ &= -2h_{k+1,k}^2\|r^0\|^2(H_k^{-1}e^1, e^k)^2\delta_{k+1}t_k^k, \\ &= -2\|r_O^k\|^2\delta_{k+1}t_k^k. \end{aligned}$$

Using again equation (2.1) the term $(AV_k u^k, AV_k u^k)$ is equal to

$$(V_k H_k u^k, V_k H_k u^k) + 2h_{k+1,k}(V_k H_k u^k, v^{k+1}(e^k)^T u^k) + h_{k+1,k}^2(v^{k+1}(e^k)^T u^k, v^{k+1}(e^k)^T u^k).$$

The second term is zero because $V_k^T v^{k+1} = 0$. The first term is equal to $(H_k u^k, H_k u^k)$ but $u^k = \delta_{k+1}H_k^{-1}H_k^{-T}e^k$, therefore,

$$(V_k H_k u^k, V_k H_k u^k) = \delta_{k+1}^2(H_k^{-T}e^k, H_k^{-T}e^k) = \delta_{k+1}^2 t_k^k.$$

For the last term we have

$$h_{k+1,k}^2(v^{k+1}(e^k)^T u^k, v^{k+1}(e^k)^T u^k) = h_{k+1,k}^2(u_k^k)^2 = h_{k+1,k}^2\delta_{k+1}^2(t_k^k)^2.$$

Hence, using the definition of δ_{k+1},

$$\begin{aligned} (AV_k u^k, AV_k u^k) &= \delta_{k+1}^2 t_k^k + h_{k+1,k}^2\delta_{k+1}^2(t_k^k)^2, \\ &= \delta_{k+1}^2 t_k^k(1 + h_{k+1,k}^2 t_k^k), \\ &= h_{k+1,k}^2\delta_{k+1}t_k^k. \end{aligned}$$

In equation (7.3) this last term is multiplied by ω_k^2. This gives

$$\omega_k^2(AV_k u^k, AV_k u^k) = \|r^0\|^2(H_k^{-1}e^1, e^k)^2 h_{k+1,k}^2\delta_{k+1}t_k^k = \|r_O^k\|^2\delta_{k+1}t_k^k.$$

Putting all these results together we obtain

$$\|r_M^k\|^2 = \|r_O^k\|^2(1 - \delta_{k+1}t_k^k).$$

But

$$1 - \delta_{k+1}t_k^k = 1 - \frac{h_{k+1,k}^2 t_k^k}{1 + h_{k+1,k}^2 t_k^k} = \frac{1}{1 + h_{k+1,k}^2 t_k^k},$$

which gives the final expression for $\|r_M^k\|^2$. □

The relation between $\|r_M^k\|^2$ and $\|r_O^k\|^2$ shows that $\|r_M^k\|^2 \leq \|r_O^k\|^2$. This was obvious since GMRES is a minimal residual method. For the residual norms of FOM and GMRES to be close we need to have either $h_{k+1,k}$ and/or t_k^k small. The result of Theorem 7.2 can be related to previous results in the literature. Expressions for the norms of residual vectors of FOM and GMRES and their relations using the sines and cosines of the rotations were given in [6] and [8].

8. Formulas for the error norm in GMRES. We use the relations between the errors in Theorem 7.1 to obtain an expression for the error norm in GMRES.

THEOREM 8.1. *Assume the block partitioning of H_n as in equation (4.1). Denote $w^k = W_k \tilde{H}_k^{-1} e^1$ and*

$$\gamma_k = \frac{h_{k+1,k}(e^k, H_k^{-1}e^1)}{1 - h_{k+1,k}(e^k, H_k^{-1}w^k)}.$$

The vector t^k being the last column of $(H_k^T H_k)^{-1}$, t_k^k its last element defined in equation (7.1),

$$\delta_{k+1} = \frac{h_{k+1,k}^2}{1 + h_{k+1,k}^2 t_k^k},$$

and $u^k = \delta_{k+1} t^k$, we have

$$\|\epsilon_M^k\|^2 = \|\epsilon_O^k\|^2 + \|r^0\|^2[2\gamma_k(e^k, H_k^{-1}e^1)(H_k^{-1}w^k, u^k) + (e^k, H_k^{-1}e^1)^2\|u^k\|^2]. \tag{8.1}$$

Proof. From Theorem 7.1 and denoting $\omega_k = (z_O^k)_k$ for simplicity we have

$$\epsilon_M^k = \epsilon_O^k + \omega_k V_k u^k.$$

The norm of ϵ_M^k is

$$\|\epsilon_M^k\|^2 = \|\epsilon_O^k\|^2 + 2\omega_k(\epsilon_O^k, V_k u^k) + \omega_k^2(V_k u^k, V_k u^k).$$

Since $(V_k u^k, V_k u^k) = (u^k, u^k)$ we are left with computing $(\epsilon_O^k, V_k u^k)$. We have

$$\epsilon_O^k = A^{-1} r_O^k = A^{-1} r^0 - V_k z_O^k.$$

But $r^0 = \|r^0\| V_n e^1$ and $A^{-1} V_n = V_n H_n^{-1}$. Therefore

$$\epsilon_O^k = \|r^0\| V_n H_n^{-1} e^1 - V_k z_O^k.$$

This gives the decomposition of the error in FOM over the vectors v^j of the orthonormal basis of the Krylov subspace. Multiplying by V_k^T we obtain

$$V_k^T \epsilon_O^k = \|r^0\|(H_n^{-1}e^1)^k - z_O^k,$$

where, once again, $(H_n^{-1}e^1)^k$ denotes the k first components of the first column of the inverse of H_n. Then,

$$(\epsilon_O^k, V_k u^k) = \|r^0\|((H_n^{-1}e^1)^k, u^k) - (z_O^k, u^k) = \|r^0\|[((H_n^{-1}e^1)^k, u^k) - (H_k^{-1}e^1, u^k)].$$

But we have seen in the proof of Lemma 4.1 that $(H_n^{-1}e^1)^k = H_k^{-1}e^1 + \gamma_k H_k^{-1} w^k$. Therefore

$$(\epsilon_O^k, V_k u^k) = \|r^0\|\gamma_k(H_k^{-1}w^k, u^k).$$

□

It is interesting to write the GMRES error norm as a function of the residual norms.

THEOREM 8.2. *Assume the block partitioning of H_n as in equation (4.1). Denote $w^k = W_k \tilde{H}_k^{-1} e^1$ and let t^k be the last column of $(H_k^T H_k)^{-1}$ and t_k^k its last element defined in equation (7.1). Then,*

$$\|\epsilon_M^k\|^2 = \|\epsilon_O^k\|^2 + 2\|r_M^k\|^2 \frac{h_{k+1,k}}{1 - h_{k+1,k}(e^k, H_k^{-1} w^k)} (H_k^{-1} w^k, t^k) + \|r_M^k\|^2 \frac{\|t^k\|^2}{1 + h_{k+1,k}^2 t_k^k}.$$

Proof. We already know that $\|\epsilon_O^k\|^2$ can be written in terms of $\|r_O^k\|^2$. Let us concentrate on the last two terms. Using the definition of γ_k and u^k we have

$$\begin{aligned} 2\|r^0\|^2 \gamma_k (H_k^{-1} e^1, e^k)(H_k^{-1} w^k, u^k) &= 2 \frac{\|r_O^k\|^2}{1 - h_{k+1,k}(e^k, H_k^{-1} w^k)} \frac{h_{k+1,k}}{1 + h_{k+1,k}^2 t_k^k} (H_k^{-1} w^k, t^k), \\ &= 2\|r_M^k\|^2 \frac{h_{k+1,k}}{1 - h_{k+1,k}(e^k, H_k^{-1} w^k)} (H_k^{-1} w^k, t^k). \end{aligned}$$

The other term is

$$\begin{aligned} \|r^0\|^2 (H_k^{-1} e^1, e^k)^2 \|u^k\|^2 &= \|r^0\|^2 \delta_{k+1}^2 (H_k^{-1} e^1, e^k)^2 \|t^k\|^2, \\ &= \|r_O^k\|^2 \frac{\|t^k\|^2}{(1 + h_{k+1,k}^2 t_k^k)^2}, \\ &= \|r_M^k\|^2 \frac{\|t^k\|^2}{1 + h_{k+1,k}^2 t_k^k}. \end{aligned}$$

□

Note that the last term in the formula of Theorem 8.2 is positive but unfortunately we do not know the sign of the middle term. Therefore we cannot tell if $\|\epsilon_M^k\|$ is smaller or larger than $\|\epsilon_O^k\|$. However, we can express everything in terms of $\|r_M^k\|$ as in the following Corollary.

COROLLARY 8.3. *The square of the norm of the error in GMRES, $\|\epsilon_M^k\|^2$, is equal to $\pi_k \|r_M^k\|^2$ where*

$$\pi_k = \frac{\|(1 + h_{k+1,k}^2 t_k^k) H_k^{-1} w^k + (1 - h_{k+1,k}(e^k, H_k^{-1} w^k)) t^k\|^2 + (1 + h_{k+1,k}^2 t_k^k)^2 \|\tilde{H}_k^{-1} e^1\|^2}{(1 - h_{k+1,k}(e^k, H_k^{-1} w^k))^2 (1 + h_{k+1,k}^2 t_k^k)}.$$

Proof. From equation (5.2) we have

$$\|\epsilon_O^k\|^2 = \|r_O^k\|^2 \frac{\|\tilde{H}_k^{-1} e^1\|^2 + \|H_k^{-1} w^k\|^2}{[1 - h_{k+1,k}(e^k, H_k^{-1} w^k)]^2}.$$

Taking the term

$$\|r_O^k\|^2 \frac{\|H_k^{-1} w^k\|^2}{[1 - h_{k+1,k}(e^k, H_k^{-1} w^k)]^2} = \|r_M^k\|^2 (1 + h_{k+1,k}^2 t_k^k) \frac{\|H_k^{-1} w^k\|^2}{[1 - h_{k+1,k}(e^k, H_k^{-1} w^k)]^2},$$

and adding it to the two last terms of Theorem 8.2, the numerator is

$$(1+h_{k+1,k}^2 t_k^k)^2 \|H_k^{-1}w^k\|^2 + 2(1+h_{k+1,k}^2 t_k^k)(1-h_{k+1,k}(e^k, H_k^{-1}w^k))(H_k^{-1}w^k, t^k) \\ + (1-h_{k+1,k}(e^k, H_k^{-1}w^k))^2 \|t^k\|^2,$$

and the denominator is

$$(1-h_{k+1,k}(e^k, H_k^{-1}w^k))^2(1+h_{k+1,k}^2 t_k^k).$$

This ratio is to be multiplied by $\|r_M^k\|^2$. The remaining term is

$$\|r_M^k\|^2 \frac{1+h_{k+1,k}^2 t_k^k}{(1-h_{k+1,k}(e^k, H_k^{-1}w^k))^2} \|\tilde{H}_k^{-1}e^1\|^2.$$

□

Fortunately $\pi_k > 0$. However it seems difficult to know if it is smaller or larger than 1.

9. Estimates of the error norm in GMRES. To estimate the GMRES error norm we use equation (8.1) and the estimate for the FOM error norm defined in section 6. Using an integer delay d, the additional term for the estimate of the square of the error norm at iteration $k-d$ is

$$\|r^0\|^2[2\gamma_{k-d}(H_{k-d}^{-1}e^1, e^{k-d})(H_{k-d}^{-1}w^{k-d}, u^{k-d}) + (H_{k-d}^{-1}e^1, e^{k-d})^2\|u^{k-d}\|^2],$$

where the definitions are similar to the ones in Theorem 8.1.

For the numerical experiments we use the same examples as in section 6. The results for example E1 are given in figure 9.1. Of course the residual norm curve is monotonic even though it is almost stagnating for more than 150 iterations. The error norm curve is not monotonic but quite smooth and the estimate well capture the error norm level at least after 100 iterations. The estimate is smoother than the one for FOM. This shows that the additional term has some smoothing effect. One can remark that the error norms for FOM and GMRES are close except that the FOM error norm has small peaks. The residual norms are more different since the FOM residual is widely oscillating.

We also compare our estimate with some estimates proposed by Brezinski (see [4], [5]). The first one is

$$\epsilon_{Br1}^k = \frac{\|r^k\|^2}{\|A^T r^k\|}, \; r^k = b - Ax^k.$$

In all the formulas proposed in [5], this one corresponds to a formula proposed by Auchmuty [3]. We also used the general formula in equation (9) in [5] that is,

$$\epsilon_{Br2}^k = \left(c_0^{\nu_1}(c_1^2)^{3-\nu}c_2^{\nu-4}\right)^{1/2},$$

with $c_0 = (r^k, r^k)$, $c_1 = (r^k, Ar^k)$ and $c_2 = (Ar^k, Ar^k)$. We chose $\nu = 4$. These quantities are shown in the right part of figure 9.1. We can see that on this problem Brezinki's first estimate underestimates the level of the error norm, but it reflects the near stagnation of the error norm. The other estimate with $\nu = 4$ oscillates but is

close to our result. Figure 9.2 compares the backward error (see [1], [7], [16]) defined as

$$\epsilon_B^k = \frac{\|b - Ax^k\|}{\|A\| \, \|x^k\| + \|b\|},$$

to the relative norm of the error $\|x - x^k\|/\|x\|$ and our estimate normalized by the norm of the exact solution. The backward error is almost always decreasing. In such a case basing the stopping criterion on the backward error may also be misleading.

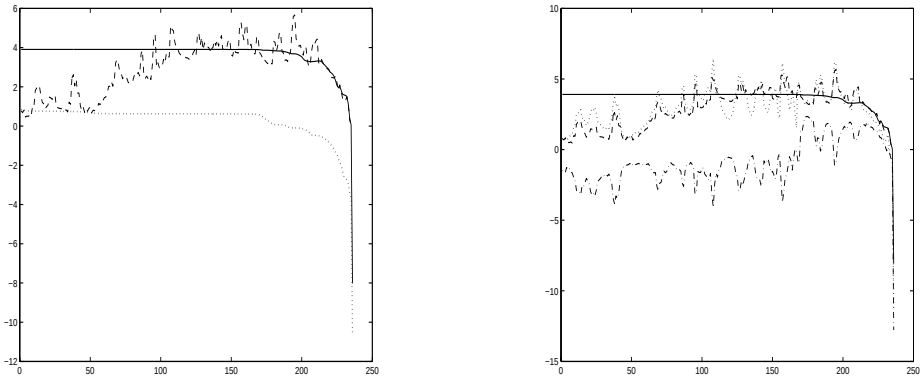

FIG. 9.1. *GMRES E1:* $\log_{10}$ *of the error norm (solid), (left) estimate* $d = 1$ *(dashed) and residual norm (dotted), (right) estimate* $d = 1$ *(dashed), Brezinski's first estimate (dash-dotted) and second estimate with* $\nu = 4$ *(dotted)*

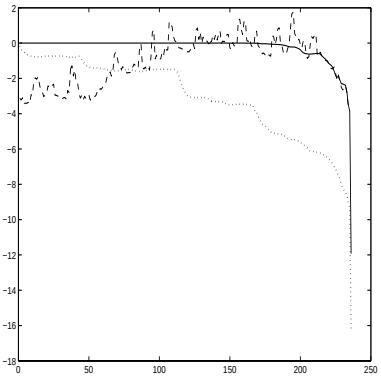

FIG. 9.2. *GMRES E1:* $\log_{10}$ *of the norm of the relative error (solid), relative norm estimate* $d = 1$ *(dashed), backward error (dotted)*

Figure 9.3 displays the results for example E2. The estimate with $d = 1$ is much better than with FOM since the error norm is well approximated after 50 iterations. Increasing d improves the result. The error norm curves of FOM and GMRES are almost identical even though the residual norms are much different.

The results for example E3 are given in figure 9.4. Again the result with $d = 1$ is much smoother than with FOM giving a good estimate of the error norm. Using $d = 10$ gives a very good estimate of the error. The error curves for FOM and GMRES are not much different even though the one for FOM is less regular.

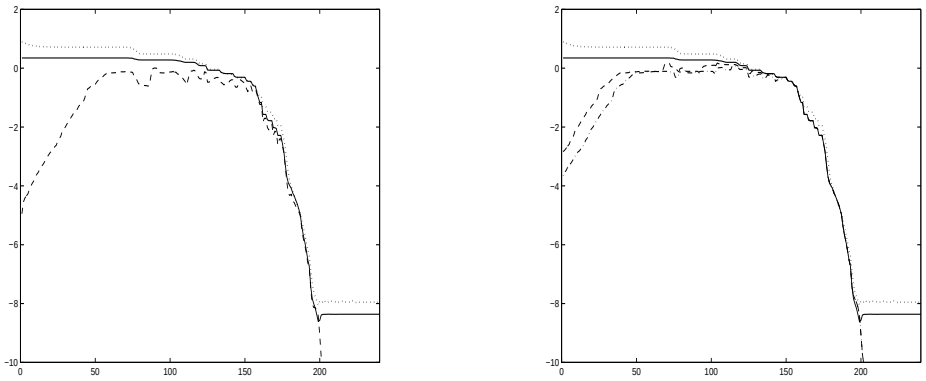

FIG. 9.3. *GMRES E2:* $\log_{10}$ *of the error norm (solid) and residual norm (dotted), (left) estimate* $d = 1$ *(dashed) , (right) estimates* $d = 10$ *(dash-dotted) and* $d = 20$ *(dashed)*

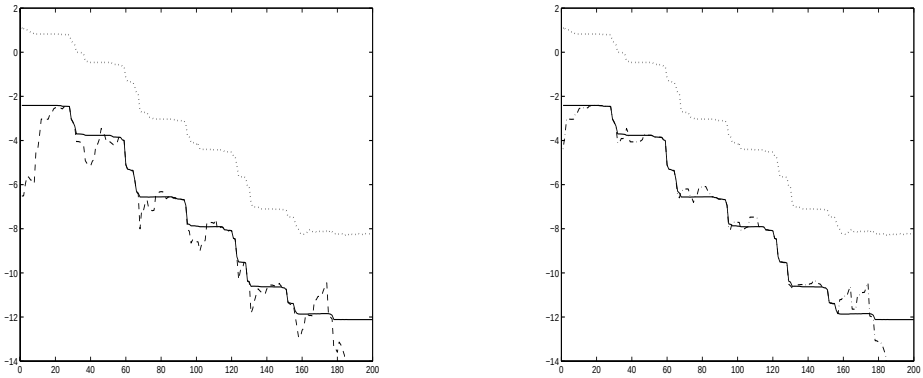

FIG. 9.4. *GMRES E3:* $\log_{10}$ *of the error norm (solid) and residual norm (dotted), (left) estimate* $d = 1$ *(dashed) , (right) estimate* $d = 10$ *(dash-dotted)*

Finally we consider example E4. Figure 9.5 shows that the estimate of the GMRES error norm with $d = 1$ is much closer to the exact error than in the FOM case. Therefore, increasing d does not improve too much the estimate. In this example the curves for the error norms in FOM and GMRES as well as the residual curves are not much different.

10. Conclusions. In this paper we have given expressions for the error l_2 norm in FOM and GMRES. We have shown how to use these results to compute estimates of the error norm during the iterations. This is done by introducing a delay d and computing an estimate d iterations before the current one. Numerical experiments have shown that this generally gives good approximations of the error norm (even for small values of d) particularly in the phase where convergence takes place. It would be interesting to use these estimates to set up a reliable stopping criterion for FOM and GMRES in combination with backward error-type criterion since using the residual norm can sometimes be misleading. This technique can also be used for preconditioned FOM or GMRES since we simply have to apply the formulas developed in this paper to the linear system $M^{-1}Ax = M^{-1}b$ when using a left preconditioner. We can also use the same technique for restarted iterations like in FOM(m) or GMRES(m).

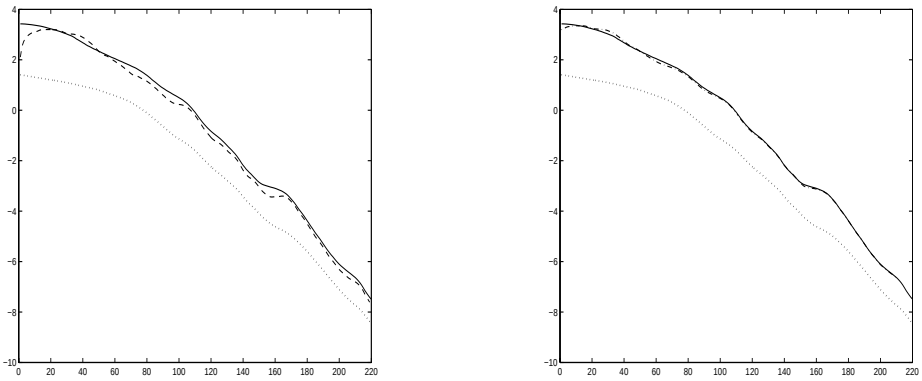

FIG. 9.5. *GMRES E4:* $\log_{10}$ *of the error norm (solid) and residual norm (dotted), (left) estimate* $d = 1$ *(dashed) , (right) estimate* $d = 10$ *(dash-dotted)*

However, the delay d which is used has to be smaller than m and it is not always guaranteed that good estimates of the error norm can be obtained if m is small since this constraints d.

Several remaining questions have to be addressed. First it would be important to know if the expressions for the error norms are still valid in finite precision arithmetic up to terms proportional to the unit roundoff. This question has been thoroughly considered for CG in [27] where it was shown that some formulas which are equivalent in exact arithmetic can have a different behavior in finite precision computations. Numerical experiments seem to show that the proposed estimates still work in finite precision but it would probably be enlightening to prove it. Secondly it would be interesting to study if the exact expressions for the error norms can lead to a better understanding of FOM and GMRES convergence.

Acknowledgments.

This work arises from studies started in 2002 but it was mainly written in 2008 and 2009 while visiting the Institute of Computer Science of the Czech Academy of Sciences in Prague with the support of a GAAS grant IAA100300802 and of the Institutional Research Plan AV0Z10300504. The revised version was written in 2010 during a visit to the Nečas Center of Charles University in Prague supported by the grant Jindrich Nečas Center for mathematical modeling, project LC06052 financed by MSMT. The author thanks particularly Miroslav Rozložník and Zdeněk Strakoš for their kind hospitality. Detailed remarks and comments from the referees lead to improvements in the presentation.

REFERENCES

[1] M. ARIOLI, I.S. DUFF AND D. RUIZ, *Stopping criteria for iterative solvers*, SIAM J. Matrix Anal. Appl., v 13, (1992), pp 138–144.

[2] W.E. ARNOLDI, *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quarterly of Appl. Math., v 9, (1951), pp 17–29.

[3] G. AUCHMUTY, *A posteriori error estimates for linear equations*, Numer. Math., v 61, (1992), pp 1–6.

[4] C. BREZINSKI, *Error estimates in the solution of linear systems*, SIAM J. Sci. Comput., v 21 n 2, (1999), pp 764–781.

[5] C. BREZINSKI, G. RODRIGUEZ AND S. SEATZU, *Error estimates for linear systems with applications to regularization*, Numerical Algorithms, v 49, (2008), pp 85–104.

[6] P.N. Brown, *A theoretical comparison of the Arnoldi and GMRES algorithms*, SIAM J. Sci. Stat. Comput., v 12 n 1, (1991), pp 58–78.
[7] F. Chaitin-Chatelin and V. Frayssé, *Lectures on finite precision computations*, SIAM, (1996).
[8] J. Cullum and A. Greenbaum, *Relation between Galerkin and norm–minimizing iterative methods for solving linear systems*, SIAM J. Matrix Anal. Appl., v 17 n 2, (1996), pp 223–247.
[9] M. Eiermann and O. Ernst, *Geometric aspects of the theory of Krylov subspace methods*, Acta Numerica v 10, Cambridge University Press, (2001), pp 251–312.
[10] B. Fischer, *Polynomial based iteration methods for symmetric linear systems*, Wyley–Teubner, (1996).
[11] G.H. Golub and G. Meurant, *Matrices, moments and quadrature with applications*, Princeton University Press, (2010).
[12] G.H. Golub and Z. Strakoš, *Estimates in quadratic formulas*, Numerical Algorithms, v 8, (1994), pp 241–268.
[13] G.H. Golub and C. Van Loan, *Matrix Computations*, Third Edition, Johns Hopkins University Press, (1996).
[14] A. Greenbaum, *Iterative methods for solving linear systems*, SIAM, (1997).
[15] M.R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards, v 49 n 6, (1952), pp 409–436.
[16] N.J. Higham, *Accuracy and stability of numerical algorithms*, second edition, SIAM, (2002).
[17] M. Hochbruck and C. Lubich, *Error analysis of Krylov methods in a nutshell*, SIAM J. Sci. Comput., v 19, (1998), pp 695–701.
[18] G. Meurant, *The computation of bounds for the norm of the error in the conjugate gradient algorithm*, Numerical Algorithms, v 16, (1997), pp 77–87.
[19] G. Meurant, *Numerical experiments in computing bounds for the norm of the error in the preconditioned conjugate gradient algorithm*, Numerical Algorithms, v 22, (1999), pp 353–365.
[20] G. Meurant, *Estimates of the l_2 norm of the error in the conjugate gradient algorithm*, Numerical Algorithms, v 40 n 2, (2005), pp 157–169.
[21] G. Meurant, *The Lanczos and Conjugate Gradient algorithms, from theory to finite precision computations*, SIAM, (2006).
[22] C.C. Paige, M. Rozložník and Z. Strakoš, *Modified Gram-Schmidt (MGS), least squares, and backward stability of MGS-GMRES*, SIAM J. Matrix Anal. Appl., v 28, (2006), pp 264–284.
[23] Y. Saad, *Iterative methods for sparse linear systems*, 2nd edition, SIAM, (2003).
[24] Y. Saad, *Krylov subspace methods for solving large unsymmetric linear systems*, Math. Comp., v 37 (1981), pp 105–126.
[25] Y. Saad, *Practical use of some Krylov subspace methods for solving indefinite and unsymmetric linear systems*, SIAM J. Sci. Stat. Comput., v 5, (1984), pp 203–228.
[26] Y. Saad and M.H. Schultz, *GMRES: a generalized minimum residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., v 7 n 3, (1986), pp 856–869.
[27] Z. Strakoš and P. Tichý, *On error estimates in the conjugate gradient method and why it works in finite precision computations*, Elec. Trans. Numer. Anal., v 13, (2002), pp 56–80.
[28] H.A. van der Vorst, *Iterative Krylov methods for large linear systems*, Cambridge University Press, (2003).