SIAM J. SCI. COMPUT.
Vol. 23, No. 6, pp. 1898–1923

RESIDUAL AND BACKWARD ERROR BOUNDS IN MINIMUM RESIDUAL KRYLOV SUBSPACE METHODS*

CHRISTOPHER C. PAIGE† AND ZDENĚK STRAKOŠ‡

Abstract. Minimum residual norm iterative methods for solving linear systems $Ax = b$ can be viewed as, and are often implemented as, sequences of least squares problems involving Krylov subspaces of increasing dimensions. The minimum residual method (MINRES) [C. Paige and M. Saunders, *SIAM J. Numer. Anal.*, 12 (1975), pp. 617–629] and generalized minimum residual method (GMRES) [Y. Saad and M. Schultz, *SIAM J. Sci. Statist. Comput.*, 7 (1986), pp. 856–869] represent typical examples. In [C. Paige and Z. Strakoš, *Bounds for the least squares distance using scaled total least squares*, Numer. Math., to appear] revealing upper and lower bounds on the residual norm of any linear least squares (LS) problem were derived in terms of the total least squares (TLS) correction of the corresponding scaled TLS problem. In this paper theoretical results of [C. Paige and Z. Strakoš, *Bounds for the least squares distance using scaled total least squares*, Numer. Math., to appear] are extended to the GMRES context. The bounds that are developed are important in theory, but they also have fundamental practical implications for the finite precision behavior of the modified Gram–Schmidt implementation of GMRES, and perhaps for other minimum norm methods.

Key words. linear equations, eigenproblem, large sparse matrices, iterative solution, Krylov subspace methods, Arnoldi method, generalized minimum residual method, modified Gram–Schmidt, least squares, total least squares, singular values

AMS subject classifications. 65F10, 65F20, 65F25, 65F50, 65G05, 15A42

PII. S1064827500381239

1. Introduction.

Consider a system of linear algebraic equations $Ax = b$, where A is a given n by n (unsymmetric) nonsingular matrix and b an n-dimensional vector. Given an initial approximation x_0, one approach to finding x is to first compute the initial residual $r_0 = b - Ax_0$. Using this, derive a sequence of Krylov subspaces $\mathcal{K}_k(A, r_0) \equiv \text{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\}$, $k = 1, 2, \ldots$, in some way, and look for approximate solutions $x_k \in x_0 + \mathcal{K}_k(A, r_0)$. Various principles are used for constructing x_k which determine various Krylov subspace methods for solving $Ax = b$. Similarly, Krylov subspaces for A can be used to obtain eigenvalue approximations or to solve other problems involving A.

Krylov subspace methods are useful for solving problems involving very large sparse matrices, since these methods use these matrices only for multiplying vectors, and the resulting Krylov subspaces frequently exhibit good approximation properties. The Arnoldi method [4] is a Krylov subspace method designed for solving the eigenproblem of unsymmetric matrices. The generalized minimum residual method (GMRES) [27] uses the Arnoldi iteration and adapts it for solving the linear system $Ax = b$. GMRES can be computationally more expensive per step than some other methods; see, for example, Bi-CGSTAB [30], QMR [8, 9] for unsymmetric A, and LSQR [20, 19] for unsymmetric or even rectangular A. However, GMRES is widely

*Received by the editors November 15, 2000; accepted for publication (in revised form) October 15, 2001; published electronically February 20, 2002.
http://www.siam.org/journals/sisc/23-6/38123.html

†School of Computer Science, McGill University, Montreal, Quebec, Canada, H3A 2A7 (paige@cs.mcgill.ca). This author's work was supported by NSERC of Canada grant OGP0009236.

‡Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic (strakos@cs.cas.cz). This author's work was supported by the GA AV CR under grant A1030103. Part of this work was performed during the academic years 1998/1999 and 1999/2000 while this author was visiting Emory University, Atlanta, GA.

used for solving linear systems arising from discretization of partial differential equations, and it is also interesting to study, since it *does* in theory minimize the 2-norm of the residual $\|r_k\| = \|b - Ax_k\|$ over $x_k \in x_0 + \mathcal{K}_k(A, r_0)$ at each step. Thus, theoretical results on GMRES can, for example, provide lower bounds for the residuals of other methods using the same Krylov subspaces. GMRES is also interesting to study computationally, especially since a strong relationship has been noticed between convergence of GMRES and loss of orthogonality among the Arnoldi vectors computed via (finite precision) modified Gram–Schmidt (MGS) orthogonalization; see [11, 24]. An understanding of this will be just as important for the practical use of the Arnoldi method as it will be for GMRES itself.

This project is complicated, so we give an introduction involving simplified results. Given an initial approximation x_0 to the solution x of $Ax = b$, we form the residual

$$r_0 = b - Ax_0, \qquad \rho_0 = \|r_0\|, \qquad v_1 = r_0/\rho_0,$$

and use v_1 to initiate the Arnoldi process [4]. In theory, after k steps this produces

$$V_{k+1} = [v_1, v_2, \ldots, v_{k+1}], \quad V_{k+1}^T V_{k+1} = I_{k+1}, \quad \text{span}\{v_1, \ldots, v_{k+1}\} = \mathcal{K}_{k+1}(A, r_0).$$

At each step GMRES takes $x_k = x_0 + V_k y_k$ as the approximation to the solution x, which gives the residual $r_k = b - Ax_k$. GMRES uses that y_k which in theory minimizes the 2-norm of this residual, so

$$\|r_k\| = \min_y \|r_0 - AV_k\, y\| = \min_y \|[v_1\rho_0, AV_k] \begin{bmatrix} 1 \\ -y \end{bmatrix} \|.$$

So far this is rigorous and well known, but now we give some ideas in approximate form, so that they will be easier to follow. It is the purpose of this paper to show for the ratio of the largest to smallest singular value (condition number) $\kappa([v_1\rho_0, AV_k])$, which increases with k, and the normwise relative backward error

$$\beta(x_k) \equiv \frac{\|r_k\|}{\|b\| + \|A\| \cdot \|x_k\|}, \tag{1.1}$$

which tends to decrease with k until it is eventually zero, that with exact arithmetic we have something like the intriguing relationship

$$\beta(x_k)\, \kappa([v_1\rho_0, AV_k]) = O(1). \tag{1.2}$$

In later sections we will develop rigorous theory for the more precise version of this. There the columns of $[v_1\rho_0, AV_k]$ in $\kappa(\cdot)$ are scaled, and a certain condition must be satisfied. We will argue that the precise version probably also holds even in finite precision arithmetic and present convincing numerical examples supporting this hypothesis.

Now we explain why (1.2) is important. An efficient, and the most usual way of computing the Arnoldi vectors $v_1, v_2, \ldots, v_{k+1}$ for large sparse unsymmetric A, is to use the MGS orthogonalization. Unfortunately, in finite precision computations this leads to loss of orthogonality among these MGS Arnoldi vectors. If these MGS Arnoldi vectors are used in GMRES we have MGS GMRES. We want to show that MGS GMRES succeeds despite the loss of orthogonality among the computed MGS Arnoldi vectors. A similar hypothesis was published in [11, 24] with a justification based on the link between loss of orthogonality among the Arnoldi vectors and the

size of the GMRES relative residual. Here is how we hope to prove a significantly stronger statement in [17] by using what is essentially the result (1.2) of this paper as a fundamental intermediate step.

Following the important work [5] of Björck, and that of Walker [32], the papers [7] and [11] showed a relationship between the *finite precision* loss of orthogonality in the MGS Arnoldi vectors and the condition number $\kappa([v_1\rho_0, AV_k])$. In particular, unless A is extremely ill-conditioned (close to numericaly singular), for computed quantities

$$\|I - V_{k+1}^T V_{k+1}\|_F \leq \kappa([v_1\rho_0, AV_k])\, O(\epsilon), \quad \epsilon \text{ the computer roundoff unit} \tag{1.3}$$

(where subscript F denotes the Frobenius norm). Combining it with a finite precision version of (1.2) would show

$$\frac{\|r_k\| \cdot \|I - V_{k+1}^T V_{k+1}\|_F}{\|b\| + \|A\| \cdot \|x_k\|} \;\leq\; \beta(x_k)\, \kappa([v_1\rho_0, AV_k])\, O(\epsilon) \;=\; O(\epsilon). \tag{1.4}$$

This would imply that it is impossible to have a significant loss of orthogonality until the normwise relative backward error is very small. It could then be shown that there would be no meaningful deterioration in the rate of convergence, and significant loss of orthogonality would imply convergence and backward stability of MGS GMRES. These results would then be somewhat analogous to those shown for the Lanczos method for the symmetric eigenproblem, where significant loss of orthogonality implied that at least one eigenvalue had been found to about machine precision, and the first eigenvalues to converge did so with no meaningful deterioration in rate of convergence; see [16]. Perhaps the ideas here could be combined with some of those from [16] to prove how the MGS Arnoldi method is affected by rounding errors.

If we can prove a result like (1.4), we will be able to justify theoretically the well-known observation that, unless the matrix A is extremely ill-conditioned, MGS GMRES competes successfully in both the rate of convergence and the final accuracy with the more expensive GMRES implementation based on the Householder reflections (HH GMRES)[31]. HH GMRES was proved backward stable in [7]. That proof relied upon the fact that the Householder reflections keep the loss of orthogonality among the computed Arnoldi vectors close to the machine precision. Orthogonality among the Arnoldi vectors can be lost using MGS GMRES finite precision computations. Therefore the results from [7] could not be extended to MGS GMRES, and a different approach had to be used.

Despite its backward stability, HH GMRES is not widely used. A popular justification for this is based on the *numerical stability* versus *computational efficiency* argument: It is generally believed that HH GMRES is favorable numerically, but the cheaper MGS GMRES is accepted (sometimes with a fear of a possible unspecified loss of accuracy) as a standard for practical computations. One aim of our work is to eliminate that fear.

This paper is the third of a sequence starting with [22], which revised the fundamentals of the scaled total least squares theory. The subsequent paper [21] produced general purpose bounds we will use here and in [17]. The present paper proves theoretical results *motivated* by the abovementioned finite precision behavior of MGS GMRES but assumes *exact* arithmetic in all the proofs. Finite precision analogies of the statements proven here will require detailed rounding error analyses, and these are intended for the planned paper [17]. Thus, when completed, we think the work in [21], in here, and in [17] will represent a substantial step forward in our understanding of MGS orthogonalization in Krylov subspace methods and will also lead to a full

justification for MGS GMRES computations. We also hope it will produce tools that will help in the analysis of MGS Arnoldi computations. We would like to investigate whether the MGS Arnoldi method still gives accurate approximations to eigenvalues, but we will not consider this here.

Since the results in this paper assume exact arithmetic, they are independent of any particular implementation of the GMRES method. They apply to any mathematically equivalent residual minimizing Krylov subspace method (such as the MINRES method for symmetric indefinite systems). Some mathematically equivalent variants of the GMRES method are described in [15, 25]. In most practical applications some acceleration technique must be applied to improve convergence of the basic method. For historical reasons such acceleration techniques are frequently and imprecisely called preconditioning. Assuming exact arithmetic, preconditioning of a given method is equivalent to the application of the (basic) method to some modified (preconditioned) system. In this paper we assume, with no loss of generality, that A represents the matrix and b the right-hand side of the preconditioned system. For simplicity of notation we assume that A and b are real. Reformulation to the general complex case is obvious.

The paper is organized as follows. In section 2 we will give the necessary mathematics of GMRES, while in section 3, which represents the main connection with the preceding papers [22] and [21], we will present bounds for the GMRES residual (Theorem 3.1). Section 4 will give an extreme example which shows that the assumption (3.5) required in Theorem 3.1 need not hold up until the very last step of the GMRES iteration. This is, of course, a highly contrived situation and not indicative of any realistic problem we have encountered. Section 5 will explain in more detail just why the bounds from section 3 are so important for our understanding of GMRES and related methods. We will prove Theorem 5.1, which is the precise version of (1.2) and represents the main result of this paper. Section 6 will discuss its consequences in light of possible scalings. Section 7 will display some computational results and section 8 will present concluding remarks.

In the paper we will use $\sigma_i(X)$ to denote the ith largest singular value of X, use $\kappa(X)$ to be the ratio of the largest to the smallest singular value of X, and refer to $\kappa(X)$ briefly as the condition number of X. The vector of elements i to j of a vector y will be denoted $y_{i:j}$, and e_j denotes the jth column of the unit matrix I. We will use $\|\cdot\|$ to denote the 2-norm and $\|\cdot\|_F$ to denote the Frobenius norm. Several quantities used in our bounds will depend on the iteration step k. For simplicity of notation we sometimes omit the explicit reference to the iteration step when the dependence is clear from the context and need not be stressed for any particular reason.

As explained above, this paper proves the precise version of (1.2), which is the fundamental intermediate step of the whole project, and it assumes exact arithmetic in all the proofs. However, the underlying discussion of MGS GMRES finite precision behavior motivates the whole work and affects most of the particular considerations in this paper. Though we separate the exact arithmetic results from the finite precision arithmetic discussion as much as possible, we cannot split them entirely. Scaling, for example, affects both (exact precision) bounds for the GMRES residual norm developed in this paper and finite precision bounds for loss of orthogonality in the Arnoldi process. Any discussion of scaling must consider both aspects, which are generally in conflict. When it will be helpful, we will use the word "ideally" to refer to a result that would hold using exact arithmetic, and "computationally" or "numerically" to a result of a finite precision computation.

2. The GMRES method. For a given n by n (usually unsymmetric) nonsingular matrix A and n-vector b, we wish to solve $Ax = b$. Given an initial approximation x_0 we form the residual

$$r_0 = b - Ax_0, \qquad \rho_0 = \|r_0\|, \qquad v_1 = r_0/\rho_0, \tag{2.1}$$

and use v_1 to initiate the Arnoldi process [4]. At step k this forms Av_k, orthogonalizes it against $v_1, v_2, \ldots, v_k$, and if the resulting vector is nonzero, normalizes it to give v_{k+1}, giving ideally

$$AV_k = V_{k+1}H_{k+1,k}, \quad V_{k+1}^T V_{k+1} = I_{k+1}, \quad V_{k+1} = [v_1, v_2, \ldots, v_{k+1}]. \tag{2.2}$$

Here $H_{k+1,k}$ is a $k+1$ by k upper Hessenberg matrix with elements h_{ij}, where $h_{j+1,j} \neq 0$, $j = 1, 2, \ldots, k-1$. If at any stage $h_{k+1,k} = 0$ we would stop with $AV_k = V_k H_{k,k}$. In this case all the eigenvalues of $H_{k,k}$ are clearly eigenvalues of A. When $h_{k+1,k} \neq 0$ the eigenvalues of $H_{k,k}$ are approximations to some of those of A, and this gives the Arnoldi method [4]. Computationally, we are unlikely to reach a k such that $h_{k+1,k} = 0$, and for solution of equations we stop when we assess the norm of the residual (ideally given as below in (2.7)) is small enough.

In general, at each step we take $x_k = x_0 + V_k y_k$ as our approximation to the solution x, which gives the residual

$$\begin{aligned} r_k = b - Ax_k &= r_0 - AV_k y_k = v_1\rho_0 - V_{k+1}H_{k+1,k}\, y_k \\ &= V_{k+1}(e_1\rho_0 - H_{k+1,k}\, y_k). \end{aligned} \tag{2.3}$$

GMRES seeks y_k which minimizes this residual by solving the linear least squares problem

$$\|r_k\| = \min_y \|r_0 - AV_k\, y\| = \min_y \|v_1\rho_0 - AV_k\, y\|. \tag{2.4}$$

Using (2.2) and (2.3), (2.4) can be formulated as the least squares problem with the upper Hessenberg matrix $H_{k+1,k}$

$$\|r_k\| = \min_y \|e_1\rho_0 - H_{k+1,k}\, y\|. \tag{2.5}$$

To solve (2.5) we apply orthogonal rotations (J_i being the rotation in the $i, i+1$ plane through the angle θ_i) sequentially to $H_{k+1,k}$ to bring it to upper triangular form S_k:

$$J_k \cdots J_2 J_1 H_{k+1,k} = Q_k^T H_{k+1,k} = \begin{pmatrix} S_k \\ 0 \end{pmatrix}.$$

The vectors y_k and r_k ideally then satisfy

$$S_k y_k = (Q_k^T e_1\rho_0)_{1:k}, \tag{2.6}$$

$$\begin{aligned} \|r_k\| &= |e_{k+1}^T Q_k^T e_1\rho_0| \\ &= |\xi_1\xi_2\cdots\xi_k|\, \|r_0\|, \qquad \xi_i = \sin\theta_i. \end{aligned} \tag{2.7}$$

The measure (2.7) of the (nonincreasing) residual norm is available without determining y_k, and since y_{k+1} will usually differ in every element from y_k, it would seem preferable to avoid determining y_k or x_k until we decide the residual norm (2.7) is

small enough to stop. Computationally, however, it is not clear that we can base the stopping criterion on (2.7) alone. The step from (2.4) to (2.5) requires orthogonality of the columns of V_{k+1}. However, even if orthogonality of the Arnoldi vectors computed using finite precision arithmetic is well preserved (as in HH GMRES), (2.7) will not hold for the computed quantities after the residual norm drops near the final accuracy level; see [7].

Finally, little has been published about the choice of the initial approximation x_0. In many cases $x_0 = 0$ is recommended or considered. For $x_0 = 0$ we have $r_0 = b$ and trivially $\|r_0\| \leq \|b\|$. This last condition seems very natural and should always be imposed. For a nonzero x_0 it may easily happen that $\|r_0\| > \|b\|$ (even $\gg$ for some problems), and any such x_0 is a poor initial approximation to the solution x. Hegedüs [13] suggested that a simple way around this difficulty is to rescale the initial approximation. Given a preliminary initial guess x_p, it is easy to determine the scaling parameter $\zeta_{\min}$ such that

$$\|r_0\| = \|b - Ax_p\zeta_{\min}\| = \min_{\zeta} \|b - Ax_p\zeta\|, \qquad \zeta_{\min} = \frac{b^T Ax_p}{\|Ax_p\|^2}. \tag{2.8}$$

Thus, by setting $x_0 = x_p\zeta_{\min}$ we ensure $\|r_0\| \leq \|b\|$. The extra cost for implementing this little trick is negligible; it should be used in GMRES computations whenever a nonzero x_0 is considered. For some related comments see the discussion concerning the experiments in section 7.

We point out that the previous paragraph does not mean that an arbitrary x_p with (2.8) gives a proper initial approximation x_0. Our general feeling is that, even with (2.8), a nonzero x_0 should not be used unless there is a good reason for preferring it over $x_0 = 0$. It has been observed that without such additional justification, a choice of nonzero x_0 satisfying $\|r_0\| \leq \|b\|$ can significantly slow down GMRES convergence [28].

3. Bounds for the GMRES residuals. From the previous section it is clear that GMRES can be seen as a sequence of least squares problems (2.4) involving Krylov subspaces of increasing dimensions. In [21] we considered the overdetermined approximate linear system $Bu \approx c$ and bounded the least squares (LS) residual

$$\text{LS residual} \equiv \min_{r,y} \|r\|_2 \quad \text{subject to} \quad By = c - r \tag{3.1}$$

from above and from below in terms of the scaled total least squares (STLS) distance

$$\text{STLS distance} \equiv \min_{s,E,z} \|[s, E]\|_F \quad \text{subject to} \quad (B + E)z\gamma = c\gamma - s, \tag{3.2}$$

where $\gamma > 0$ is the scaling parameter. The bounds from [21] say nothing about an iterative method, or where B or c come from, and so they are general results. In order to apply the results from [21] to GMRES we have to identify B, c, and γ with the proper quantities in GMRES. We have several choices, but as yet there is no choice which is clearly superior to the others. Therefore we will formulate the bounds in the following theorem and in section 5 in a general way. Particular scalings (γ and D_k in the theorem) will be discussed in section 6.

To obtain useful bounds for the kth step of GMRES, we consider $c = r_0 = v_1\rho_0$ and $B = B_k = AV_kD_k$, where D_k is a diagonal matrix of positive scaling coefficients ($D_k > 0$). Note that the column scaling by the diagonal matrix D_k does not change

the optimal residual r_k (see (2.4)) and

$$\|r_k\| = \min_y \|v_1\rho_0 - AV_k\, y\| = \min_{D_k^{-1}y} \|c - B_k\,(D_k^{-1}y)\|\,. \tag{3.3}$$

Clearly, for this c and B_k the solution of (3.1) is $D_k^{-1}y_k$, where y_k is the solution of the LS problem (2.4). The column scaling matrix D_k will prove useful later. Note that, by construction, B_k has full column rank.

We now give bounds on the $\|r_k\|$ in GMRES, together with bounds on an important ratio δ_k.

THEOREM 3.1. *Given a scalar $\gamma > 0$ and a positive diagonal matrix D_k, use $\sigma(\cdot)$ to denote singular values and $\|\cdot\|$ to denote 2-norms. Let the n by n nonsingular matrix A, the vectors r_0, y_k, and r_k, the scalar ρ_0, and the matrix V_k be as in the GMRES algorithm (2.1)–(2.5) using exact arithmetic, and let AV_k have rank k. Denote $B_k = AV_kD_k$, $c = v_1\rho_0$, and define*

$$\delta_k \equiv \delta_k(\gamma, D_k) \equiv \sigma_{k+1}([c\gamma, B_k])/\sigma_k(B_k) = \sigma_{k+1}([v_1\rho_0\gamma, AV_kD_k])/\sigma_k(AV_kD_k). \tag{3.4}$$

If

$$v_1 \not\perp \{\textit{left singular vector subspace of } B_k \textit{ corresponding to } \sigma_{min}(B_k)\}, \tag{3.5}$$

then $\delta_k < 1$ and

$$\begin{aligned} \mu_L \equiv \sigma_{k+1}([c\gamma, B_k])\,\{\gamma^{-2} + \|D_k^{-1}y_k\|^2\}^{\frac{1}{2}} \quad &\le \quad \|r_k\| \\ \le\ \mu_U \equiv \sigma_{k+1}([c\gamma, B_k])\,\{\gamma^{-2} + (1-\delta_k^2)^{-1}\|D_k^{-1}y_k\|^2\}^{\frac{1}{2}}, & \end{aligned} \tag{3.6}$$

$$\frac{\|r_k\|}{\left\{\gamma^{-2} + \frac{\|D_k^{-1}y_k\|^2}{1-\delta_k^2}\right\}^{\frac{1}{2}}\sigma_k(B_k)} \le \delta_k \le \frac{\|r_k\|}{\{\gamma^{-2} + \|D_k^{-1}y_k\|^2\}^{\frac{1}{2}}\sigma_k(B_k)}, \tag{3.7}$$

$$\frac{\gamma\|r_k\|}{\|[c\gamma, B_k]\|} \le \delta_k \le \frac{\gamma\|r_k\|}{\sigma_k([c\gamma, B_k])} \le \frac{\gamma\|r_k\|}{\sigma_k(B_k)} \le \frac{\gamma\|r_k\|}{\sigma_n(A)\sigma_k(D_k)}. \tag{3.8}$$

Proof. We see $c\gamma = v_1\rho_0\gamma$ and $B_k = AV_kD_k$ satisfy the conditions and assumptions of Theorem 4.1 of [21] for any $\gamma > 0$, and from (3.3) we see that r_k and $D_k^{-1}y_k$ correspond to r and y in (3.1); so the theorem holds with [21, (4.4)] giving (3.6) and its equivalent (3.7), while Corollary 6.1 of [21] gives all but the last inequality in (3.8), which holds since $V_k^HV_k = I$. ∎

Note that apart from the last inequality in (3.8) the result does not depend on orthogonality of the columns of V_k, since Theorem 4.1 of [21] requires nothing of $B = B_k = AV_kD_k$ here except that it has full column rank. The only requirement is for $\|r_k\|$ to be a minimum (see (2.4), (3.1), and (3.3)) at each step. It should also be pointed out that due to monotonicity of $\|r_k\|$ from GMRES, possible oscillations in the upper bound (3.6) can be eliminated by taking the minimum

$$\|r_k\| \ \le\ \min_{j=1,\ldots,k}\{\sigma_{j+1}([v_1\rho_0\gamma, B_j])\,\{\gamma^{-2} + (1-\delta_j^2)^{-1}\|D_j^{-1}y_j\|^2\}^{\frac{1}{2}}\}. \tag{3.9}$$

In the paper [21] we compared the bounds for the LS residual used here with other existing bounds. For example, [21, Corollary 5.1] gives

$$\begin{aligned} \gamma\|r_k\| &\le \delta_k\,\{\|c\|^2\gamma^2 + \sigma_k^2(B_k) - \sigma_{k+1}^2([c\gamma, B_k])\}^{\frac{1}{2}} \\ &\le \delta_k\,\{\|c\|^2\gamma^2 + \sigma_k^2(B_k)\}^{\frac{1}{2}}. \end{aligned} \tag{3.10}$$

As stated in [21, section 5], our bounds in (3.6) can be significantly better than those from (3.10). They are also easily applicable to the problem investigated in this paper. We will therefore not examine (3.10) and the other possible bounds which can be derived from (3.10) here.

It will be important to examine the tightness of the bounds (3.6). The following corollary is an immediate consequence of [21, Corollary 4.2].

COROLLARY 3.2. *Under the conditions and assumptions of Theorem* 3.1*, and using the notation there together with*

$$\eta \equiv \frac{\|r_k\| - \mu_L}{\|r_k\|}, \qquad \zeta \equiv \frac{\mu_U - \mu_L}{\|r_k\|}, \tag{3.11}$$

we have the following bound on η *and* ζ*:*

$$0 \le \eta \ \le \ \zeta \ \le \ \frac{\gamma^2 \|D_k^{-1} y_k\|^2}{2 + \gamma^2 \|D_k^{-1} y_k\|^2} \cdot \frac{\delta_k^2}{1 - \delta_k^2} \to 0 \quad \textit{as} \quad \gamma \to 0, \tag{3.12}$$

where the upper bound goes to zero at least as fast as $O(\gamma^4)$ *(see* (3.8)*).* ∎

The assumption (3.5) is not necessary for proving the bounds (3.6)–(3.8) and (3.12). From the proof of [21, Theorem 4.1] it is clear that these bounds require only $\delta_k < 1$, and, moreover, the lower bound in (3.6), the upper bound in (3.7) and the bounds in (3.8) also hold if $\delta_k = 1$. (The upper bound in (3.6) and the lower bound (3.7) become ∞ and 0 when $\delta_k = 1$, and so hold trivially.) Using (3.5), however, makes the theory clean and consistent. The assumption (3.5) is independent of scaling and it ensures that the bounds do not contain irrelevant quantities; see [22, Remark 4.3].

From (3.12) and (3.8) we see that small δ_k, γ, $\|r_k\|$ or $\|D_k^{-1} y_k\|/(1 - \delta_k^2)$ ensures that the bounds (3.6) are not only very tight, but very tight in a relative sense. The tightness of the bounds depends in an important way on δ_k; for $\delta_k \ll 1$ we get the strong relationship from (3.6)

$$\|r_k\| \ \approx \ \sigma_{min}([v_1 \rho_0 \gamma, AV_k D_k]) \, \{\gamma^{-2} + \|D_k^{-1} y_k\|^2\}^{\frac{1}{2}}. \tag{3.13}$$

We know $0 \le \delta_k \le 1$ from (3.4). If $\delta_k \approx 1$ the bounds in (3.6) and (3.7) become weak, so we need to see if $\delta_k \approx 1$ is possible. In the GMRES context δ_k will necessarily be small as $\|r_k\| \to 0$ (see (3.8)). Proper scaling can always ensure $\delta_k \ll 1$. (For a fixed D_k it was shown in [22, Corollary 4.1] that if (3.5) holds, then $\delta_k < 1$, δ_k increases and decreases with γ, and (3.8) shows $\gamma \to 0 \Rightarrow \delta_k \to 0$.) Using this argument, it appears at first that the disturbing case $\delta_k \approx 1$ can easily be eliminated from our discussion. It turns out, however, that this is not entirely true because the use of scaling also has disadvantages. We will see that we cannot use an arbitrarily small γ to ensure $\delta_k \ll 1$ without (potentially) damaging the tightness of the bounds for the loss of orthogonality among the Arnoldi vectors (the tightness of the scaled version of (1.3)). On the other hand, a scaling which might be appropriate from the point of view of the formulation of the main result (a scaled version of (1.2); see the following section) might at the same time increase the value of δ_k. The choice of scaling therefore represents a delicate task. Despite these subtle details, we will see that $\delta_k \approx 1$ represents a technical problem but not a serious conceptual difficulty. We will return to the detailed discussion of this point in section 6.

4. Delayed convergence of GMRES. It is possible for convergence of GMRES to be very slow and stagnate entirely even with exact arithmetic. Suppose

$$A = [e_2 \gamma_2, e_3 \gamma_3, \ldots, e_n \gamma_n, e_1 \gamma_1], \qquad b = e_1 \|b\|, \qquad x_0 = 0,$$

for some $\gamma_i \neq 0$, $i = 1, \ldots, n$; then in (2.1) and (2.2) for $k < n$

$$V_{k+1} = [e_1, e_2, \ldots, e_{k+1}], \qquad H_{k+1,k} = [e_2\gamma_2, e_3\gamma_3, \ldots, e_{k+1}\gamma_{k+1}],$$

and in (2.3) and (2.5)

$$y_k = 0, \qquad x_k = 0, \qquad r_k = r_0, \qquad k = 1, 2, \ldots, n-1;$$

so any convergence at all is delayed until the solution is obtained at step $k = n$. Here we have $v_1 = e_1 \perp \mathcal{R}(AV_k)$ for $k < n$, so (3.5) does not hold and $\delta_k = 1$ for $k = 1, 2, \ldots, n-1$. In fact (3.6) degenerates to $\|r_k\| = \|r_0\|$ for $k < n$.

5. Backward error theorem. Now we show why we consider the bounds from Theorem 3.1 to be so important. This provides the scaled versions of (1.2)–(1.4). Remember that the scaled equivalents of the finite precision results (1.3)–(1.4) are only for motivation here, and the full proofs of these will be left to [17].

As noticed in [32] and used in [7] (see also [3]), the Arnoldi process (2.2) with (2.1) ideally gives the QR factorization of $[r_0, AV_k]$, since on defining upper triangular $R_{k+1} \equiv [e_1\rho_0, H_{k+1,k}]$ we see

$$[r_0, AV_k] = V_{k+1}[e_1\rho_0, H_{k+1,k}] = V_{k+1}R_{k+1}, \qquad V_{k+1}^T V_{k+1} = I_{k+1}. \tag{5.1}$$

By comparing this with (2.1) and (2.2), we see we may now refer to (5.1) as the Arnoldi process.

If the orthogonalization in (2.2) is carried out by the MGS technique, then it is straightforward to show that this MGS Arnoldi process provides V_{k+1} and R_{k+1}, which are *computationally* identical to those produced by the QR factorization of $[r_0, \widetilde{AV}_k]$ by MGS. Here, $\widetilde{AV}_k$ indicates that the multiplications Av_j, $j = 1, \ldots, k$, are computed numerically. A parallel statement holds when classical Gram–Schmidt orthogonalization is used in (2.2).

With a computer using finite precision with unit roundoff ϵ, the computed vectors $v_1, v_2, \ldots$ tend to lose orthogonality. It was shown by Björck [5] that using MGS in the QR factorization $C = QR$ computationally leads to Q such that

$$\|I - Q^T Q\|_F \leq \kappa(C)\, O(\epsilon).$$

(For convenience in numerical experiments we use the Frobenius norm.)

Thus from the discussion following (5.1), for the finite precision version of (2.2) using MGS we have (see (1.3))

$$\|I - V_{k+1}^T V_{k+1}\|_F \leq \kappa([v_1\rho_0, AV_k])\, O(\epsilon). \tag{5.2}$$

Note that $\kappa([v_1\rho_0, AV_k])$ is used here instead of $\kappa([r_0, \widetilde{AV}_k])$. Using $\kappa([v_1\rho_0, AV_k])$ simplifies further considerations; the difference between $\kappa([v_1\rho_0, AV_k])$ and $\kappa([r_0, \widetilde{AV}_k])$ is absorbed in the multiplicative factor $O(\epsilon)$. For the detailed justification see [7] and [11].

When MGS is used with exact arithmetic in (5.1), the resulting matrix V_{k+1} is invariant with respect to the column scaling in $[v_1\rho_0\gamma, AV_kD_k]$, where $\gamma > 0$ and D_k is a positive diagonal k by k matrix. It appears that, ignoring a small additional error of $O(\epsilon)$, the matrix V_{k+1} resulting from the *finite precision* MGS Arnoldi process (5.1) is invariant with respect to positive column scaling. This important result was noticed in [11, p. 711], and was partially exploited there. It can be justified by

the following argument (which is a variant of the argument attributed to Bauer; see [33, pp. 129–130]). If the scaling factors are always powers of the base of the floating point arithmetic (powers of 2 for the IEEE FP arithmetic), then the resulting V_{k+1} computed in finite precision arithmetic using the MGS Arnoldi process (5.1) will be exactly the same as the V_{k+1} computed in finite precision arithmetic using the same MGS Arnoldi process for the scaled data $[r_0\gamma, AV_kD_k]$. If the scaling factors are not powers of the base of the floating point arithmetic, then there will be additional rounding errors proportional to unit roundoff ϵ. Apparently no formal proof of the last part has been given, so we hope to include one in [17].

If all the above are true, the loss of orthogonality among the MGS Arnoldi vectors computed via (5.1) with a computer using finite precision arithmetic with unit roundoff ϵ is bounded by

$$\|I - V_{k+1}^T V_{k+1}\|_F \le \kappa([v_1\rho_0\gamma, AV_kD_k])\, O(\epsilon) \tag{5.3}$$

for all $\gamma > 0$ and positive diagonal k by k matrices D_k. One possibility is to scale the columns of $[v_1\rho_0\gamma, AV_kD_k]$ so they have unit length. That is, take

$$\gamma = \rho_0^{-1}, \qquad D_k = \operatorname{diag}(\|Av_1\|^{-1}, \ldots, \|Av_k\|^{-1}) \equiv \operatorname{diag}(\|Av_j\|^{-1}). \tag{5.4}$$

The corresponding condition number and the bound (5.3) would then be no more than a factor $\sqrt{k+1}$ away from its minimum (see [29]), so this is nearly optimal scaling. Other convenient choices will be discussed in the next section. Extensive experimental evidence suggests that for the nearly optimal scaling (5.4), the bound (5.3) is tight, and usually

$$\|I - V_{k+1}^T V_{k+1}\|_F \approx \kappa([v_1\rho_0\gamma, AV_kD_k])\, O(\epsilon). \tag{5.5}$$

It was observed that when MGS was used in (2.2), leading to the MGS GMRES method (2.1)–(2.6), loss of orthogonality in V_{k+1} was accompanied by a small relative residual norm $\|r_k\|/\rho_0$; see [11]. That is, significant loss of orthogonality in MGS GMRES apparently did not occur before convergence measured by $\|r_k\|/\rho_0$ occurred. This fortuitous behavior was analyzed numerically in [11] and a partial explanation was offered there. A much stronger and more complete theoretical explanation of the observed behavior can be derived from the bounds (3.6)–(3.8). As a first step, $\|r_k\|/\rho_0$ must be replaced by a more appropriate convergence characteristic.

We will use the terminology (such as *normwise*) and results reported in [14, section 7.1]. The *backward error* for x_k as an approximate solution for $Ax = b$ is a measure of the amounts by which A and b have to be perturbed so that x_k is the exact solution of the perturbed system $(A + \Delta A)x_k = b + \Delta b$. The *normwise relative backward error* of x_k defined by

$$\beta(x_k) \equiv \min_{\beta, \Delta A, \Delta b} \{\beta \ : \ (A + \Delta A)x_k = b + \Delta b,\ \|\Delta A\| \le \beta\|A\|,\ \|\Delta b\| \le \beta\|b\|\}$$

was shown by Rigal and Gaches [23] (see [14, Theorem 7.1, p. 132]), to satisfy

$$\beta(x_k) = \frac{\|r_k\|}{\|b\| + \|A\| \cdot \|x_k\|} = \frac{\|\Delta A_{\min}\|}{\|A\|} = \frac{\|\Delta b_{\min}\|}{\|b\|}. \tag{5.6}$$

We strongly believe that if no other (more relevant and more sophisticated) criterion is available (such as in [1]), this relative backward error should always be

preferred to the (relative) residual norm $\|r_k\|/\|r_0\| = \|r_k\|/\rho_0$ in (2.1) when measuring convergence of iterative methods. In practice $\|A\|$ has to be replaced by its approximation—when available—or simply by the Frobenius norm of A. The theoretical reasons for preferring the relative backward error are well known; see, for example, [2] and [14]. We will add some more practical arguments in section 7. In particular the residual norm can be very misleading and easily misinterpreted. It is surprising and somewhat alarming that $\|r_k\|/\rho_0$ remains in use as the main (and usually the only) indicator of convergence of iterative processes. This statement applies to the majority of computational results published by numerical analysts. Our results will put a new emphasis on the importance of the backward error. For GMRES and the other residual minimizing methods, this raises a key question. If the residual norm is somewhat in doubt as a measure of convergence, how does this affect the position of the minimal residual principle as one of the main principles on which practical Krylov subspace methods are based? The answer needs work, and its further discussion is beyond the scope of this paper. However, we do not expect that the position of the minimal residual principle will be considerably shaken by such an analysis; rather we think it will be reaffirmed. It seems that GMRES, though based on the minimal residual principle, also produces a very good (nearly optimal) backward error.

We will now describe our main observation. This illustrates and supports the main goal of our work on MGS GMRES, which is to prove a scaled version of (1.4). Consider a plot with two lines obtained from the MGS GMRES finite precision computation. One line represents the relative backward error $\|r_k\|/(\|b\| + \|A\| \cdot \|x_k\|)$ and the other the loss of orthogonality $\|I - V_{k+1}^T V_{k+1}\|_F$ (both plotted on the same logarithmic scale) as a function of the iteration step k. We have observed that these two lines are always very nearly reflections of each other through the horizontal line defined by their intersection. For a clear example of this, see the dashed lines in Figure 7.1. In other words, *in finite precision MGS GMRES computations, the product of the normwise relative backward error and the loss of orthogonality is (as a function of the iteration step) almost constant and equal to the order of the machine precision* ϵ. The goal of this paper and [17] is to present a theoretical proof of this observed fact, and its fundamental consequences, which are that orthogonality among the computed MGS Arnoldi vectors is effectively maintained until convergence and total loss of orthogonality implies convergence of the normwise relative backward error to $O(\epsilon)$, which is equivalent to (normwise) backward stability of MGS GMRES.

Using the results presented in [21] the main ideas are simple and elegant. The proof itself (as yet incomplete) is, however, technical and tedious. Therefore in this paper we restrict ourselves to proving and discussing exact arithmetic results about the product of the normwise relative backward error and the condition number $\kappa([v_1\rho_0\gamma, AV_kD_k])$; with finite precision arithmetic this condition number controls the numerical loss of orthogonality via (5.5). A detailed rounding error analysis, together with the results relating the genuine loss of orthogonality $\|I - V_{k+1}^T V_{k+1}\|_F$ to the relative backward error, is intended for [17].

In the following theorem the product of the normwise relative backward error of GMRES and the condition number of the scaled matrix $[v_1\rho_0\gamma, AV_kD_k]$ is bounded from below and from above. Note that the theorem assumes exact arithmetic and therefore the result holds for GMRES in general. The theorem is formulated for any $\gamma > 0$ and any positive diagonal D_k; bounds corresponding to the specific choices of γ and D_k will be given in section 6.

THEOREM 5.1. *Under the conditions and assumptions of Theorem* 3.1*, and using the notation there, let* $\sigma_1 \equiv \sigma_1([v_1\rho_0\gamma, AV_kD_k]) = \|[v_1\rho_0\gamma, AV_kD_k]\|$, $\kappa_k \equiv \kappa([v_1\rho_0\gamma, AV_kD_k])$. *Then*

$$\begin{aligned} \frac{\sigma_1}{\sqrt{2}} \cdot \frac{\{\gamma^{-2} + \|D_k^{-1}y_k\|^2\}^{\frac{1}{2}}}{\{\|b\|^2 + \|A\|^2\|x_k\|^2\}^{\frac{1}{2}}} &\le \sigma_1 \frac{\{\gamma^{-2} + \|D_k^{-1}y_k\|^2\}^{\frac{1}{2}}}{\|b\| + \|A\| \cdot \|x_k\|} \\ &\le \kappa_k \frac{\|r_k\|}{\|b\| + \|A\| \cdot \|x_k\|} \\ \le \sigma_1 \frac{\{\gamma^{-2} + (1-\delta_k^2)^{-1}\|D_k^{-1}y_k\|^2\}^{\frac{1}{2}}}{\|b\| + \|A\| \cdot \|x_k\|} &\le \sigma_1 \frac{\{\gamma^{-2} + (1-\delta_k^2)^{-1}\|D_k^{-1}y_k\|^2\}^{\frac{1}{2}}}{\{\|b\|^2 + \|A\|^2\|x_k\|^2\}^{\frac{1}{2}}}. \end{aligned} \tag{5.7}$$

Proof. The tighter lower and upper bounds follow immediately from (3.6) in Theorem 3.1. However,

$$\frac{1}{\sqrt{2}} \le f\left(\frac{\|A\| \cdot \|x_k\|}{\|b\|}\right) = \frac{\{\|b\|^2 + \|A\|^2\|x_k\|^2\}^{\frac{1}{2}}}{\|b\| + \|A\| \cdot \|x_k\|} \le 1, \tag{5.8}$$

since for $\omega \ge 0$, $f(\omega) \equiv (1+\omega^2)^{\frac{1}{2}}/(1+\omega)$ satisfies $f(0) = 1$, $f(\omega) < 1$ for $\omega > 0$, $f(\omega) \to 1$ for $\omega \to \infty$, and $f(\omega)$ has for $\omega > 0$ a single minimum $f(1) = \sqrt{2}/2$. This gives the weaker lower and upper bounds in (5.7). ∎

Note that the ratio of the tighter upper and lower bounds is (exactly as in (3.6))

$$\nu \equiv \frac{\{\gamma^{-2} + (1-\delta_k^2)^{-1}\|D_k^{-1}y_k\|^2\}^{\frac{1}{2}}}{\{\gamma^{-2} + \|D_k^{-1}y_k\|^2\}^{\frac{1}{2}}} \tag{5.9}$$

and the corresponding ratio of the weaker bounds is $\sqrt{2}\,\nu$. We will prefer the weaker bounds because they are convenient for the discussion of the particular scalings in the next section, and the factor $\sqrt{2}$ does not affect our considerations.

6. Scaling choices. There is no easy preference for the choice of scaling, since we have to consider several aspects that are unfortunately in conflict.

As described before, our ultimate goal is to relate the loss of orthogonality among the Arnoldi vectors to the convergence of MGS GMRES measured by the normwise relative backward error by obtaining a scaled version of (1.4). Considering (5.3) it seems that the role of scaling is to minimize $\kappa([v_1\rho_0\gamma, AV_kD_k])$, and the nearly optimal scaling (5.4) seems to be the right choice. Scaling decreasing $\kappa([v_1\rho_0\gamma, AV_kD_k])$ may, however, increase the value of $\delta(\gamma, D_k)$ and therefore act against the tightness of the bounds in Theorem 5.1; see (3.8) and (3.12). While decreasing γ decreases δ_k [22, Corollary 4.1], decreasing entries in D_k increase the upper bounds in (3.8) and potentially also δ_k. In order to describe this in more detail we denote, for the moment, $\vartheta \equiv (\sigma_k(D_k))^{-1}, D_k' \equiv \vartheta D_k, \sigma_k(D_k') = 1$. Now $\vartheta\sigma_1([v_1\rho_0\gamma, AV_kD_k]) = \sigma_1([v_1\rho_0\gamma\vartheta, AV_kD_k'])$, $\kappa([v_1\rho_0\gamma, AV_kD_k]) = \kappa([v_1\rho_0\gamma\vartheta, AV_kD_k'])$, and for δ_k in (3.4)

$$\begin{aligned} \delta_k \equiv \delta_k(\gamma, D_k) &= \frac{\sigma_{k+1}([v_1\rho_0\gamma, AV_kD_k])}{\sigma_k(AV_kD_k)} \\ &= \frac{\sigma_{k+1}([v_1\rho_0\gamma\vartheta, AV_kD_k'])}{\sigma_k(AV_kD_k')} = \delta_k(\gamma\vartheta, D_k'). \end{aligned} \tag{6.1}$$

This shows the bounds in Theorem 5.1 rescale trivially, giving the same results for the scaling γ, $D_k = \vartheta^{-1}D_k'$, as for the scaling $\gamma\vartheta$, D_k'. It is clear from [22, Corollary

4.1] that, for a fixed D_k', $\delta_k(\gamma\vartheta, D_k')$ increases monotonically with $\gamma\vartheta$, and in some circumstances it can be close to unity. (We assume that the assumptions of Theorem 3.1 hold and therefore $\delta_k < 1$ always.) It follows that if ϑ is very large (resulting in large $\gamma\vartheta$), then $\delta_k(\gamma\vartheta, D_k')$ can be close to unity. This negatively affects the tightness of the bounds in Theorem 5.1. Consequently, the near optimal tightness in (5.3) might be achieved at the cost of weakening (5.7). Similarly, weakening (5.3) may result in a tighter (5.7).

Please notice that varying ϑ (for a fixed D_k') has, due to (6.1), the same effect on $\delta_k(\gamma, D_k) = \delta_k(\gamma, \vartheta^{-1}D_k') = \delta_k(\gamma\vartheta, D_k')$ as varying the scaling parameter γ. It therefore need not be considered here.

To study further the effects of scaling, we will discuss three specific cases: no scaling ($\gamma = 1$, $D_k = I$), the nearly optimal column scaling $\gamma = \rho_0^{-1}$, $D_k = \mathrm{diag}(\|Av_j\|^{-1})$, and the norm scaling $\gamma = \|b\|^{-1}$, $D_k = \|A\|^{-1}I$. We will consider only the weaker bounds given by Theorem 5.1.

PROPOSITION 6.1. *Under the conditions and assumptions of Theorem* 3.1 *and using the notation of Theorem* 5.1*, we have the following bounds:*
With no scaling ($\gamma = 1$, $D_k = I$) we have $\delta_k \equiv \delta_k(1, I)$, $\sigma_1 \equiv \sigma_1([r_0, AV_k])$, $\kappa_k \equiv \kappa([r_0, AV_k])$, and the weaker bounds from (5.7) *give*

$$\begin{aligned} \chi_{L1} \;\equiv\; \frac{\sigma_1}{\sqrt{2}} \cdot \frac{\{1+\|y_k\|^2\}^{\frac{1}{2}}}{\{\|b\|^2+\|A\|^2\|x_k\|^2\}^{\frac{1}{2}}} \;&\le\; \kappa_k \, \frac{\|r_k\|}{\|b\|+\|A\|\cdot\|x_k\|} \\ &\le\; \sigma_1 \, \frac{\{1+(1-\delta_k^2)^{-1}\|y_k\|^2\}^{\frac{1}{2}}}{\{\|b\|^2+\|A\|^2\|x_k\|^2\}^{\frac{1}{2}}} \;\equiv\; \chi_{U1}. \end{aligned} \tag{6.2}$$

The nearly optimal column scaling $\gamma = \rho_0^{-1}$, $D_k = \mathrm{diag}(\|Av_j\|^{-1})$ gives $\delta_k \equiv \delta_k(\rho_0^{-1}, D_k)$, $\sigma_1 \equiv \sigma_1([v_1, AV_kD_k])$, $\kappa_k \equiv \kappa([v_1, AV_kD_k])$, and

$$\begin{aligned} \chi_{L2} \;\equiv\; \frac{\sigma_1}{\sqrt{2}} \cdot \frac{\{\rho_0^2+\|D_k^{-1}y_k\|^2\}^{\frac{1}{2}}}{\{\|b\|^2+\|A\|^2\|x_k\|^2\}^{\frac{1}{2}}} \;&\le\; \kappa_k \, \frac{\|r_k\|}{\|b\|+\|A\|\cdot\|x_k\|} \\ &\le\; \sigma_1 \, \frac{\{\rho_0^2+(1-\delta_k^2)^{-1}\|D_k^{-1}y_k\|^2\}^{\frac{1}{2}}}{\{\|b\|^2+\|A\|^2\|x_k\|^2\}^{\frac{1}{2}}} \;\equiv\; \chi_{U2}. \end{aligned} \tag{6.3}$$

Finally, the scaling $\gamma = \|b\|^{-1}$, $D_k = \|A\|^{-1}I$ gives

$$\delta_k \equiv \delta_k\left(\frac{\|A\|}{\|b\|}, I\right), \quad \sigma_1 \equiv \sigma_1\left(\left[\frac{v_1\rho_0}{\|b\|}, \frac{AV_k}{\|A\|}\right]\right), \quad \kappa_k \equiv \kappa\left(\left[\frac{v_1\rho_0}{\|b\|}, \frac{AV_k}{\|A\|}\right]\right), \tag{6.4}$$

$$\begin{aligned} \chi_{L3} \equiv \frac{\sigma_1}{\sqrt{2}} \cdot \frac{\{\|b\|^2+\|A\|^2\|y_k\|^2\}^{\frac{1}{2}}}{\{\|b\|^2+\|A\|^2\|x_k\|^2\}^{\frac{1}{2}}} \;&\le\; \kappa_k \, \frac{\|r_k\|}{\|b\|+\|A\|\cdot\|x_k\|} \\ &\le\; \sigma_1 \, \frac{\{\|b\|^2+(1-\delta_k^2)^{-1}\|A\|^2\|y_k\|^2\}^{\frac{1}{2}}}{\{\|b\|^2+\|A\|^2\|x_k\|^2\}^{\frac{1}{2}}} \;\equiv\; \chi_{U3}. \end{aligned} \tag{6.5}$$

∎

Throughout this discussion of GMRES in exact arithmetic, we could have replaced $\|y_k\|_2$ by $\|x_k - x_0\|_2$, since $x_k = x_0 + V_k y_k$. However, we chose not to do this in order that the results be relevant to the finite precision case as well, where V_k may lose orthogonality. The exception which will follow will allow us to write the result (6.5)

in a very simple form. Consider for the moment $x_0 = 0$. Then (ideally) $\|x_k\| = \|y_k\|$ and when $\delta_k(\|b\|^{-1}\|A\|, I) \ll 1$, (6.5) reduces to (with the definitions in (6.4))

$$\frac{\sigma_1}{\sqrt{2}} \;\le\; \kappa_k \,\frac{\|r_k\|}{\|b\| + \|A\| \cdot \|x_k\|} \;\lesssim\; \sigma_1 \,. \tag{6.6}$$

For the scaling in (6.3), each of the $k+1$ columns of $[v_1, AV_kD_k]$ has a 2-norm of 1, so

$$1 \le \sigma_1 \equiv \|[v_1\rho_0\gamma, AV_kD_k]\| \le \sqrt{k+1}\,. \tag{6.7}$$

For the scaling in (6.4) and x_0 chosen from (2.8), the 2-norm of each column of $[v_1\rho_0\|b\|^{-1}, AV_k\|A\|^{-1}]$ is bounded above by 1, so the upper bound in (6.7) holds. If we assume $x_0 = 0$ as well, then $v_1\rho_0 = r_0 = b$, so the first column has the 2-norm of 1, and all of (6.7) holds. However, generalizing a suggestion by Ruiz [26], for any matrix partitioned into two submatrices

$$\begin{aligned}
\max\{\|W\|, \|Z\|\} \;\le\; \|[W, Z]\| &= \max_{\|w\|^2+\|z\|^2=1} \|Ww + Zz\| \\
&\le \max_{\|w\|^2+\|z\|^2=1} \{\|W\|\|w\| + \|Z\|\|z\|\} \\
&= \max_{\|w\|^2+\|z\|^2=1} (\|W\|, \|Z\|)(\|w\|, \|z\|)^T \\
&\le \{\|W\|^2 + \|Z\|^2\}^{\frac{1}{2}}\,.
\end{aligned} \tag{6.8}$$

Applying these bounds to $\|[v_1\rho_0\gamma, AV_kD_k]\|$ with scaling in (6.4) and $x_0 = 0$ gives

$$1 \;\le\; \sigma_1 \equiv \|[v_1\rho_0\gamma, AV_kD_k]\| \;\le\; \sqrt{2}\,. \tag{6.9}$$

Thus for the scaling in (6.4) with $x_0 = 0$, both (6.6) and (6.9) hold, which gives the simplest form of our main result—the correct version of (1.2).

PROPOSITION 6.2. *Under the conditions and assumptions of Theorem* 3.1*, using the notation of Theorem* 5.1 *and assuming that* $\delta_k \ll 1$*, we have with* $\gamma = \|b\|^{-1}$*,* $D_k = \|A\|^{-1}$*, and* $x_0 = 0$*:*

$$\frac{1}{\sqrt{2}} \le \kappa_k \,\frac{\|r_k\|}{\|b\| + \|A\| \cdot \|x_k\|} \;\lesssim\; \sqrt{2}, \tag{6.10}$$

which can be written as

$$\kappa_k \,\frac{\|r_k\|}{\|b\| + \|A\| \cdot \|x_k\|} \;=\; O(1)\,. \qquad \square \tag{6.11}$$

The last results hold even for nonzero x_0 whenever $\|y_k\| = O(\|x_k\|)$. (Numerical experiments suggest that for well chosen x_0 (see (2.8)), this assumption is very realistic.) Because of the simple form of (6.10) we call the scaling in (6.4) scaling for elegance. In the experiments in section 7 we will compare the bounds χ_{L1}, χ_{L2}, and χ_{L3}, and χ_{U1}, χ_{U2}, and χ_{U3}, together with the effects of the particular scalings on the tightness of (5.3).

In an iterative solution of equations with nonsingular A we expect $\|r_k\| \to 0$ so $\delta_k \to 0$ (see (3.8)), and $\delta_k \ll 1$ is necessary eventually. For a general (finite dimensional) problem this seems trivial, but there are extreme possibilities: δ_k may,

for example, be close to unity (or $\delta_k = 1$ in some special cases) for $k = 1, 2, \ldots, n-1$ and $\delta_n = 0$ (see section 4). However, in many practical problems there exists a k_0 much smaller than n such that $\delta_k \ll 1$ for $k = k_0, k_0 + 1, \ldots$, and in Corollary 3.2

$$0 \leq \eta = \eta(k) \approx 0 \tag{6.12}$$

holds for $k > k_0$. In other problems $\delta_k \ll 1$ for a number of steps, but then suddenly δ_k appears very close to unity. In these cases the smoothed upper bound (3.9) might be considered—our experiments suggest it is usually close to $\|r_k\|$ for all iteration steps k. Typical examples are shown in section 7.

Under the assumptions of Theorem 3.1 $\delta_k = \delta_k(\gamma, D_k)$ is bounded away from unity for all positive γ, D_k [21, Theorem 3.1] and, unless the projection of r_0 onto the left singular vector subspace corresponding to $\sigma_{\min}(AV_kD_k)$ of the matrix AV_kD_k is very small compared to $\|r_k\|$, we can expect that the bounds for $\|r_k\|$ given by Theorem 3.1 are sufficiently tight. Still, the choices of D_k having small elements on the diagonal seems very unfortunate because they (potentially) increase the value of δ_k. Fortunately, as shown in section 7, in practical computations small diagonal elements in D_k have a much less dramatic effect on δ_k, and on the tightness of the bounds (3.6) for $\|r_k\|$, than the weaker upper bound in (3.8) would suggest. Moreover, we will show that in our experiments the scaling $\gamma = \|b\|^{-1}$, $D_k = \|A\|^{-1}I$ (which provided the result (6.10)) indeed relaxed the tightness of the bounds (3.6) for $\|r_k\|$, but the resulting relaxed bounds always remained very acceptable.

As mentioned above, under the assumption (3.5) δ_k is bounded away from unity, but it can still get very close to unity for some k. We have observed numerically (see also section 7) that with no scaling ($\gamma = 1$, $D_k = I$), δ_k gets close to unity quite rarely. For the other scalings considered in this paper (which are important for the formulation of our results), some δ_k may be much closer to unity, and that may also happen more often. Still, as we will now show for the example of the scaling for elegance $\gamma = \|b\|^{-1}$, $D_k = \|A\|^{-1}I$, the situation $\delta_k \approx 1$ cannot occur after GMRES has converged to a reasonable accuracy, and therefore it does not represent a serious obstacle for our theory. From (3.8) we have

$$\delta_k \leq \gamma \|r_k\| / \sigma_k(AV_kD_k), \tag{6.13}$$

which with the scaling $\gamma = \|b\|^{-1}$, $D_k = \|A\|^{-1}I$, and with $\|r_0\| \leq \|b\|$ (perhaps via (2.8)) and $V_k^T V_k = I$, gives a bound in terms of the *relative residual* $\|r_k\|/\|r_0\|$:

$$\delta_k \leq \frac{\|r_k\| \cdot \|A\|}{\|b\| \cdot \sigma_k(AV_k)} \leq \frac{\|r_k\|}{\|r_0\|}\, \kappa(A). \tag{6.14}$$

Similarly (using the same scaling) with $\|x_k\| = \|y_k\|$ we can obtain a bound in terms of the *relative backward error*

$$\delta_k \leq \frac{\|r_k\|}{\|b\| + \|A\| \cdot \|x_k\|}\, \sqrt{2}\, \kappa(A) = \beta(x_k)\sqrt{2}\, \kappa(A). \tag{6.15}$$

This follows using (3.7) and (5.8), since

$$\delta_k \leq \frac{\|r_k\|}{\{\|b\|^2 + \|A\|^2\|x_k\|^2\}^{1/2} \sigma_k(AV_k)/\|A\|} \leq \beta(x_k)\sqrt{2}\, \kappa(A).$$

Thus, when the relative residual norm drops significantly below $\kappa(A)^{-1}$, or the relative backward error drops significantly below $\{\sqrt{2}\, \kappa(A)\}^{-1}$, $\delta_k = \delta_k(\|b\|^{-1}\|A\|, I) \ll 1$ and (6.11) will hold.

For any given D_k there is a particular value of the scaling parameter γ such that $\delta(\gamma, D_k)$ is related to $\|r_k\|$ even in a more tight way than described above. For a fixed D_k define $\gamma_0^{(k)} = \gamma_0^{(k)}(D_k) = \sigma_k(AV_kD_k)/\rho_0$. With the scaling $D_k, \gamma_0^{(k)}$ the first column of the matrix $[v_1\rho_0\gamma_0^{(k)}, AV_kD_k]$ is equal to $\sigma_k(AV_kD_k)v_1$ and has the norm equal to $\sigma_k(AV_kD_k)$. Moreover, for this D_k, $\delta_k(\gamma, D_k) < 1$ for all $\gamma < \gamma_0^{(k)}$, and

$$\|r_k\| = \rho_0 \quad \text{if and only if} \quad \delta_k(\gamma_0^{(k)}, D_k) = 1\,, \tag{6.16}$$

$$\delta_k(\gamma_0^{(k)}, D_k) \le \|r_k\|/\rho_0 \le \sqrt{2}\,\delta_k(\gamma_0^{(k)}, D_k); \tag{6.17}$$

see [15, (3.11) and (3.12)]. Though with this particular scaling the relationship between $\delta_k(\gamma_0^{(k)}, D_k)$ and $\|r_k\|$ is extremely simple, it will not lead to a simple form of the main result (1.2). Also, possibly small value $\gamma_0^{(k)}$ may inconveniently relax the bound (5.3) and significantly complicate the analysis left to [17]. Therefore we have not used this scaling in our paper.

The approach here might also be useful for Krylov subspace methods which minimize other norms, such as minimum error methods, as we now show. Let $V_k = [v_1, \ldots, v_k]$ be generated in some way, and $r_0 = b - Ax_0$, $\rho_0 = \|r_0\|$, $v_1 = r_0/\rho_0$, $x_k = x_0 + V_k y_k$, $r_k = b - Ax_k = r_0 - AV_k y_k$, with A nonsingular. Consider, for example, a method that minimizes $\|A^{-1}r_k\| = \|x - x_k\|$ at each step (so y_k will differ from that in GMRES). Then taking $[c, B] = A^{-1}[v_1\rho_0, AV_k] = [(x - x_0), V_k]$ and $\gamma = \rho_0^{-1}$, Theorem 4.1 of [21] gives (with $\delta_k = \sigma_{k+1}([(x - x_0)\rho_0^{-1}, V_k])/\sigma_k(V_k)$) the bounds

$$\begin{aligned} &\sigma_{k+1}([(x - x_0)\rho_0^{-1}, V_k])\,\{\rho_0^2 + \|y_k\|^2\}^{\frac{1}{2}} \quad \le \quad \|x - x_k\| \\ &\le \sigma_{k+1}([(x - x_0)\rho_0^{-1}, V_k])\,\{\rho_0^2 + (1 - \delta_k^2)^{-1}\|y_k\|^2\}^{\frac{1}{2}}, \end{aligned} \tag{6.18}$$

so at least this theory holds for more general minimum norm methods than just GMRES. Of course, if $V_k^T V_k = I$, then $\sigma_k(V_k) = 1$. We have not studied how this might be used.

It appears that the approach can also be applied to methods which minimize some norm with respect to other Krylov subspaces, such as LSQR [20, 19] for solution of equations with unsymmetric A, or LS solutions with rectangular A. It may also be useful for methods which are not based on Krylov subspaces.

7. Experimental results. We will illustrate our theoretical results with numerical experiments. We initiated these to look for possible limitations in our theory. We wished to check the validity of our assumptions in practical computations. We also wished to find out to what extent the results developed here for exact precision GMRES would hold for quantities computed in the presence of rounding errors.

In our theory $\delta_k = \sigma_{k+1}([v_1\rho_0\gamma, AV_kD_k])/\sigma_k(AV_kD_k)$ plays an important role. Section 4 showed it is possible to have $\delta_k = 1$ for all but the last step, and in that example the residual stagnated at $\|r_0\|$ until the final step. On the other hand, $\delta_k \approx 1$ cannot in general (for scalings different from $D_k, \gamma_0^{(k)}(D_k)$, see (6.16), (6.17)) be linked with the (approximate) stagnation of GMRES; the GMRES residual norm may almost stagnate while $\delta_k \ll 1$, and it can decrease rapidly while $\delta_k \approx 1$. If $\delta_k \approx 1$, then there can be a large gap between the upper and lower bounds in (3.6). This does not negate the argument that orthogonality is effectively maintained until convergence in finite precision MGS GMRES ($\delta_k \ll 1$ is necessary eventually), but it does make us question the tightness of the bounds in (3.6).

Fortunately, experiments suggest that (3.9) is always a sufficiently good (and mostly very good) upper bound. We also found that with no scaling ($\gamma = 1$, $D_k = I$) δ_k is often reasonably below unity during the entire computation. As k increases δ_k can decrease, then increase, but it must eventually become small, for from (3.7), (3.8), (6.14), and (6.15) we see the upper bound on δ_k must decrease as $\|r_k\|$ or $\|r_k\|/(\|b\| + \|A\| \cdot \|x_k\|)$ becomes sufficiently small. However, when $\delta_k \ll 1$ from the start to the end,

$$\|r_k\| \;\approx\; \sigma_{k+1}([v_1\rho_0\gamma, AV_kD_k])\,\{\gamma^{-2} + \|D_k^{-1}y_k\|^2\}^{\frac{1}{2}}$$

throughout the computation, and the lower and upper bounds are very close. Thus we can have this unexpectedly very close relationship between $\|r_k\|$ and the smallest singular value of $[v_1\rho_0\gamma, AV_kD_k]$. An interesting experience was that even when $\delta_k \approx 1$, leading to the upper bound being significantly larger than the lower bound in (3.6), it was not always the upper bound which was weak. We frequently observed that the upper bound was tight while the lower bound was a noticeable underestimate for $\|r_k\|$. Moreover, the dependence of δ_k and the tightness of the bounds (3.6) on the scaling parameters γ, D_k was quite weak. We will illustrate these observations by presenting results of numerical experiments showing different types of behavior of δ_k. These observations could also be further studied theoretically using the results of Proposition 6.1 or some other approach, but we do not wish to go into it here.

In all experiments $b = e \equiv (1, \ldots, 1)^T$. Except for the experiment shown in Figure 7.10 (where $x_0 = \text{randn}(n, 1)$ from MATLAB 5.3), x_0 is always determined from (2.8) with $x_p = \text{randn}(n, 1)$ from MATLAB 5.3. These choices of x_0 and x_p are worth a comment. We wish to illustrate our theoretical results on some nontrivial examples. The randomly generated initial vectors x_0 (or x_p in (2.8)) were chosen in our illustrations to avoid any correlation between the initial approximation and the solution. This is because we sought to illustrate cases where there were no hidden relationship that affected the computations. In practical computations, however, for the very same reason, a randomly chosen initial approximation should be avoided. Sometimes a random initial approximation x_0 is reported to give faster convergence than the other popular choice $x_0 = 0$. As we explain later, we believe that statements like that represent a serious misunderstanding caused by a superficial view of convergence. As far as genuine convergence characteristics are concerned, we argue that such statements are of no relevance.

Experiments were performed on a Silicon Graphics Origin 200 Workstation using MATLAB 5.3, $\epsilon = 1.11 \times 10^{-16}$. In all experiments matrices from the Rutherford–Boeing collection were used. Results for the matrix FS1836 with $n = 183$, $\|A\| \approx 1.2{*}10^9$, $\kappa(A) \approx 1.5{*}10^{11}$ (see Figures 7.1–7.4) illustrate improvement of the tightness of the bounds (3.6) as the residual norm drops. For the matrix WEST0132 with $n = 132$, $\|A\| \approx 3.2 * 10^5$, $\kappa(A) \approx 6.4 * 10^{11}$ (see Figures 7.5–7.8), the tightness of the bounds (3.6) oscillates during the whole computation. Results for the matrix STEAM1 with $n = 240$, $\|A\| \approx 2.2 * 10^7$, $\kappa(A) \approx 3.1 * 10^7$ (see Figures 7.9 and 7.10) represent the case when the bounds (3.6) are very tight from the start to the end.

We have given two figures for STEAM1 and four figures for the other matrices, so we now indicate what is in each figure. Figures 7.1, 7.5, 7.9, and 7.10 make the same use of lines. The dots show the norm of the *directly* computed residual divided

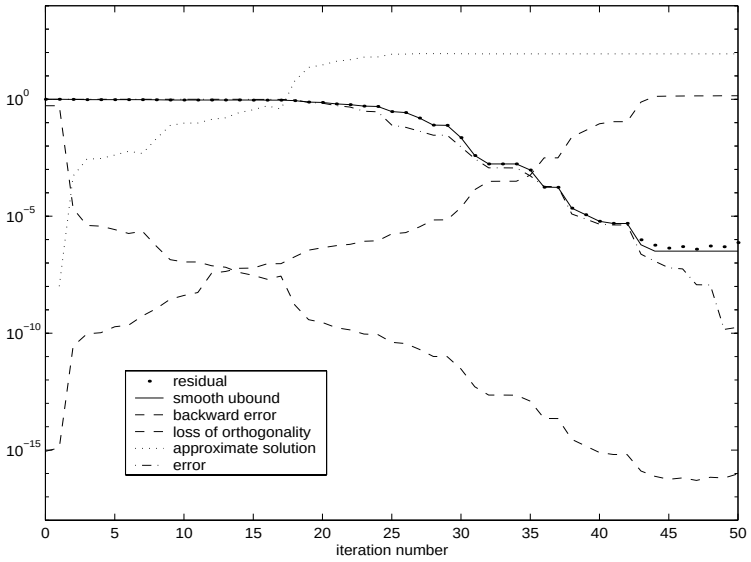

FIG. 7.1. *Norm of the directly computed relative residual (dots), the smooth upper bound (solid line), the loss of orthogonality among the Arnoldi vectors measured in the Frobenius norm (dashed line, monotonically increasing) and the normwise relative backward error (dashed line, mostly decreasing), norm of the approximate solution (dotted line), and the relative error (dashed-dotted line) for MGS GMRES applied to FS1836.*

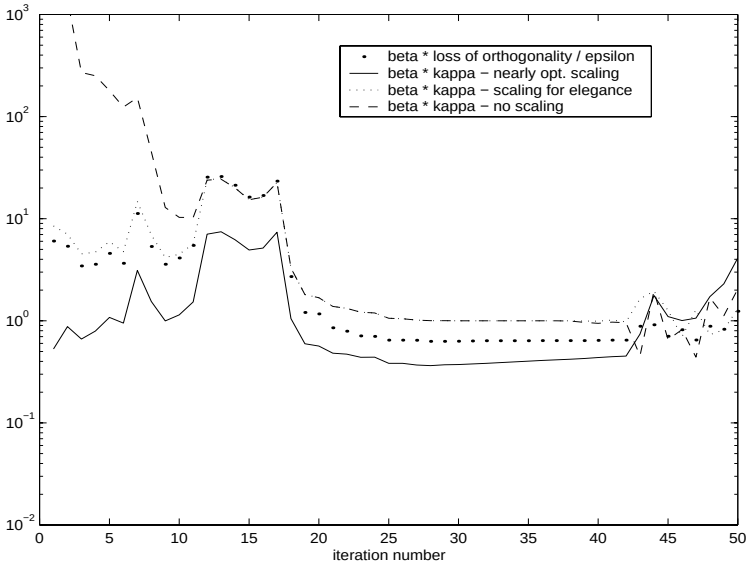

FIG. 7.2. *The product of the normwise relative backward error and the loss of orthogonality among the Arnoldi vectors measured in the Frobenius norm divided by the machine precision unit ϵ (dots), and the product of the normwise relative backward error $\beta(x_k)$ and the condition number of the matrix $[v_1\rho_0\gamma, AV_kD_k]$ for different scalings: the nearly optimal column scaling $\gamma = \rho_0^{-1}$, $D_k = \mathrm{diag}(\|Av_j\|^{-1})$ (solid line), the norm scaling (scaling for elegance) $\gamma = \|b\|^{-1}$, $D_k = \|A\|^{-1}I$ (dotted line), and no scaling $\gamma = 1$, $D_k = I$ (dashed line) for MGS GMRES applied to FS1836.*

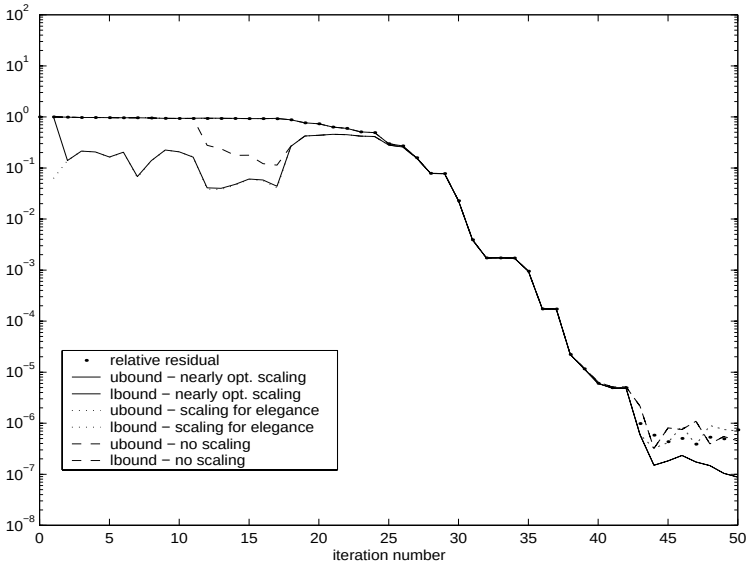

FIG. 7.3. *Norm of the directly computed relative residual (dots), and its lower and upper bounds μ_L and μ_U for different scalings: the nearly optimal column scaling (solid lines), the scaling for elegance (dotted lines), and no scaling (dashed lines) for MGS GMRES applied to FS1836. Until the orthogonality is completely lost, the upper bounds are indistinguishable from the actual quantities.*

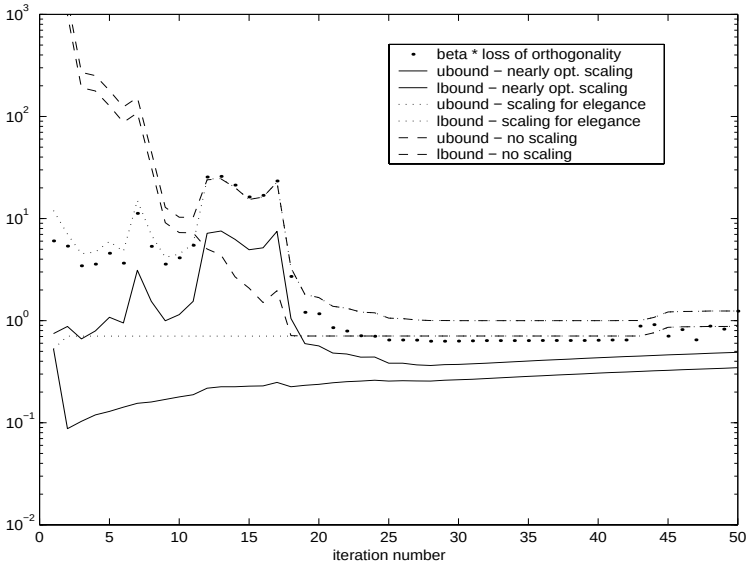

FIG. 7.4. *Product of the backward error and the loss of orthogonality among the Arnoldi vectors measured in the Frobenius norm divided by the machine precision unit ϵ (dots), and the values χ_L and χ_U for different scalings: the nearly optimal column scaling (solid lines), the scaling for elegance (dotted lines), and no scaling (dashed lines) for MGS GMRES applied to FS1836.*

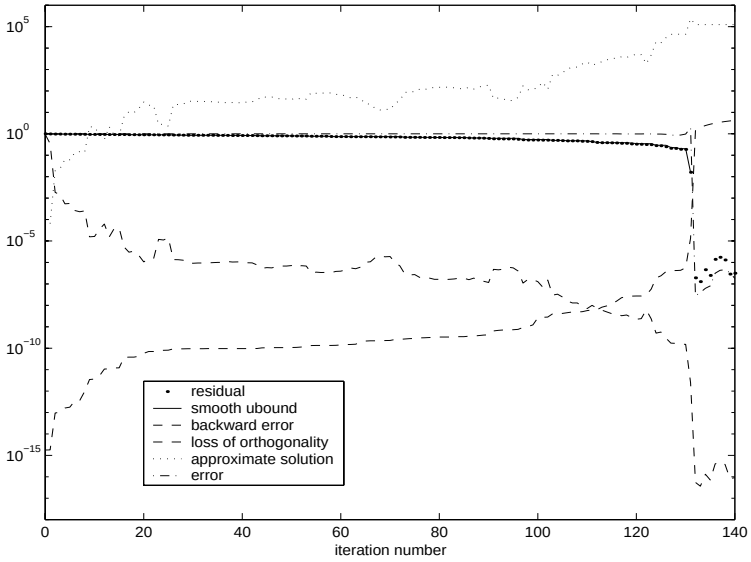

FIG. 7.5. *Norm of the directly computed relative residual (dots), the smooth upper bound (solid line), the loss of orthogonality among the Arnoldi vectors measured in the Frobenius norm (dashed line, monotonically increasing) and the normwise relative backward error (dashed line, mostly decreasing), norm of the approximate solution (dotted line), and the relative error (dashed-dotted line) for MGS GMRES applied to WEST0132.*

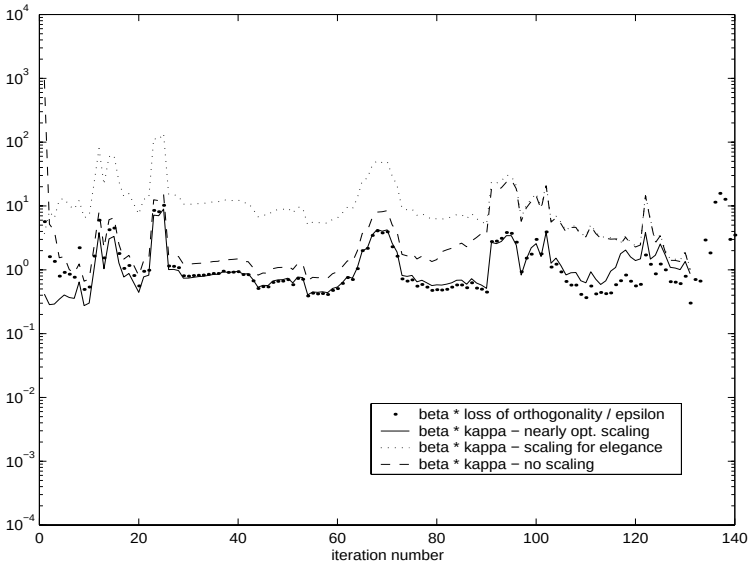

FIG. 7.6. *Product of the normwise relative backward error and the loss of orthogonality among the Arnoldi vectors measured in the Frobenius norm divided by the machine precision unit (dots), and product of the backward error and the condition number of the matrix* $[v_1\rho_0\gamma, AV_kD_k]$ *for different scalings: the nearly optimal column scaling (solid line), the scaling for elegance (dotted line), and no scaling (dashed line) for MGS GMRES applied to WEST0132.*

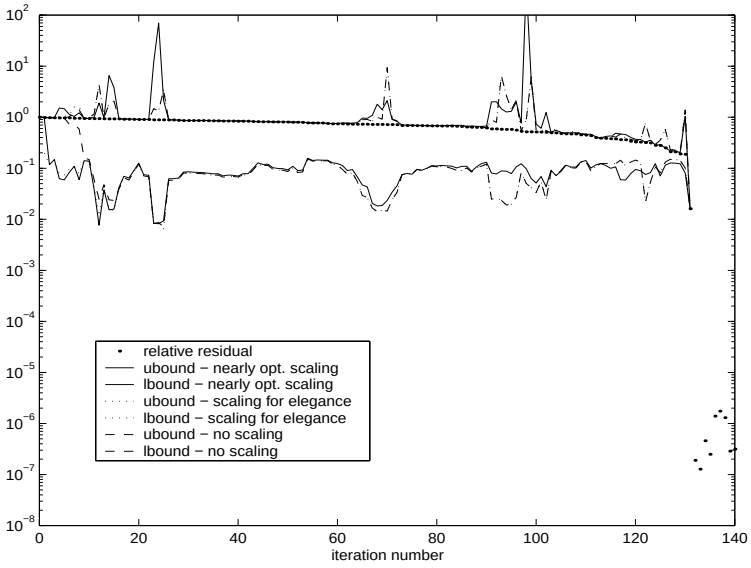

FIG. 7.7. *Norm of the directly computed relative residual(dots), and its lower and upper bounds μ_L and μ_U for different scalings: the nearly optimal column scaling (solid lines), the scaling for elegance (dotted lines), and no scaling (dashed lines) for MGS GMRES applied to WEST0132.*

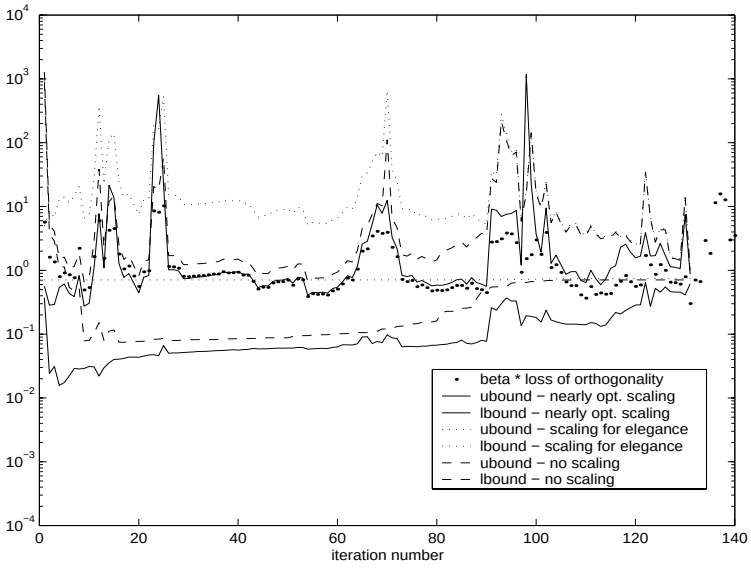

FIG. 7.8. *Product of the backward error and the loss of orthogonality among the Arnoldi vectors measured in the Frobenius norm divided by the machine precision unit ϵ (dots), and the values χ_L and χ_U for different scalings: the nearly optimal column scaling (solid lines), the scaling for elegance (dotted lines), and no scaling (dashed lines) for MGS GMRES applied to WEST0132.*

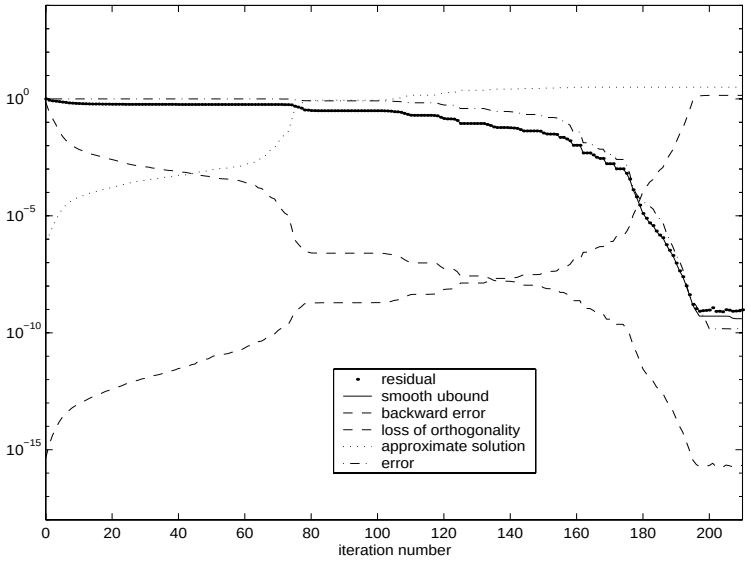

FIG. 7.9. *Norm of the directly computed relative residual (dots), the smooth upper bound (solid line), the loss of orthogonality among the Arnoldi vectors measured in the Frobenius norm (dashed line, monotonically increasing) and the normwise relative backward error (dashed line, mostly decreasing), norm of the approximate solution (dotted line), and the relative error (dashed-dotted line) for MGS GMRES applied to STEAM240.*

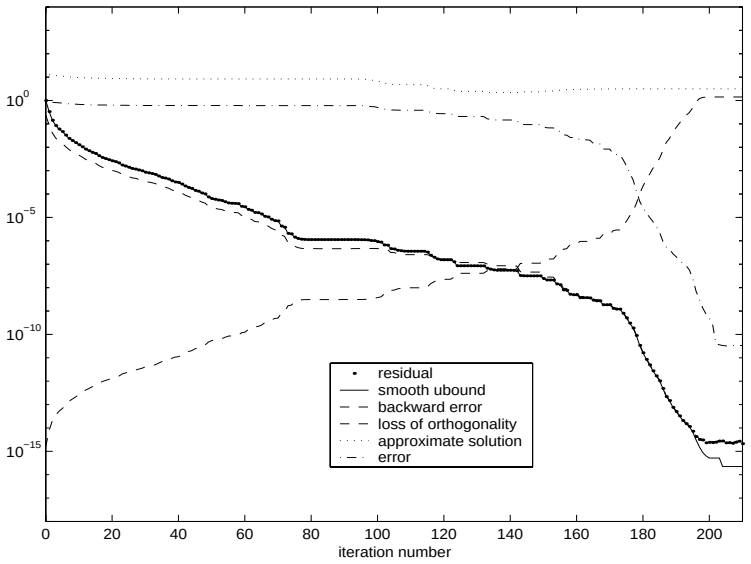

FIG. 7.10. *Norm of the directly computed relative residual (dots), the loss of orthogonality among the Arnoldi vectors measured in the Frobenius norm (dashed line, monotonically increasing) and the normwise relative backward error (dashed line, mostly decreasing), norm of the approximate solution (dotted line), and the relative error (dashed-dotted line) for MGS GMRES applied to STEAM240 with randomly chosen initial approximation* x_0.

by $\|r_0\|$, that is, $\|b - Ax_k\|/\|r_0\|$, which we call the relative residual. (We do not give the *iteratively* computed residual norm (2.7); until near convergence, it was always graphically indistinguishable from the norm of the directly computed residual.) The solid line gives the smoothed upper bound (3.9) divided by $\|r_0\|$. The dashed-dotted line gives the normalized norm of the error $\|x-x_k\|/\|x-x_0\|$; the dotted line gives the norm of the approximate solution $\|x_k\|$. The dashed lines give the loss of orthogonality among the Arnoldi vectors measured in the Frobenius norm $\|I - V_k^T V_k\|_F$ (essentially increasing), as well as the normwise relative backward error $\|r_k\|/(\|b\| + \|A\| \cdot \|x_k\|)$, which is mostly decreasing. Note the spectacular symmetry of the loss of orthogonality and the backward error in every case.

For each matrix, the remaining figures present and compare convergence characteristics, upper and lower bounds, and several quantities illustrating our theory for different scalings of the matrix $[v_1\rho_0\gamma, AV_kD_k]$. In each of Figures 7.2, 7.3, 7.4 (for FS1836) and 7.6, 7.7, 7.8 (for WEST0132) dashed lines represent results with no scaling $\gamma = 1$, $D_k = I$, solid lines the nearly optimal column scaling $\gamma = \rho_0^{-1}$, $D_k = \mathrm{diag}(\|Av_j\|^{-1})$, and dotted lines the scaling for elegance $\gamma = \|b\|^{-1}$, $D_k = \|A\|^{-1}I$.

Figures 7.2 and 7.6 are devoted to the tightness of the bound (5.3) for the loss of orthogonality among the Arnoldi vectors. The dots show the product of the normwise relative backward error and the loss of orthogonality divided by the machine precision unit $\{\|r_k\|/(\|b\| + \|A\| \cdot \|x_k\|)\} \cdot \|I - V_k^T V_k\|_F / \epsilon$, the dashed, solid, and dotted lines the product $\{\|r_k\|/(\|b\| + \|A\| \cdot \|x_k\|)\} \cdot \kappa([v_1\rho_0\gamma, AV_kD_k])$ for different scalings. The figures show that (5.5) is well justified for the nearly optimal column scaling. Replacing the actual loss of orthogonality $\|I - V_k^T V_k\|_F$ in our considerations by $\{\kappa([v_1\rho_0\gamma, AV_kD_k])\,\epsilon\}$ does not cause a significant difference (except perhaps at the beginning of the process with no scaling) even for the other scalings. Close to convergence (5.5) holds for all the scalings considered in our paper.

Figures 7.3 and 7.7 are devoted to normalized residual bounds, that is, bounds on $\|b - Ax_k\|/\|r_0\|$, which are denoted by points. The pairs of dashed, solid, and dotted lines give the upper and lower bounds μ_U and μ_L from (3.6) for different scalings. We can see that the effect of scaling on the bounds in (3.6) is quite insignificant.

Finally, Figures 7.4 and 7.8 compare the product of the normwise relative backward error and the loss of orthogonality divided by the machine precision unit, that is, $\{\|r_k\|/(\|b\| + \|A\| \cdot \|x_k\|)\} \cdot \|I - V_k^T V_k\|_F / \epsilon$ (denoted by dots), with the upper and lower bounds χ_U and χ_L from (6.2) (dashed lines), (6.3) (solid lines), and (6.5) (dotted lines). These figures reflect the possible lack of tightness of the bound for the loss of orthogonality among the Arnoldi vectors shown separately on Figures 7.2 and 7.6, as well as the lack of tightness of the bounds in (6.2)–(6.5). They demonstrate that though the results developed in this paper assume exact arithmetic, and though the form of the bounds in (6.2)–(6.5) seems a bit complicated, the simplest form of our main result (6.11) holds as convergence is approached for all our scalings and for the quantities actually computed using finite precision arithmetic.

The experiments for the figures discussed up to now (and for Figure 7.9) used $b = e$ and x_0 determined from (2.8) with $x_p = \mathrm{randn}(n, 1)$ from MATLAB 5.3. The remaining Figure 7.10 was computed for $x_0 = \mathrm{randn}(n, 1)$ from MATLAB 5.3 without using (2.8). Both Figures 7.9 and 7.10 were computed for the matrix STEAM1, and they show the same quantities as Figures 7.1 and 7.5. If we concentrate on the relative residual norm only, then it looks as if Figure 7.10 shows much better convergence (faster, and to much better accuracy) than Figure 7.9. Such a view on convergence,

though understandable, is completely wrong. We cannot give a full quantitative explanation within this paper; however, we will present an intuitive but clear argument on which such an explanation will eventually be based. By using $x_0 = \text{randn}(n, 1)$ and then computing the initial residual as $r_0 = b - Ax_0 = e - Ax_0$ (as on Figure 7.10) we correlate the initial residual strongly with the dominating parts of the operator A. (Note that all the matrices used here have some dominating components.) In all cases the norm of the resulting initial residual is large, $\|r_0\| \gg \|b\|$. At the early stage of computation this artificially created dominating information is eliminated, which creates an illusion of fast convergence. However, no real fast convergence is taking place, as you can see on the error convergence curve, and the "good final accuracy" is due to the fact that the initial residual is large. For $b = e$ and x_0 determined from (2.8) with $x_p = \text{randn}(n, 1)$ from MATLAB 5.3 we get $\|x_0\| \ll 1$ and $x_0 \approx 0$. Then r_0 contains practically no information about the dominating parts of A, the problem is difficult to solve, and the convergence is (for many steps) slow. Still, this choice (which produces results very close to those with the choice $x_0 = 0$) gives the right information about the behavior of GMRES when applied to the problem $Ax = b$, $b = e$. The illusion of fast convergence and better final accuracy for a random x_0 has evolved among some users of numerical software perhaps as a side effect of using the norm of the relative residual for displaying convergence. Our point is that the illusive role of a random x_0 can easily be revealed by using the absolute values of the residual norm for displaying convergence and by comparing the convergence curve for a random x_0 to that for the initial approximation set to zero ($x_0 = 0$). Finally, please note the correspondence of the error and the backward error when comparing Figures 7.10 and 7.9.

Now we comment on particular characteristics of each problem. For the matrix FS1836 in Figures 7.1–7.4 the value of δ_k rises until it is close to unity, stays there for a few iteration steps, then follows the descent of the residual norm. For all scalings the upper bounds μ_U (for $\|r_k\|$) are very tight until convergence, the lower bounds μ_L (for $\|r_k\|$) are weak when $\delta_k \approx 1$, but no scaling ($\gamma = 1, D_k = I$) gives a significantly tighter lower bound than the other two at the early stages of the computation (Figure 7.3). On the other hand, at the early stages of the computation the condition number of the matrix $[v_1\rho_0, AV_k]$ is for no scaling much larger than for the other scalings, which explains the difference between the dashed and the other lines on Figures 7.2 and 7.4 for k from 1 to 10. After convergence is approached, all scalings produce about the same results.

For the matrix WEST0132 (in Figures 7.5–7.8) the value of δ_k is close to unity (with some oscillations) for most iteration steps. The upper and lower bounds μ_U and μ_L differ significantly until the sharp drop of the residual. Scalings are not important. Note that despite the oscillations (we have chosen this matrix on purpose because it seems to produce challenging results; many other examples not presented here give much smoother behavior) all the lines on Figures 7.6 and 7.8 converge together as the sharp drop of the residual is approached.

For the matrix STEAM1 we omit figures analogous to Figures 7.2–7.4 for FS1836 and Figures 7.6–7.8 for WEST0132. The omitted figures would show a good agreement of the computed results with our theory; they do not offer any other information, and therefore we see no reason for extending the length of the paper by including them.

Summarizing, our experiments suggest that the equivalents of Theorems 3.1 and 5.1, of the Propositions 6.1 and 6.2 (where κ_k will be replaced by $\|I - V_{k+1}^T V_{k+1}\|_F$ and $O(1)$ by $O(\epsilon)$), hold for the *numerically computed quantities*. However, the statements

must be slightly modified to account for the effect of rounding errors, especially for the influence of the loss of orthogonality on the size of the directly computed residuals $\|b - Ax_k\|$. A rigorous proof will require further work and is intended in [17].

8. Conclusion. In Krylov subspace methods, approximate solutions to matrix problems are usually constructed by using orthogonality relations and projections. Orthogonality and projections create a mathematical elegance and beauty in this context. In the presence of rounding errors orthogonality and projection properties are gradually (and sometimes very quickly) lost. Fortunately, as was first shown for A symmetric and the Lanczos method (see, for example, [16], [10], [12]), not all the mathematical elegance need be lost with them.

This paper is devoted to GMRES, and our fundamental hypothesis is as follows. When the Arnoldi vectors are computed via the finite precision MGS process, the loss of orthogonality is related in a straightforward way to the convergence of GMRES. In particular, orthogonality among the Arnoldi vectors is effectively maintained until the normwise relative backward error converges close to the machine precision level. If we assume that the bound for the loss of orthogonality among the Arnoldi vectors is tight and (5.5) holds, then our hypothesis could be strengthened to the following: *the product of the loss of orthogonality among the Arnoldi vectors (measured in the Frobenius norm) and the normwise relative backward error is for any iteration step a small multiple of the machine precision unit.* This last statement would then imply that total loss of orthogonality among the Arnoldi vectors computed via finite precision MGS orthogonalization would mean convergence of the normwise relative backward error to machine precision level, and, consequently, it would prove backward stability of MGS GMRES. Our work can also be seen as another step on the way (probably started by Sheffield (see [6], especially the abstract and section 2, also [7])) towards the full justification of the MGS orthogonalization in competition with orthogonalization by Householder reflections for certain classes of problems.

Note that in the present paper we have not proven the finite precision versions of the statements formulated above. Our paper assumes exact arithmetic in its theoretical part and carries out groundwork for the detailed rounding error analysis of the MGS GMRES which we plan to publish in [17].

Acknowledgments. The authors would like to thank Miro Rozložník for his helpful comments. They also wish to thank Jőrg Liesen and Daniel Ruiz for many valuable suggestions which improved the content and presentation of this paper.

REFERENCES

[1] M. Arioli, *Stopping criterion for the conjugate gradient algorithm in a finite element method framework*, Numer. Math., submitted.

[2] M. Arioli, I. Duff, and D. Ruiz, *Stopping criteria for iterative solvers*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 138–144.

[3] M. Arioli and C. Fassino, *Roundoff error analysis of algorithms based on Krylov subspace methods*, BIT, 36 (1996), pp. 189–206.

[4] W. Arnoldi, *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[5] A. Björck, *Solving linear least squares problems by Gram-Schmidt orthogonalization*, BIT, 7 (1967), pp. 1–21.

[6] A. Björck and C. C. Paige, *Loss and recapture of orthogonality in the modified Gram–Schmidt algorithm*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 176–190.

[7] J. Drkošová, A. Greenbaum, M. Rozložník, and Z. Strakoš, *Numerical stability of the GMRES method*, BIT, 35 (1995), pp. 308–330.

[8] R. W. Freund and N. M. Nachtigal, *QMR: A quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.
[9] R. W. Freund and N. M. Nachtigal, *An implementation of the QMR method based on coupled two-term recurrences*, SIAM J. Sci. Comput., 15 (1994), pp. 313–337.
[10] A. Greenbaum, *Behavior of slightly perturbed Lanczos and conjugate gradient recurrences*, Linear Algebra Appl., 113 (1989), pp. 7–63.
[11] A. Greenbaum, M. Rozložník, and Z. Strakoš, *Numerical behavior of the modified Gram-Schmidt GMRES implementation*, BIT, 37 (1997), pp. 706–719.
[12] A. Greenbaum and Z. Strakos, *Predicting the behavior of finite precision Lanczos and conjugate gradient computations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 121–137.
[13] C. Hegedüs, *private communication*, 1998.
[14] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 1996.
[15] J. Liesen, M. Rozložník, and Z. Strakoš, *Least squares residuals and minimal residual methods*, SIAM J. Sci. Comput., 23 (2002), pp. 1503–1525.
[16] C. C. Paige, *Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem*, Linear Algebra Appl., 34 (1980), pp. 235–258.
[17] C. C. Paige, M. Rozložník, and Z. Strakoš, *Rounding error analysis of the modified Gram-Schmidt GMRES*, in preparation.
[18] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.
[19] C. C. Paige and M. A. Saunders, *Algorithm* 583 *LSQR: Sparse linear equations and least squares problems*, ACM Trans. Math. Software, 8 (1982), pp. 195–209.
[20] C. C. Paige and M. A. Saunders, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software, 8 (1982), pp. 43–71.
[21] C. C. Paige and Z. Strakoš, *Bounds for the least squares distance using scaled total least squares*, Numer. Math., to appear.
[22] C. C. Paige and Z. Strakoš, *Scaled total least squares fundamentals*, Numer. Math., to appear.
[23] M. Rigal and J. Gaches, *On the compatibility of a given solution with the data of a given system*, J. Assoc. Comput. Mach., 14 (1967), pp. 543–548.
[24] M. Rozložník, *Numerical Stability of the GMRES Method*, Ph.D. Thesis, Institute of Computer Science, Academy of Sciences, Prague, 1997.
[25] M. Rozložník and Z. Strakoš, *Variants of the residual minimizing Krylov space methods*, in Proceedings of the XIth Summer School "Software and Algorithms of Numerical Mathematics", I. Marek, ed., Železná Ruda, University of West Bohemia, Plzen, Czech Republic, 1996, pp. 208–225.
[26] D. Ruiz, *private communication*, 2001.
[27] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[28] Z. Strakoš, *Theory of Convergence and Effects of Finite Precision Arithmetic in Krylov Subspace Methods*, D.Sc. Thesis, Institute of Computer Science, Academy of Sciences, Prague, 2001.
[29] A. van der Sluis, *Condition numbers and equilibration matrices*, Numer. Math., 14 (1969), pp. 14–23.
[30] H. A. van der Vorst, *Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 631–644.
[31] H. F. Walker, *Implementation of the GMRES method using Householder transformations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 152–163.
[32] H. F. Walker, *Implementation of the GMRES method*, J. Comput. Phys., 53 (1989), pp. 311–320.
[33] D. Watkins, *Fundamentals of Matrix Computations*, John Wiley, New York, 1991.