MODIFIED GRAM-SCHMIDT (MGS), LEAST SQUARES, AND BACKWARD STABILITY OF MGS-GMRES

CHRISTOPHER C. PAIGE *, MIROSLAV ROZLOŽNÍK †, AND ZDENĚK STRAKOŠ †

Abstract. The generalized minimum residual method (GMRES) [Y. Saad and M. Schultz, *SIAM J. Sci. Statist. Comput.*, 7 (1986), pp. 856–869] for solving linear systems $Ax = b$ is implemented as a sequence of least squares problems involving Krylov subspaces of increasing dimensions. The most usual implementation is Modified Gram-Schmidt GMRES (MGS-GMRES). Here we show that MGS-GMRES is backward stable. The result depends on a more general result on the backward stability of a variant of the MGS algorithm applied to solving a linear least squares problem, and uses other new results on MGS and its loss of orthogonality, together with an important but neglected condition number, and a relation between residual norms and certain singular values.

Key words. rounding error analysis, backward stability, linear equations, condition numbers, large sparse matrices, iterative solution, Krylov subspace methods, Arnoldi method, generalized minimum residual method, modified Gram–Schmidt, QR factorization, loss of orthogonality, least squares, singular values

AMS subject classifications. 65F10, 65F20, 65F25, 65F35, 65F50, 65G50, 15A12, 15A42

1. Introduction. Consider a system of linear algebraic equations $Ax = b$, where A is a given n by n (unsymmetric) nonsingular matrix and b a nonzero 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 [2] is a Krylov subspace method designed for solving the eigenproblem of unsymmetric matrices. The generalized minimum residual method (GMRES) [20] 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 [24] and QMR [9] for unsymmetric A, and LSQR [16] for unsymmetric or rectangular A. However GMRES is widely used for solving linear systems arising from discretization of partial differential equations, and as we show, it is backward stable and it does effectively minimize the 2-norm of the residual at each step.

The most usual way of applying the Arnoldi method for large sparse unsymmetric A is to use modified Gram–Schmidt orthogonalization (MGS). Unfortunately in finite precision computations this leads to loss of orthogonality among the MGS Arnoldi

*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 (miro@cs.cas.cz), (strakos@cs.cas.cz). The work of the last two authors was supported by the project 1ET400300415 within the National Program of Research "Information Society" and by the Institutional Research Plan AV0Z10300504 "Computer Science for the Information Society: Models, Algorithms, Applications".

vectors. If these vectors are used in GMRES we have MGS-GMRES. Experience suggested that MGS-GMRES succeeds despite this loss of orthogonality, see [12]. For this reason we examine the MGS version of Arnoldi's algorithm and use this to show that the MGS–GMRES method does eventually produce a backward stable approximate solution when applied to any member of the following class of linear systems with floating point arithmetic unit roundoff ϵ (σ means singular value):

$$Ax = b \neq 0, \quad A \in \mathbf{R}^{n\times n}, \quad b \in \mathbf{R}^n, \quad \sigma_{min}(A) \gg n^2\epsilon\|A\|_F. \tag{1.1}$$

(See also Appendix A. The restriction here is deliberately imprecise, see below.) Moreover we show that MGS–GMRES gives backward stable solutions for its least squares problems at all iteration steps, thus answering important open questions. The proofs depend on new results on the loss of orthogonality and backward stability of the MGS algorithm, as well as the application of the MGS algorithm to least squares problems, and a lot of this paper is devoted to first obtaining these results.

While the k-th step of MGS produces the k-th orthonormal vector v_k, it is usual to say v_k is produced by step $k-1$ in the Arnoldi and MGS-GMRES algorithms. We will attempt to give a consistent development while avoiding this confusion. Thus step $k-1$ of MGS-GMRES is essentially the k-th step of MGS applied to $[b, AV_{k-1}]$ to produce v_k in $[b, AV_{k-1}] = V_kR_k$, where $V_k \equiv [v_1, \ldots, v_k]$ and R_k is upper triangular. In practice, if we reach a solution at step $m-1$ of MGS-GMRES, then numerically b must lie in the range of AV_{m-1}, so that $B_m \equiv [b, AV_{m-1}]$ is numerically rank deficient. But this means we have to show that our rounding error analysis of MGS holds for rank deficient B_m — and this requires an extension of some results in [5].

In Section 2 we describe our notation and present some of the tools we need which may be of more general use. For example we show the importance of the condition number $\tilde{\kappa}_F(A)$ in (2.1), prove the existence of a nearby vector in Lemma 2.3, and provide a variant of the singular value–residual norm relations of [17] in Theorem 2.4. In Sections 3.1–3.2 we review MGS applied to $n\times m$ B of rank m, and its numerical equivalence to the Householder QR reduction of B augmented by an $m \times m$ matrix of zeros. In Section 3.3 we show how the MGS rounding error results extend to the case of $m > n$, while in Section 4 we show how these results apply to the Arnoldi algorithm. In Section 5 we analyze the loss of orthogonality in MGS and the Arnoldi algorithm and how it is related to the near rank deficiency of the columns of B or its Arnoldi equivalent, refining a nice result of Giraud and Langou [10, 14]. Section 6 introduces the key step used to prove convergence of these iterations. In Section 7.1 we prove the backward stability of the MGS algorithm applied to solving linear least squares problems of the form required by the MGS-GMRES algorithm, and in Section 7.2 we show how loss of orthogonality is directly related to new normwise relative backward errors of a sequence of *different* least squares problems, supporting a conjecture on the convergence of MGS-GMRES and its loss of orthogonality, see [18]. In Section 8.1 we show that at every step MGS-GMRES computes a backward stable solution for that step's linear least squares problem, and in Section 8.2 we show that one of these solutions is also a backward stable solution for (1.1) in at most $n+1$ MGS steps.

The restriction on A in (1.1) is essentially a warning to be prepared for difficulties in using the basic MGS-GMRES method on singular systems, see for example [6, 23]. The imprecise nature of the condition (using $\gg$ instead of $>$ with some constant) was chosen to make the presentation easier. A constant could be provided (perhaps closer to 100 than 10), but since the long bounding sequence used was so loose, it would be meaningless. Appendix A suggests that the form $n^2\epsilon\|A\|_F$ might be optimal, but

since for large n rounding errors tend to combine in a semi-random fashion, it is reasonable to replace n^2 by n, and a more practical requirement than (1.1) might be:

$$\text{For large } n, \qquad n\epsilon\|A\|_F/\sigma_{min}(A) \le 0.1. \tag{1.2}$$

2. Notation and mathematical basics. We describe the notation we will use, together with some generally useful results. We use "≡" to mean "is defined as" in the first occurrence of an expression, but in any later occurrences of this expression it means "is equivalent to (by earlier definition)". A bar above a symbol will denote a computed quantity, so if V_k is an ideal mathematical quantity, $\bar{V}_k$ will denote its actual computed value. The floating point arithmetic unit roundoff will be denoted by ϵ (half the machine epsilon, see [13, pp.37–38]), I_n denotes the $n\times n$ unit matrix, e_j will be the j-th column of a unit matrix I, so Be_j is the j-th column of B, and $B_{i:j} \equiv [Be_i, \ldots, Be_j]$. We will denote the absolute value of a matrix B by $|B|$, its Moore-Penrose generalized inverse by $B^\dagger$, $\|\cdot\|_F$ will denote the Frobenius norm, $\sigma(\cdot)$ will denote a singular value, and $\kappa_2(B) \equiv \sigma_{max}(B)/\sigma_{min}(B)$. See (2.1) for $\tilde{\kappa}_F(\cdot)$. Matrices and vectors whose first symbol is Δ, such as ΔV_k, will denote rounding error terms. For the rounding error analyses we will use Higham's notation [13, pp.63–68]: $\tilde{c}$ will denote a small integer ≥ 1 whose exact value is unimportant, ($\tilde{c}$ might have a different value at each appearance) and $\gamma_n \equiv n\epsilon/(1-n\epsilon)$, $\tilde{\gamma}_n \equiv \tilde{c}n\epsilon/(1-\tilde{c}n\epsilon)$. Without mentioning it again, we will always assume the conditions are such that the denominators in objects like this (usually bounds) are positive, see for example [13, (19.6)]. We see $\tilde{\gamma}_n/(1-\tilde{\gamma}_n) = \tilde{c}n\epsilon/(1-2\tilde{c}n\epsilon)$, and might write $\tilde{\gamma}_n/(1-\tilde{\gamma}_n) = \tilde{\gamma}'_n$ for mathematical correctness, but will refer to the right hand side as $\tilde{\gamma}_n$ thereafter. E_m, $\hat{E}_m$, $\tilde{E}_m$ will denote matrices of rounding errors (see just before Theorem 3.3), and $\|E_m e_j\|_2 \le \gamma\|Be_j\|_2$ implies this holds for $j = 1, \ldots, m$ unless otherwise stated.

REMARK 2.1. *(See also Appendix A). An important idea used throughout this paper is that column bounds of the above form lead to several results which are independent of column scaling, and we take advantage of this by using the following choice of condition number. Throughout the paper, D will represent* any *positive definite diagonal matrix.*

The choice of norms is key to making error analyses readable, and fortunately there is a compact column-scaling-independent result with many uses. Define

$$\tilde{\kappa}_F(A) \equiv \min_{\text{diagonal } D>0} \|AD\|_F/\sigma_{min}(AD). \tag{2.1}$$

This leads to some useful new results.

LEMMA 2.1. *If E and B have m columns then for any positive definite diagonal matrix D: $\|Ee_j\|_2 \le \gamma\|Be_j\|_2,\ j = 1, \ldots, m, \ \Rightarrow\ \|ED\|_F \le \gamma\|BD\|_F$;*

$$\|Ee_j\|_2 \le \gamma\|Be_j\|_2 \text{ for } j = 1, \ldots, m \ \ \&\ \ \operatorname{rank}(B) = m \ \ \Rightarrow\ \ \|EB^\dagger\|_F \le \gamma\tilde{\kappa}_F(B).$$

With the QR factorization $B = Q_1 R$, this leads to $\|ER^{-1}\|_F \le \gamma\tilde{\kappa}_F(B) = \gamma\tilde{\kappa}_F(R)$.

Proof. $\|Ee_j\|_2 \le \gamma\|Be_j\|_2$ implies $\|EDe_j\|_2 \le \gamma\|BDe_j\|_2$ so $\|ED\|_F \le \gamma\|BD\|_F$. For B of rank m, $(BD)^\dagger = D^{-1}B^\dagger$, $\|(BD)^\dagger\|_2 = \sigma_{min}^{-1}(BD)$, and so

$$\|EB^\dagger\|_F = \|ED(BD)^\dagger\|_F \le \|ED\|_F\|(BD)^\dagger\|_2 \le \gamma\|BD\|_F/\sigma_{min}(BD).$$

Since this is true for all such D, we can take the minimum, proving our results. ∎

LEMMA 2.2. *If $m \times m$ $\bar{R}$ is nonsingular and $P_1^T P_1 = I$ in $P_1\bar{R} = B + E$, and $\gamma\tilde{\kappa}_F(B) < 1$, then*

$$\|Ee_j\|_2 \le \gamma\|Be_j\|_2, \quad j = 1, \ldots, m, \quad \Rightarrow \quad \|E\bar{R}^{-1}\|_F \le \gamma\tilde{\kappa}_F(B)/(1-\gamma\tilde{\kappa}_F(B)).$$

Proof. For any D in (2.1), $\|Ee_j\|_2 \le \gamma\|Be_j\|_2 \Rightarrow \|ED\|_F \le \gamma\|BD\|_F$, and then $\sigma_{min}(\bar{R}D) \ge \sigma_{min}(BD) - \gamma\|BD\|_F$, so $\|E\bar{R}^{-1}\|_F = \|ED(\bar{R}D)^{-1}\|_F$ is bounded by

$$\|ED\|_F\|(\bar{R}D)^{-1}\|_2 \le \frac{\gamma\|BD\|_F}{\sigma_{min}(BD) - \gamma\|BD\|_F} = \frac{\gamma\|BD\|_F/\sigma_{min}(BD)}{1 - \gamma\|BD\|_F/\sigma_{min}(BD)}.$$

Taking the minimum over D proves the result. ∎

Suppose $\bar{V}_m \equiv [\bar{v}_1, \ldots, \bar{v}_m]$ is an $n \times m$ matrix whose columns have been *computationally normalized* to have 2-norms of 1, and so have norms in $[1-\tilde{\gamma}_n, 1+\tilde{\gamma}_n]$. Now define $\tilde{V}_m \equiv [\tilde{v}_1, \ldots, \tilde{v}_m]$ where $\tilde{v}_j$ is just the correctly normalized version of $\bar{v}_j$, so

$$\begin{aligned} &\bar{V}_m = \tilde{V}_m(I + \Delta_m), \quad \Delta_m \equiv \mathrm{diag}(\nu_j), \quad \text{where } |\nu_j| \le \tilde{\gamma}_n, \quad j = 1, \ldots, m; \\ &\bar{V}_m^T\bar{V}_m = \tilde{V}_m^T\tilde{V}_m + \tilde{V}_m^T\tilde{V}_m.\Delta_m + \Delta_m.\tilde{V}_m^T\tilde{V}_m + \Delta_m.\tilde{V}_m^T\tilde{V}_m.\Delta_m, \\ &\|\bar{V}_m^T\bar{V}_m - \tilde{V}_m^T\tilde{V}_m\|_F / \|\tilde{V}_m^T\tilde{V}_m\|_F \le \tilde{\gamma}_n(2 + \tilde{\gamma}_n) \equiv \tilde{\gamma}_n'. \end{aligned} \tag{2.2}$$

From now on we will not document the analogs of the last step $\tilde{\gamma}_n(2 + \tilde{\gamma}_n) \equiv \tilde{\gamma}_n'$, but finish with $\le \tilde{\gamma}_n$. In general it will be as effective to consider $\tilde{V}_m$ as $\bar{V}_m$, and we will develop our results in terms of $\tilde{V}_m$ rather than $\bar{V}_m$. The following will be useful here

$$\|[\tilde{V}_m, I_n]\|_2^2 = \|I_n + \tilde{V}_m\tilde{V}_m^H\|_2 = 1 + \|\tilde{V}_m\tilde{V}_m^H\|_2 = 1 + \|\tilde{V}_m\|_2^2 \le 1 + \|\tilde{V}_m\|_F^2 = 1 + m. \tag{2.3}$$

Lemma 2.3 deals with the problem: Suppose we have $d \in \mathbf{R}^n$ and we know for some unknown perturbation $f \in \mathbf{R}^{(m+n)}$ that $\| \begin{bmatrix} 0 \\ d \end{bmatrix} + f\|_2 = \rho$. Is there a perturbation g *of the same dimension as* d, and having a similar norm to that of f, such that $\|d + g\|_2 = \rho$ also? Here we show such a g exists in the form $g = Nf$, $\|N\|_2 \le \sqrt{2}$.

Lemma 2.3. *For a given $d \in \mathbf{R}^n$ and unknown $f \in \mathbf{R}^{(m+n)}$, if*

$$\begin{bmatrix} f_1 \\ d + f_2 \end{bmatrix} \equiv \begin{bmatrix} 0 \\ d \end{bmatrix} + f = p\rho \equiv \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} \rho, \quad \textit{where} \quad \|p\|_2 = 1,$$

then there exists $0 \le \sigma \le 1$, $v \in \mathbf{R}^n$ with $\|v\|_2 = 1$, and $n \times (m+n)$ N of the form

$$N \equiv [v(1 + \sigma)^{-1}p_1^T, I_n], \tag{2.4}$$

so that

$$d + Nf = v\rho. \tag{2.5}$$

This gives

$$\| \begin{bmatrix} 0 \\ d \end{bmatrix} + f\|_2 = \|d + Nf\|_2 = \rho, \quad 1 \le \|N\|_2 \le \sqrt{2}. \tag{2.6}$$

Proof. Define $\sigma \equiv \|p_2\|_2$. If $\sigma = 0$ take any $v \in \mathbf{R}^n$ with $\|v\|_2 = 1$. Otherwise define $v \equiv p_2/\sigma$ so $\|v\|_2 = 1$. In either case $p_2 = v\sigma$ and $p_1^Tp_1 = 1 - \sigma^2$. Now define N as in (2.4), so

$$\begin{aligned} &d + Nf = d + v(1 + \sigma)^{-1}\|p_1\|_2^2\rho + f_2 = p_2\rho + v(1 - \sigma)\rho = v\rho \\ &NN^T = I + v(1 + \sigma)^{-2}(1 - \sigma^2)v^T, \\ &1 \le \|N\|_2^2 = \|NN^T\|_2 = 1 + (1 - \sigma)/(1 + \sigma) \le 2, \end{aligned}$$

proving (2.5) and (2.6). ∎

This is a refinement of a special case of [5, Lem.3.1], see also [13, Ex.19.12]. The fact that the perturbation g in d has the form of N times the perturbation f is important, as we shall see in Section 7.1.

Finally we give a general result on the relation between least squares residual norms and singular values. The bounds below were given in [17, Thm.4.1], but subject to a certain condition [17, (1.4)] that we cannot be sure will hold here. To prove that our results here hold subject to the different condition (1.1), we need to prove a related result. In order not to be too repetitive, we will prove a slightly more general result than we considered before, or need here, and make the theorem and proof brief.

THEOREM 2.4. *Let $B \in \mathbf{R}^{n\times k}$ have rank s and singular values $\sigma_1 \geq \cdots \geq \sigma_s > 0$. For $c \in \mathbf{R}^n$ and a scalar $\phi \geq 0$ define $\hat{y} \equiv B^\dagger c$, $\hat{r} \equiv c - B\hat{y}$, $\sigma(\phi) \equiv \sigma_{s+1}([c\phi, B])$ and $\delta(\phi) \equiv \sigma(\phi)/\sigma_s$. If $\hat{r}\phi \neq 0$ then $\sigma(\phi) > 0$, and if $\phi_0 \equiv \sigma_s/\|c\|$ then $\delta(\phi) < 1$ $\forall \phi \in [0, \phi_0)$,*

$$\sigma^2(\phi)(\phi^{-2} + \|\hat{y}\|_2^2) \leq \|\hat{r}\|_2^2 \leq \sigma^2(\phi)(\phi^{-2} + \|\hat{y}\|_2^2/[1-\delta^2(\phi)]), \quad \forall \phi > 0 \text{ s.t. } \delta(\phi) < 1.$$

Proof. $\hat{r}$ is the least squares residual for $By \approx c$, so $\hat{r}\phi \neq 0$ means $[c\phi, B]$ has rank $s+1$ and $\sigma(\phi) > 0$. If $0 \leq \phi < \phi_0$ then $\|c\phi\| < \|c\phi_0\| = \sigma_s$, so via Cauchy's interlacing theorem, $0 \leq \sigma(\phi) \equiv \sigma_{s+1}([c\phi, B]) < \sigma_s$, giving $0 \leq \delta(\phi) < 1$. Using the singular value decomposition $B = W\,\mathrm{diag}(\Sigma, 0)Z^T$, $W^T = W^{-1}$, $Z^T = Z^{-1}$, write

$$W^T[c, BZ] = \begin{bmatrix} a_1 & \Sigma & 0 \\ a_2 & 0 & 0 \end{bmatrix}, \quad \Sigma \equiv \begin{bmatrix} \sigma_1 & & \\ & \cdot & \\ & & \sigma_s \end{bmatrix}, \quad a_1 \equiv \begin{bmatrix} \alpha_1 \\ \cdot \\ \alpha_s \end{bmatrix}, \quad \hat{y} = Z \begin{bmatrix} \Sigma^{-1} a_1 \\ 0 \end{bmatrix}.$$

It is then straightforward to show, see for example [26, (39.4)], [17, (2.6)], [15, pp.1508–10], *et al.*, that for all ϕ such that $\phi > 0$ and $\delta(\phi) < 1$, $\sigma(\phi)$ is the smallest root of

$$\|\hat{r}\|^2 = \sigma(\phi)^2 \left[\phi^{-2} + \sum_{i=1}^{s} \frac{\alpha_i^2/\sigma_i^2}{1 - \sigma(\phi)^2/\sigma_i^2} \right].$$

But then
$$\|\hat{y}\|_2^2 = \sum_{i=1}^{s} \frac{\alpha_i^2}{\sigma_i^2} \leq \sum_{i=1}^{s} \frac{\alpha_i^2/\sigma_i^2}{1-\sigma(\phi)^2/\sigma_i^2} \leq \sum_{i=1}^{s} \frac{\alpha_i^2/\sigma_i^2}{1-\sigma(\phi)^2/\sigma_s^2} = \frac{\|\hat{y}\|_2^2}{1-\delta^2(\phi)}$$

while $\delta(\phi) \equiv \sigma(\phi)/\sigma_s < 1$, and the result follows. ∎

We introduced ϕ_0 to show $\delta(\phi) < 1$ for some $\phi > 0$. For results related to Theorem 2.4 we refer to [15, pp.1508–1510], which first introduced this useful value ϕ_0.

3. The Modified Gram-Schmidt (MGS) algorithm. In order to understand the numerical behavior of the MGS-GMRES algorithm, we first need a very deep understanding of the MGS algorithm. Here this is obtained by a further study of the numerical equivalence between MGS and the Householder QR factorization of an augmented matrix, see [5], and also [13, §19.8].

We do not give exact bounds, but work with terms of the form $\tilde{\gamma}_n$ instead, see [13, pp.63–68] and our Section 2. The exact bounds will not even be approached for the large n we are interested in, so there is little reason to include such fine detail. In Sections 3.1–3.3 we will review the MGS-Householder equivalence and extend some of the analysis that was given in [5] and [13, §19.8].

3.1. The basic MGS algorithm. Given a matrix $B \in \mathbf{R}^{n\times m}$ with rank $m \leq n$, the Modified Gram-Schmidt algorithm (MGS) in theory produces V_m and nonsingular R_m in the QR factorization

$$B = V_m R_m, \qquad V_m^T V_m = I_m, \qquad R_m \text{ upper triangular}, \tag{3.1}$$

where $V_m \equiv [v_1, \ldots, v_m]$, and $m \times m$ $R_m \equiv (\rho_{ij})$. The version of the MGS algorithm which immediately updates all columns computes a sequence of matrices $B = B^{(1)}, B^{(2)}, \ldots, B^{(m+1)} = V_m \in \mathbf{R}^{n\times m}$, where $B^{(i)} = [v_1, \ldots, v_{i-1}, b_i^{(i)}, \ldots, b_m^{(i)}]$. Here the first $(i-1)$ columns are final columns in V_m, and $b_i^{(i)}, \ldots, b_m^{(i)}$ have been made orthogonal to $v_1, \ldots, v_{i-1}$. In the i-th step we take

$$\rho_{ii} := \|b_i^{(i)}\|_2 \neq 0 \ \text{ since } \operatorname{rank}(B) = m, \quad v_i := b_i^{(i)}/\rho_{ii}, \tag{3.2}$$

and orthogonalize $b_{i+1}^{(i)}, \ldots, b_m^{(i)}$ against v_i using the orthogonal projector $I - v_i v_i^T$,

$$\rho_{ij} := v_i^T b_j^{(i)}, \qquad b_j^{(i+1)} := b_j^{(i)} - v_i \rho_{ij}, \qquad j = i+1, \ldots, m. \tag{3.3}$$

We see $B^{(i)} = B^{(i+1)} R^{(i)}$ where $R^{(i)}$ has the same i-th row as R_m, but is the unit matrix otherwise. Note that in the m-th step no computation is performed in (3.3), so that after m steps we have obtained the factorization

$$B = B^{(1)} = B^{(2)} R^{(1)} = B^{(3)} R^{(2)} R^{(1)} = B^{(m+1)} R^{(m)} \ldots R^{(1)} = V_m R_m, \tag{3.4}$$

where in exact arithmetic the columns of V_m are orthonormal by construction.

This formed R_m a row at a time. If the j-th column of B is only available after v_{j-1} is formed, as in MGS-GMRES, then we usually form R_m a column at a time. This does not alter the numerical values if we produce $\rho_{1,j}$, $b_j^{(2)}$; $\rho_{2,j}$, $b_j^{(3)}$; *etc.*.

It was shown in [3] that for the computed $\bar{R}_m$ and $\bar{V}_m$ in MGS

$$B + E = \bar{V}_m \bar{R}_m, \quad \|E\|_2 \leq c_1(m,n)\epsilon \|B\|_2, \quad \|I - \bar{V}_m^T \bar{V}_m\|_2 \leq c_2(m,n)\epsilon\kappa_2(B), \tag{3.5}$$

where $c_i(m,n)$ denoted a scalar depending on m, n and the details of the arithmetic. We get a deeper understanding by examining the MGS-Householder QR relationship.

3.2. MGS as a Householder method. The modified Gram-Schmidt algorithm for the QR factorization of B can be interpreted as an orthogonal transformation applied to the matrix B augmented with a square matrix of zero elements on top. This is true in theory for *any* method of QR factorization, but for Householder's method *it is true in the presence of rounding errors as well*. This observation was made by Charles Sheffield, and relayed to the authors of [5] by Gene Golub.

First we look at the theoretical result. Let $B \in \mathbf{R}^{n\times m}$ have rank m, and let $O_m \in \mathbf{R}^{m\times m}$ be a zero matrix. Consider the QR factorization

$$\tilde{B} \equiv \begin{bmatrix} O_m \\ B \end{bmatrix} = P_m \begin{bmatrix} R \\ 0 \end{bmatrix} \equiv \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix}, \quad P_m^T = P_m^{-1}. \tag{3.6}$$

Since B has rank m, P_{11} is zero, P_{21} is an $n \times m$ matrix of orthonormal columns, and, see (3.1), $B = V_m R_m = P_{21} R$. If upper triangular R_m and R are both chosen to have positive diagonal elements in $B^T B = R_m^T R_m = R^T R$, then $R_m = R$ by uniqueness, so $P_{21} = V_m$ can be found from any QR factorization of the augmented matrix $\tilde{B}$. The last n columns of P_m are then arbitrary up to an $n \times n$ orthogonal multiplier, but in theory the Householder reduction produces, see [5, (2.7)–(2.8)], the (surprisingly symmetric) orthogonal matrix

$$P_m = \begin{bmatrix} O_m & V_m^T \\ V_m & I - V_m V_m^T \end{bmatrix}, \tag{3.7}$$

showing that in this case P_m is fully defined by V_m.

A crucial result for this paper is that the Householder QR factorization giving (3.6) is also *numerically* equivalent to MGS applied to B. A close look at this Householder reduction, see for example [5, (2.6)–(2.7)], shows that for the computed version

$$\bar{P}_m^T \equiv \bar{P}^{(m)} \cdots \bar{P}^{(1)}, \qquad \bar{P}^{(j)} = I - \bar{p}_j \bar{p}_j^T, \qquad \bar{p}_j = \begin{bmatrix} -e_j \\ \bar{v}_j \end{bmatrix}, \qquad j = 1, \ldots, m, \tag{3.8}$$

where the $\bar{v}_j$ are *numerically identical* to the computed $\bar{v}_j$ in (3.2), so for example after the first two Householder transformations, our computed equivalent of $\bar{P}^{(2)}\bar{P}^{(1)}\tilde{B}$ is

$$\begin{bmatrix} \bar{\rho}_{11} & \bar{\rho}_{12} & \bar{\rho}_{13} & \cdots & \bar{\rho}_{1m} \\ 0 & \bar{\rho}_{22} & \bar{\rho}_{23} & \cdots & \bar{\rho}_{2m} \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \bar{b}_3^{(3)} & \cdots & \bar{b}_m^{(3)} \end{bmatrix}, \tag{3.9}$$

where the $\bar{\rho}_{jk}$ and $\bar{b}_k^{(j)}$ are also *numerically identical* to the corresponding computed values in (3.2) and (3.3). That is, in practical computations, the $\bar{v}_j$, $\bar{\rho}_{jk}$ and $\bar{b}_k^{(j)}$ are *identical* in both algorithms, see [5, p.179]. Note that the j-th row of $\bar{R}_m$ is completely formed in the j-th step and not touched again, while $\bar{b}_j^{(j)}$ is eliminated.

3.3. MGS applied to $n\times m$ B with $m>n$. The paper [5] was written assuming that $m \le n$ and $n \times m$ B in (3.1) had rank m, but it was mentioned in [5, p.181] that the rank condition was not necessary for proving the equivalence mentioned in the last paragraph of Section 3.2 above. For computations involving $n \times m$ B with $m > n$, Householder QR on B will stop in at most $n-1$ steps, but both MGS on B, and Householder QR on $\tilde{B}$ in (3.6), can nearly always be carried on for the full m steps. The MGS–Householder QR equivalence also holds for $m > n$, since the MGS and augmented Householder methods, being identical theoretically and numerically, either both stop with some $\bar{\rho}_{kk} = 0$, $k < m$, see (3.2), or both carry on to step m. It is this $m > n$ case we need here, and we extend the results of [5] to handle this. Because of this numerical equivalence, the backward error analysis for the Householder QR factorization of the augmented matrix in (3.6) can also be applied to the modified Gram-Schmidt algorithm on B. Two basic lemmas contribute to Theorem 3.3 below.

LEMMA 3.1. *In dealing with Householder transformations such as (3.8), Wilkinson [26, §4.2] pointed out that it is perfectly general to analyze operations with $P = I - pp^T$ for p having no zero elements. (This means we can drop the zero elements of p and the corresponding elements of the unit matrix and vector that P is applied to. In (3.8) each p has at most $n+1$ nonzero elements that we need to consider).*

LEMMA 3.2. *[13, Lem.19.3]. In practice, if j Householder transformations are applied to a vector $b \in \mathbf{R}^n$, the computed result $\bar{c}$ satisfies*

$$\bar{c} = P_j \cdots P_2 P_1 (b + \Delta b), \qquad \|\Delta b\|_2 \le j\tilde{\gamma}_n \|b\|_2.$$

In Theorem 3.3, E_m will refer to rounding errors in the basic MGS algorithm, while later $\hat{E}_m$ will refer to errors in the basic MGS algorithm applied to solving the equivalent of the MGS-GMRES least squares problem, and $\tilde{E}_m$ will refer to errors in the MGS-GMRES algorithm. All these matrices will be of the following form:

$$E_m \in \mathbf{R}^{(m+n)\times m}, \quad E_m \equiv \begin{bmatrix} E'_m \\ E''_m \end{bmatrix} \begin{matrix} \}m \\ \}n \end{matrix}. \tag{3.10}$$

THEOREM 3.3. *Let $\bar{R}_m$ and $\bar{V}_m = [\bar{v}_1, \ldots, \bar{v}_m]$ be the computed results of MGS applied to $B \in \mathbf{R}^{n\times m}$ as in (3.1)–(3.4), but now allow $m > n$. For $j = 1, \ldots, m$, step j computes $\bar{v}_j$ and the j-th row of $\bar{R}_m$ and $\bar{b}_{j+1}^{(j+1)}, \ldots, \bar{b}_m^{(j+1)}$ (see (3.9)). Define*

$$\bar{p}_j = \begin{bmatrix} -e_j \\ \bar{v}_j \end{bmatrix}, \quad \bar{P}^{(j)} = I - \bar{p}_j \bar{p}_j^T, \quad \bar{P}_m = \bar{P}^{(1)} \bar{P}^{(2)} \ldots \bar{P}^{(m)}, \tag{3.11}$$

$$\tilde{v}_j = \bar{v}_j / \|\bar{v}_j\|_2, \quad \tilde{p}_j = \begin{bmatrix} -e_j \\ \tilde{v}_j \end{bmatrix}, \quad \tilde{P}^{(j)} = I - \tilde{p}_j \tilde{p}_j^T, \quad \tilde{P}_m = \tilde{P}^{(1)} \tilde{P}^{(2)} \ldots \tilde{P}^{(m)}.$$

Then $\tilde{P}^{(j)}$ is the orthonormal equivalent of the computed version $\bar{P}^{(j)}$ of the Householder matrix applied in the j-th step of the Householder QR factorization of $\tilde{B}$ in (3.6), so that $\tilde{P}_m^T \tilde{P}_m = I$, and for the computed version $\bar{R}_m$ of $R = R_m$ in (3.6), and any positive definite diagonal matrix D, see Lemma 2.1, (here $j = 1, \ldots, m$)

$$\tilde{P}_m \begin{bmatrix} \bar{R}_m \\ 0 \end{bmatrix} = \begin{bmatrix} E'_m \\ B + E''_m \end{bmatrix}; \quad \tilde{P}_m \text{ orthogonal}; \quad \bar{R}_m, E'_m \in \mathbf{R}^{m\times m}; \tag{3.12}$$

$$E_m \equiv \begin{bmatrix} E'_m \\ E''_m \end{bmatrix}; \quad \|E_m e_j\|_2 \le j\tilde{\gamma}_n \|Be_j\|_2, \|E_m D\|_F \le m\tilde{\gamma}_n \|BD\|_F;$$

$$\|\bar{R}_m e_j\|_2 \le \|Be_j\|_2 + \|E_m e_j\|_2 \le (1 + j\tilde{\gamma}_n)\|Be_j\|_2; \tag{3.13}$$

$$E'_m e_1 = 0, \quad \|E'_m e_j\|_2 \le j^{\frac{1}{2}} \tilde{\gamma}_n \|Be_j\|_2, \quad j = 2, \ldots, m; \tag{3.14}$$

$$\|E'_m D\|_F \le m^{\frac{1}{2}} \tilde{\gamma}_n \|(BD)_{2:m}\|_F;$$

$$\tilde{P}_m = \begin{bmatrix} \tilde{S}_m & (I - \tilde{S}_m)\tilde{V}_m^T \\ \tilde{V}_m(I - \tilde{S}_m) & I - \tilde{V}_m(I - \tilde{S}_m)\tilde{V}_m^T \end{bmatrix}, \qquad \tilde{P}_m \tilde{P}_m^T = I. \tag{3.15}$$

where $m\times m$ E'_m and $\tilde{S}_m$ are strictly upper triangular. The j-th row of E'_m is wholly produced in step j, just as the j-th row of $\bar{R}_m$ is. The j-th column of $\tilde{S}_m$ is not defined until step j, and is not altered thereafter. (If MGS stops with $\bar{\rho}_{kk} = 0$, see (3.2), rows $k, \ldots, m$ of $\bar{R}_m$ and E'_m are zero, and columns $k, \ldots, m$ of $\bar{V}_m$ and $\tilde{S}_m$ are nonexistent, so we replace m above by k).

Proof. The MGS–augmented Householder QR equivalence for the case of $m \le n$ was proven in [5], and that this extends to $m > n$ is proven in the first paragraph of Section 3.3. As a result we can apply Lemmas 3.1 & 3.2 to give (3.12)–(3.13). The ideal P in (3.6) has the structure in (3.7), but it was shown in [5, Thm.4.1, & (4.5)] (which did not require $n \ge m$ in our notation) that $\tilde{P}_m$ in (3.11) and (3.12) has the extremely important structure of (3.15), for some strictly upper triangular $m\times m$ $\tilde{S}_m$. Since $E'_m = \tilde{S}_m \bar{R}_m$, this is strictly upper triangular too.

The rest follow with Lemmas 3.1 & 3.2. We have used $\tilde{\gamma}_n = \tilde{\gamma}'_{n+1}$ rather than $\tilde{\gamma}_{m+n}$ because in each step, $\bar{p}_j$ in (3.11) has only $n+1$ elements, see (3.9) and Lemma 3.1. Row j in $\bar{R}_m$ is not touched again after it is formed in step j, see (3.9), and so the same is true for row j in E'_m in (3.12), see Lemma 3.1. Since $E'_m = \tilde{S}_m \bar{R}_m$, the j-th column of $\tilde{S}_m$ is not defined until $\bar{\rho}_{jj}$ is computed in step j, and since these three matrices are all upper triangular, it is not altered in later steps. Finally we obtain new bounds in (3.14). The element $\bar{\rho}_{ij}$ is formed by the *one* transformation $\bar{P}^{(i)}$ in (3.11) applied to $\bar{b}_j^{(i)}$ in (3.9), and so from Lemma 3.2 we can say (remember $(E'_m)_{ii} = 0$)

$$|(E'_m)_{ij}| \le \tilde{\gamma}_n \|\bar{b}_j^{(i)}\|_2 \le \tilde{\gamma}'_n \|Be_j\|_2, \quad j = i+1, \ldots, m,$$

which is quite loose, but leads to the the bounds in (3.14). ∎

Note that (3.14) involves $j^{\frac{1}{2}}$, rather than the j in previous publications.

REMARK 3.1. *It is counter-intuitive that E'_m is* strictly *upper triangular, so we will explain it. We need only consider the first augmented Householder-MGS transformation of the first vector to form $\bar{\rho}_{11}$ in (3.9). We can rewrite the relevant part of the first transformation ideally as, see (3.11) and Lemma 3.1,*

$$P\begin{bmatrix}0\\ b\end{bmatrix} = \begin{bmatrix}\rho\\ 0\end{bmatrix}, \qquad P = \begin{bmatrix}0 & v^T\\ v & I - vv^T\end{bmatrix}, \qquad b = v\rho, \qquad \|v\|_2 = 1.$$

From b we compute $\bar{\rho}$ and $\bar{v}$, then define $\tilde{v} \equiv \bar{v}/\|\bar{v}\|_2$ so $\|\tilde{v}\|_2 = 1$. In order for $E'_m e_1 = 0$ in (3.12), there must exist a backward error term Δb such that

$$\begin{bmatrix}0 & \tilde{v}^T\\ \tilde{v} & I - \tilde{v}\tilde{v}^T\end{bmatrix}\begin{bmatrix}0\\ b + \Delta b\end{bmatrix} = \begin{bmatrix}\bar{\rho}\\ 0\end{bmatrix},$$

which looks like $n+1$ conditions on the n-vector Δb. But multiplying throughout by P shows there is *a solution $\Delta b = \tilde{v}\bar{\rho} - b$. The element above Δb is* forced *to be zero, so that there are actually $n+1$ conditions on $n+1$ unknowns. An error analysis (see Lemma 3.2) then bounds $\|\Delta b\|_2 \le \tilde{\gamma}_n \|b\|_2$.*

4. The Arnoldi algorithm *as* MGS. The Arnoldi algorithm [2] is the basis of MGS-GMRES. We assume that the initial estimate of x in (1.1) is $x_0 = 0$, so that the initial residual $r_0 = b$, and use the Arnoldi algorithm with $\rho \equiv \|b\|_2$, $v_1 \equiv b/\rho$, to sequentially generate the columns of $V_{k+1} \equiv [v_1, \ldots, v_{k+1}]$ via the ideal process:

$$AV_k = V_k H_{k,k} + v_{k+1} h_{k+1,k} e_k^T = V_{k+1} H_{k+1,k}, \qquad V_{k+1}^T V_{k+1} = I_{k+1}. \tag{4.1}$$

Here $k \times k$ $H_{k,k} = (h_{ij})$ is upper Hessenberg, and we stop at the first $h_{k+1,k} = 0$. Because of the orthogonality, this ideal algorithm must stop for some $k \le n$. Then $AV_k = V_k H_{k,k}$ where $H_{k,k}$ has rank at least $k-1$. If $h_{k+1,k} = 0$ and $H_{k,k}$ has rank $k-1$, there exists a nonzero z such that $AV_k z = V_k H_{k,k} z = 0$, so that A must be singular. Thus when A is nonsingular so is $H_{k,k}$, and so in MGS-GMRES, solving $H_{k,k} y = e_1 \rho$ and setting $x = V_k y$ solves (1.1). But if A is singular, this might not provide a solution even to consistent $Ax = b$:

$$A = \begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}, \quad x = \begin{bmatrix}0\\ 1\end{bmatrix}, \quad v_1 = b = Ax = \begin{bmatrix}1\\ 0\end{bmatrix}, \quad AV_1 = V_1 H_{1,1}, \quad H_{1,1} = 0.$$

Thus it is no surprise that we will require a restriction of the form (1.1) to ensure that the *numerical* MGS-GMRES algorithm always obtains a meaningful solution.

To relate the Arnoldi and MGS-GMRES algorithms to the MGS algorithm, we now replace $k+1$ by m and say that in the m-th MGS step these produce v_m, and MGS-GMRES also produces the approximation $x_{m-1} = V_{m-1} y_{m-1}$ to the solution x of (1.1). Then apart from forming the Av_j, the algorithm we use to give (4.1) is identical to (3.2)–(3.3) with the same vectors v_j, and

$$b_1 \equiv b, \ \rho_{11} \equiv \rho; \quad \text{and for } j = 1, \ldots, m-1, \ b_{j+1} \equiv Av_j, \ \rho_{i,j+1} \equiv h_{i,j} \ i = 1, \ldots, j+1,$$

except that Av_j cannot be formed and orthogonalized against $v_1, \ldots, v_j$ until v_j is available. This does not alter the numerical values. Thus with upper triangular R_m,

$$B_m \equiv A\,[x, V_{m-1}] = [b, AV_{m-1}] = V_m\,[e_1\rho, H_{m,m-1}] \equiv V_m R_m, \quad V_m^T V_m = I. \tag{4.2}$$

So in theory *the Arnoldi algorithm obtains the QR factorization of* $B_m \equiv [b, AV_{m-1}]$ *by applying MGS to* B_m. Computationally we can see that we have applied MGS to $\bar{B}_m \equiv [b, fl(A\bar{V}_{m-1})]$ where $\bar{V}_{m-1} \equiv [\bar{v}_1, \ldots, \bar{v}_{m-1}]$ is the matrix of supposedly orthonormal vectors computed by MGS, and see for example [13, §3.5],

$$fl(A\bar{v}_j) = (A+\Delta A_j)\bar{v}_j, \quad |\Delta A_j| \le \gamma_n|A|, \quad \text{so} \quad fl(A\bar{V}_{m-1}) = A\tilde{V}_{m-1}+\Delta V_{m-1},$$
$$|\Delta V_{m-1}| \le \gamma_n|A|.|\bar{V}_{m-1}|, \quad \|\Delta V_{m-1}\|_F \le m^{\frac{1}{2}}\gamma_n\||A|\|_2 \le m^{\frac{1}{2}}\gamma_n\|A\|_F, \tag{4.3}$$

gives the computed version of $A\bar{V}_{m-1}$. We could replace n by the maximum number of non-zeros per row, while users of preconditioners, or less simple multiplications, could insert their own bounds on ΔV_{m-1} here.

REMARK 4.1. *The bounds in (4.3) are not column-scaling independent. Also any scaling applies to the columns of* $A\bar{V}_{m-1}$*, not to* A*, and so would not be of such an advantage for MGS-GMRES as for ordinary MGS. Therefore it would seem important to ensure the columns of* A *are reasonably scaled for MGS-GMRES — e.g. to approach the minimum over positive diagonal* D *of* $\|AD\|_F/\sigma_{min}(AD)$*, see Appendix A.*

The rounding error behavior of the Arnoldi algorithm is as follows.

THEOREM 4.1. *For the computational version of the Arnoldi algorithm (4.1) (with* $m \equiv k+1$*) with floating point arithmetic unit roundoff* ϵ *producing* $\bar{V}_m$ *and* $\bar{R}_m \equiv [e_1\bar{\rho}, \bar{H}_{m,m-1}]$*, see (4.2), there exists an* $n+m$ *square orthogonal matrix* $\tilde{P}_m$ *of the form (3.15) where* $\tilde{V}_m$ *is* $\bar{V}_m$ *with its columns correctly normalized, such that if*

$$\bar{B}_m \equiv [b, fl(A\bar{V}_{m-1})] = [b, A\tilde{V}_{m-1}] + [0, \Delta V_{m-1}], \tag{4.4}$$

where we can use the bounds on ΔV_{m-1} *in (4.3), then all the results of Theorem 3.3 apply with* B *there replaced by* $\bar{B}_m$ *here.*

Thus whatever we say for MGS will hold for the Arnoldi algorithm if we simply replace B by $\bar{B}_m \equiv [b, fl(A\bar{V}_{m-1})] = [b, A\tilde{V}_{m-1}]+[0, \Delta V_{m-1}]$. The key idea of viewing the Arnoldi algorithm as MGS applied to $[b, AV_n]$ appeared in [25]. It was used in [8] and [1], and in particular in [18], in which we outlined another possible approach to backward stability analysis of MGS–GMRES. Here we have chosen a different way of proving the backward stability result, and this follows the spirit of [5] and [10].

5. Loss of orthogonality of $\bar{V}_m$ from MGS and the Arnoldi algorithm. The analysis here is applicable to both the MGS and Arnoldi algorithms. B will denote the given matrix in MGS, or $\bar{B}_m \equiv [b, fl(A\bar{V}_{m-1})]$ in the Arnoldi algorithm. Unlike [10, 14], we do not base the theory on [5, Lem.3.1], since a direct approach is cleaner and gives nicer results. It is important to be aware that our bounds will be of a different nature to those in [10, 14]. Even though the rounding error analysis of MGS in [10, 14] is based on the ideas in [5], the bounds obtained in [10] and [14, pp.32–38] are unexpectedly strong compared with our results based on [5]. This is because [10, (18)–(19)] and [14, (1.68)–(1.69)] leading to [10, Thm.3.1] and [14, Thm.1.4.1] follow from [26, p.160, (45.3)]. But in Wilkinson's [26], (45.3) follows from his (45.2), (45.1) and (44.6), where this last is clearly for fl_2 arithmetic (double precision accumulation of inner products). Since double precision is used in [10, 14], their analysis is essentially assuming what could be called fl_4 — quadruple precision accumulation of inner products. This is not stated in [10, 14], and the result is that their bounds appear to be much better (tighter) and the conditions much easier (less strict) than those that would have been obtained using standard floating point arithmetic. We will now obtain refined bounds based on our standard floating point arithmetic analysis, and attempt to correct this misunderstanding.

REMARK 5.1. *The $\tilde\gamma_n$ in each expression in (3.12)–(3.14) is essentially the same $\tilde\gamma_n$, that from Lemma 3.2, so we will call it $\hat\gamma_n$. We could legitimately absorb various small constants into a series of new $\tilde\gamma_n$, but that would be less transparent, so we will develop a sequence of loose bounds based on this fixed $\hat\gamma_n$.*

To simplify our bounds, we use "$\{\le\}$" to mean "$\le$" under the assumption that $m\hat\gamma_n\tilde\kappa_F(B)\le 1/8$. Note that this has the following consequences.

$$m\hat\gamma_n\tilde\kappa_F(B)\le 1/8 \quad\Rightarrow\quad \{\ (1-m\hat\gamma_n\tilde\kappa_F(B))^{-1}\le 8/7 \quad \&$$
$$\mu\equiv m^{\frac12}\hat\gamma_n\tilde\kappa_F(B)8/7\le 1/7 \quad \&\quad (1+\mu)/(1-\mu)\le 4/3\ \}. \tag{5.1}$$

The basic bound is for $\tilde S_m = E_m'\bar R_m^{-1}$, see (3.12), (3.15). This is part of an orthogonal matrix so $\|\tilde S_m\|_2\le 1$. From (3.12) and (3.14) for any $m\times m$ diagonal matrix $D>0$,

$$\|\tilde S_m\|_F = \|E_m'D(\bar R_mD)^{-1}\|_F \le \|E_m'D\|_F\|(\bar R_mD)^{-1}\|_2 = \|E_m'D\|_F/\sigma_{min}(\bar R_mD)$$
$$\le \frac{\|E_m'D\|_F}{\sigma_{min}(BD)-\|E_mD\|_2} \le \frac{m^{\frac12}\hat\gamma_n\|(BD)_{2:m}\|_F}{\sigma_{min}(BD)-m\hat\gamma_n\|BD\|_F}, \tag{5.2}$$

$$\|\tilde S_m\|_F \le m^{\frac12}\hat\gamma_n\tilde\kappa_F(B)/(1-m\hat\gamma_n\tilde\kappa_F(B))\ \{\le\}\ \frac87 m^{\frac12}\hat\gamma_n\tilde\kappa_F(B)\ \{\le\}\ \frac17. \tag{5.3}$$

with obvious restrictions. The bounds (5.3) took a minimum over D.

$\bar V_m\equiv[\bar v_1,\dots,\bar v_m]$ is the $n\times m$ matrix of vectors computed by m steps of MGS, $\tilde V_m\equiv[\tilde v_1,\dots,\tilde v_m]$ is the correctly normalized version of $\bar V_m$, so $\tilde V_m$ satisfies (2.2)–(2.3). Since $I-\tilde S_m$ is nonsingular upper triangular, the first m rows of $\tilde P_m$ in (3.15) give

$$(I-\tilde S_m)\tilde V_m^T\tilde V_m(I-\tilde S_m)^T = I-\tilde S_m\tilde S_m^T$$
$$= (I-\tilde S_m)(I-\tilde S_m)^T + (I-\tilde S_m)\tilde S_m^T + \tilde S_m(I-\tilde S_m)^T,$$
$$\tilde V_m^T\tilde V_m = I + \tilde S_m^T(I-\tilde S_m)^{-T} + (I-\tilde S_m)^{-1}\tilde S_m, \tag{5.4}$$
$$(I-\tilde S_m)^{-1}\tilde S_m = \tilde S_m(I-\tilde S_m)^{-1} = \text{strictly upper triangular part}(\tilde V_m^T\tilde V_m). \tag{5.5}$$

Since $\tilde V_{m-1}^T\tilde v_m$ is the above diagonal part of the last column of symmetric $\tilde V_m^T\tilde V_m - I$, (5.5) and (5.3) give the key bound (at first using $2m\hat\gamma_n\tilde\kappa_F(B)<1$, see (5.1))

$$\sqrt2\|\tilde V_{m-1}^T\tilde v_m\|_2 \le \|I-\tilde V_m^T\tilde V_m\|_F = \sqrt2\|(I-\tilde S_m)^{-1}\tilde S_m\|_F$$
$$\le \sqrt2\|\tilde S_m\|_F/(1-\|\tilde S_m\|_2) \le (2m)^{\frac12}\hat\gamma_n\tilde\kappa_F(B)/[1-(m+m^{\frac12})\hat\gamma_n\tilde\kappa_F(B)],$$
$$\{\le\}\ \frac43(2m)^{\frac12}\hat\gamma_n\tilde\kappa_F(B), \qquad (\textit{cf.}\ [3],\ [5,\ (5.3)]), \tag{5.6}$$

and similarly for $\bar V_m$, see (2.2). This is superior to the bound in [5], but the scaling idea is not new. Higham [13, p.373] (and in the 1996 first edition) argued that $\kappa_2(B)$ in [5] and [3], see (3.5), might be replaced by the minimum over positive diagonal matrices D of $\kappa_2(BD)$, which is almost what we have proven using $\tilde\kappa_F(B)$ in (2.1).

One measure of the extent of loss of orthogonality of $\tilde V_m$ is $\kappa_2(\tilde V_m)$.

LEMMA 5.1. *If $\tilde V_m^T\tilde V_m = I+\tilde F_m+\tilde F_m^T$ with strictly upper triangular $\tilde F_m$ and $\tilde S_m$ in $\tilde F_m\equiv\tilde S_m(I-\tilde S_m)^{-1}$, see (5.4), then for all singular values $\sigma_i(\tilde V_m)$*

$$\frac{1-\|\tilde S_m\|_2}{1+\|\tilde S_m\|_2}\le\sigma_i^2(\tilde V_m)\le\frac{1+\|\tilde S_m\|_2}{1-\|\tilde S_m\|_2}, \qquad \kappa_2(\tilde V_m)\le\frac{1+\|\tilde S_m\|_2}{1-\|\tilde S_m\|_2}.$$

Proof. Obviously $\|\tilde F_m\|_2\le\|\tilde S_m\|_2/(1-\|\tilde S_m\|_2)$. For any $y\in\mathbf{R}^k$ such that $\|y\|_2=1$, $\|\tilde V_my\|_2^2 = 1+2y^T\tilde F_my\le 1+2\|\tilde F_m\|_2\le(1+\|\tilde S_m\|_2)/(1-\|\tilde S_m\|_2)$, which gives

the upper bound on every $\sigma_i^2(\tilde{V}_m)$. From (5.4) $(I-\tilde{S}_m)\tilde{V}_m^T\tilde{V}_m(I-\tilde{S}_m)^T = I-\tilde{S}_m\tilde{S}_m^T$, so for any $y \in \mathbf{R}^k$ such that $\|y\|_2 = 1$, define $z \equiv (I-\tilde{S}_m)^T y$ so $\|z\|_2 \le 1+\|\tilde{S}_m\|_2$ and then

$$\frac{z^T\tilde{V}_m^T\tilde{V}_m z}{z^T z} = \frac{1-y^T\tilde{S}_m\tilde{S}_m^T y}{z^T z} \ge \frac{1-\|\tilde{S}_m\|_2^2}{(1+\|\tilde{S}_m\|_2)^2} = \frac{1-\|\tilde{S}_m\|_2}{1+\|\tilde{S}_m\|_2},$$

giving the lower bound on every $\sigma_i^2(\tilde{V}_m)$. The bound on $\kappa_2(\tilde{V}_m)$ follows. ■

Combining Lemma 5.1 with (5.1) and (5.3) gives the major result

$$\text{(5.7)}\quad \text{for } j=1,\dots,m,\ \ j\hat{\gamma}_n\tilde{\kappa}_F(B_j) \le 1/8 \ \ \Rightarrow\ \ \|\tilde{S}_j\|_F \le 1/7$$
$$\Rightarrow\ \ \kappa_2(\tilde{V}_j),\ \sigma_{min}^{-2}(\tilde{V}_j),\ \sigma_{max}^2(\tilde{V}_j)\ \le 4/3.$$

At this level the distinction between $\kappa_2(\bar{V}_m)$ and $\kappa_2(\tilde{V}_m)$ is miniscule, see (2.2), and by setting $j = m$ we can compare this with the elegant result which was the main theorem of Giraud and Langou [10], see [14, Thm.1.4.1]. In our notation:

THEOREM 5.2. *[10, Thm.3.1], [14, Thm.1.4.1]. Let $B \in \mathbf{R}^{n\times m}$ be a matrix with full rank $m \le n$ and condition number $\kappa_2(B)$ such that*

$$\text{(5.8)}\qquad 2.12(m+1)\epsilon < 0.01 \quad \textit{and} \quad 18.53m^{\frac{3}{2}}\epsilon\kappa_2(B) \le 0.1.$$

Then MGS in floating point arithmetic (Present comment in 2005: actually fl_2, or fl_4 if we use double precision) *computes $\bar{V}_m \in \mathbf{R}^{n\times m}$ as*

$$\kappa_2(\bar{V}_m) \le 1.3. \qquad ■$$

Note that the conditions (5.8) do not involve the dimension n of each column of $\bar{V}_m$, and this is the result of their analysis using fl_2. We can assume m satisfying the second condition in (5.8) will also satisfy the first.

To compare Theorem 5.2 with $j = m$ in (5.7), note that $m\tilde{\gamma}_n$ essentially means $\tilde{c}mn\epsilon$ for some constant $\tilde{c} > 1$, probably less than the 18.53 in Theorem 5.2. We assumed standard (IEEE) floating point arithmetic, but if we had assumed fl_2 arithmetic, that would have eliminated the n from our condition in (5.7). We used (2.1), which involves $\|BD\|_F \le m^{\frac{1}{2}}\|BD\|_2$. If we inserted this upper bound, that would mean our condition would be like that in Theorem 5.2, except we have the optimal result over column scaling, see (2.1). So if the same arithmetic is used, (5.7) is more revealing than Theorem 5.2. It is worth noting that with the introduction of XBLAS [7], the fl_2 and fl_4 options may become available in the near future.

6. A critical step in the Arnoldi and MGS-GMRES iterations. It will simplify the analysis if we use (5.7) to define a distinct value $\hat{m}$ of m. This value will depend on the problem and the constants we have chosen, but it will be sufficient for us to prove convergence and backward stability of MGS-GMRES in $\hat{m}-1 \le n$ steps. For the ordinary MGS algorithm remember $\bar{B}_m = B_m$, and think of m as increasing.

$$\text{(6.1)}\qquad \text{Let } \hat{m} \text{ be the } \textit{first} \text{ integer such that } \kappa_2(\tilde{V}_{\hat{m}}) > 4/3,$$

then we know from (5.7) that for $\bar{B}_{\hat{m}}$ in the Arnoldi algorithm, see (4.4) and (2.1),

$$\text{(6.2)}\qquad \hat{m}\hat{\gamma}_n\tilde{\kappa}_F(\bar{B}_{\hat{m}}) > 1/8, \ \text{ so } \ \sigma_{min}(\bar{B}_{\hat{m}}D) < 8\hat{m}\hat{\gamma}_n\|\bar{B}_{\hat{m}}D\|_F \ \forall \text{ diagonal } D > 0.$$

But since $\sigma_{min}(\tilde{V}_j) \le \sigma_1(\tilde{v}_1) = \|\tilde{v}_1\|_2 = 1 \le \sigma_{max}(\tilde{V}_j)$, (6.1) also tells us that

$$\text{(6.3)}\qquad \kappa_2(\tilde{V}_j),\ \sigma_{min}^{-1}(\tilde{V}_j),\ \sigma_{max}(\tilde{V}_j)\ \le 4/3, \quad j = 1,\dots,\hat{m}-1.$$

The above reveals the philosophy of the present approach to proving backward stability of MGS-GMRES. Other approaches have been tried. Here all is based on $\tilde{\kappa}_F(\bar{B}_m)$ rather than the backward error or residual norm. In [12, Thm.3.2, p.713] a different approach was taken — the assumption was directly related to the norm of the residual. The present approach leads to very compact and elegant formulations, and it is hard to say now whether the earlier approaches (see [18]) would have succeeded.

7. Least squares solutions via MGS. The linear least squares problem

$$\hat{y} \equiv \arg\min_y \|b - Cy\|_2, \qquad \hat{r} \equiv b - C\hat{y}, \qquad C \in \mathbf{R}^{n\times(m-1)}, \tag{7.1}$$

may be solved via MGS in different ways. Here we discuss two of these ways, but first we remind the reader how this problem appears in MGS-GMRES with $C = AV_{m-1}$.

After carrying out step $m-1$ of the Arnoldi algorithm as in Section 4 to produce $[b, AV_{m-1}] = V_m R_m$, see (4.2), the MGS-GMRES algorithm in theory minimizes the 2-norm of the residual $\|r_{m-1}\|_2 = \|b - Ax_{m-1}\|_2$ over $x_{m-1} \in x_0 + \mathcal{K}_{m-1}(A, r_0)$, where for simplicity we are assuming $x_0 = 0$ here. It does this by using V_{m-1} from (4.1) to provide an approximation $x_{m-1} \equiv V_{m-1}y_{m-1}$ to the solution x of (1.1). Then the corresponding residual is

$$r_{m-1} \equiv b - Ax_{m-1} = [b, AV_{m-1}] \begin{bmatrix} 1 \\ -y_{m-1} \end{bmatrix} = V_m R_m \begin{bmatrix} 1 \\ -y_{m-1} \end{bmatrix}, \tag{7.2}$$

where $R_m \equiv [e_1\rho, H_{m,m-1}]$. The ideal least squares problem is

$$y_{m-1} = \arg\min_y \|[b, AV_{m-1}] \begin{bmatrix} 1 \\ -y \end{bmatrix} \|_2, \tag{7.3}$$

but (in theory) the MGS-GMRES least squares solution is found by solving

$$y_{m-1} \equiv \arg\min_y \|R_m \begin{bmatrix} 1 \\ -y \end{bmatrix} \|_2. \tag{7.4}$$

7.1. The MGS least squares solution used in MGS-GMRES. If $B = [C, b]$ in (3.1)–(3.4), and C has rank $m-1$, then it was shown in [5, (6.3)], see also [13, §20.3], that MGS can be used to compute $\hat{y}$ in (7.1) in a backward stable way. Here we need to show that we can solve (7.1) in a stable way with MGS applied to $B = [b, C]$ (note the reversal of C and b) in order to prove the backward stability of MGS-GMRES. Just remember $B = [b, C] \equiv \bar{B}_m$ in (4.4), for MGS-GMRES. The analysis could be based directly on [5, Lem.3.1], but the following is more precise.

Let MGS on B in (3.1) lead to the computed $\bar{R}_m$ (we can assume $\bar{R}_m$ is nonsingular, see later) satisfying (3.12), where $B = [b, C]$. Then (3.12) and (7.1) give

$$\tilde{P}_m \begin{bmatrix} \bar{R}_m \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ [b, C] \end{bmatrix} + E_m; \quad \|E_m e_j\|_2 \le j\tilde{\gamma}_n \|[b, C]e_j\|_2, \quad j = 1, \ldots, m, \tag{7.5}$$

$$\hat{y} \equiv \arg\min_y \|B \begin{bmatrix} 1 \\ -y \end{bmatrix} \|_2, \qquad \hat{r} = B \begin{bmatrix} 1 \\ -\hat{y} \end{bmatrix}. \tag{7.6}$$

To solve this latter computationally, having applied MGS to B to give $\bar{R}_m$, we

$$\text{carry out a backward stable solution of } \min_y \|\bar{R}_m \begin{bmatrix} 1 \\ -y \end{bmatrix} \|_2 \tag{7.7}$$

by orthogonal reduction followed by solution of a triangular system. With (3.13) we will see this leads to

$$\hat{Q}^T(\bar{R}_m + \Delta R_m) = \begin{bmatrix} \bar{t} & \bar{U} + \Delta U \\ \bar{\tau} & 0 \end{bmatrix}, \qquad (\bar{U} + \Delta U)\bar{y} = \bar{t}, \tag{7.8}$$
$$\|\Delta R_m e_j\|_2 \le \tilde{\gamma}'_m \|\bar{R} e_j\|_2 \le \tilde{\gamma}_m \|B e_j\|_2 = \tilde{\gamma}_m \|[b, C] e_j\|_2, \quad j = 1, \ldots, m,$$

where $\hat{Q}$ is an orthogonal matrix while $\bar{\tau}$, $\bar{t}$, nonsingular upper triangular $\bar{U}$, and $\bar{y}$ are computed quantities. Here ΔU is the backward rounding error in the solution of the upper triangular system to give $\bar{y}$, see for example [13, Thm.8.3], and ΔR_m was obtained by combining ΔU with the backward rounding error in the QR factorization that produced $\bar{\tau}$, $\bar{t}$ and $\bar{U}$, see for example [13, Thm.19.10] (where here there are $m-1$ stages, each of one rotation). Clearly $\bar{y}$ satisfies

$$\bar{y} = \arg\min_y \|(\bar{R}_m + \Delta R_m) \begin{bmatrix} 1 \\ -y \end{bmatrix} \|_2. \tag{7.9}$$

In order to relate this least squares solution back to the MGS factorization of B, we add the error term ΔR_m to (7.5) to give (replacing $j\tilde{\gamma}_n + \tilde{\gamma}_m$ by $j\tilde{\gamma}_n$)

$$\tilde{P}_m \begin{bmatrix} (\bar{R}_m + \Delta R_m) \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ [b, C] \end{bmatrix} + \hat{E}_m, \quad \hat{E}_m \equiv E_m + \tilde{P}_m \begin{bmatrix} \Delta R_m \\ 0 \end{bmatrix}, \tag{7.10}$$
$$\|\hat{E}_m e_j\|_2 \le j\tilde{\gamma}_n \|[b, C] e_j\|_2, \quad j = 1, \ldots, m.$$

Now we can write for any $y \in \mathbf{R}^{m-1}$

$$r = r(y) \equiv b - Cy, \quad p = p(y) \equiv \tilde{P}_m \begin{bmatrix} (\bar{R}_m + \Delta R_m) \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ -y \end{bmatrix} = \begin{bmatrix} 0 \\ r \end{bmatrix} + \hat{E}_m \begin{bmatrix} 1 \\ -y \end{bmatrix}, \tag{7.11}$$

and we see from (2.6) in Lemma 2.3 that for any $y \in \mathbf{R}^{m-1}$ there exists $N(y)$ so that

$$\|p(y)\|_2 = \|(\bar{R}_m + \Delta R_m) \begin{bmatrix} 1 \\ -y \end{bmatrix} \|_2 = \|b - Cy + N(y)\hat{E}_m \begin{bmatrix} 1 \\ -y \end{bmatrix} \|_2, \quad \|N(y)\|_2 \le \sqrt{2}.$$

Defining $[\Delta b(y), \Delta C(y)] \equiv N(y)\hat{E}_m$ shows that for all $y \in \mathbf{R}^{m-1}$

$$\|(\bar{R}_m + \Delta R_m) \begin{bmatrix} 1 \\ -y \end{bmatrix} \|_2 = \|b + \Delta b(y) - [C + \Delta C(y)]y\|_2. \tag{7.12}$$

Thus $\bar{y}$ in (7.9) also satisfies

$$\bar{y} = \arg\min_y \|b + \Delta b(y) - [C + \Delta C(y)]y\|_2, \tag{7.13}$$
$$\|[\Delta b(y), \Delta C(y)] e_j\|_2 \le j\tilde{\gamma}_n \|[b, C] e_j\|_2, \qquad j = 1, \ldots, m,$$

where the bounds are independent of y, so that $\bar{y}$ is a backward stable solution for (7.1). That is, MGS applied to $B = [b, C]$ followed by (7.7) is backward stable as long as the computed $\bar{R}_m$ from MGS is nonsingular (we can stop early to ensure this). The almost identical analysis and result applies wherever b is in B, but we just gave the $B = [b, C]$ case for clarity.

Since we have a backward stable solution $\bar{y}$, we expect various related quantities to have reliable values, and we now quickly show two cases of this. If $\|E\|_F \le \gamma \|B\|_F$

then $\|Ey\|_2^2 = \sum_i \|e_i^T Ey\|_2^2 \le \sum_i \|e_i^T E\|_2^2 \|y\|_2^2 = \|E\|_F^2 \|y\|_2^2 \le \gamma^2 \|B\|_F^2 \|y\|_2^2$. So from the bounds in (7.10) we have for any $y \in \mathbf{R}^{m-1}$ the useful basic bound

$$\|\hat{E}_m \begin{bmatrix} 1 \\ -y \end{bmatrix}\|_2 \le \tilde{\gamma}_{mn} \psi_m(y), \quad \psi_m(y) \equiv \|b\|_2 + \|C\|_F \|y\|_2. \tag{7.14}$$

Multiplying (7.8) and (7.10) on the right by $\begin{bmatrix} 1 \\ -\bar{y} \end{bmatrix}$ shows that the residual $\bar{r}$ satisfies

$$\bar{r} \equiv b - C\bar{y}, \quad \tilde{P}_m \begin{bmatrix} \hat{Q} e_m \bar{\tau} \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \bar{r} \end{bmatrix} + \hat{E}_m \begin{bmatrix} 1 \\ -\bar{y} \end{bmatrix}, \quad |\, \|\bar{r}\|_2 - |\bar{\tau}| \,| \le \tilde{\gamma}_{mn} \psi_m(\bar{y}), \tag{7.15}$$

so that $|\tau|$ approximates $\|\bar{r}\|_2$ with a good relative error bound. Multiplying the last equality in this on the left by $[\tilde{V}_m, I_n]$, and using (3.15), (3.12), (7.10), (7.8), (3.14), and (2.3) with the argument leading to (7.14), we see

$$\tilde{V}_m \hat{Q} e_m \bar{\tau} = \bar{r} + [\tilde{V}_m, I_n] \hat{E}_m \begin{bmatrix} 1 \\ -\bar{y} \end{bmatrix} = \bar{r} + [\tilde{V}_m (E'_m + \Delta R_m) + E''_m] \begin{bmatrix} 1 \\ -\bar{y} \end{bmatrix}, \tag{7.16}$$

$$\|\bar{r} - \tilde{V}_m \hat{Q} e_m \bar{\tau}\|_2 \le \tilde{\gamma}_{mn} \psi_m(\bar{y}) \text{ for } m < \hat{m} \text{ in (6.1)}.$$

Thus $\bar{V}_m \hat{Q} e_m \bar{\tau}$ also approximates $\bar{r} \equiv b - C\bar{y}$ with a good relative error bound, see (2.2) and its following sentence.

7.2. Least squares solutions and loss of orthogonality in MGS. An apparently strong relationship was noticed between convergence of finite precision MGS-GMRES and loss of orthogonality among the Arnoldi vectors, see [12, 19]. It was thought that if this relationship was fully understood, we might use it to prove that finite precision MGS-GMRES would necessarily converge, see for example [18]. A similar relationship certainly *does* exist — it is the relationship between the loss of orthogonality in ordinary MGS applied to B, and the residual norms for what we will call the last vector least squares (LVLS) problems involving B, and we will derive this here. It adds to our understanding but it is not necessary for our other proofs, and could initially be skipped.

Because this is a theoretical tool, we will only consider rounding errors in the MGS part of the computation. We will do the analysis for MGS applied to any matrix $B = [b_1, \ldots, b_m]$. After step j we have $n \times j$ $\bar{V}_j$ and $j \times j$ $\bar{R}_j$, so that

$$\bar{R}_j \equiv \begin{bmatrix} \bar{U}_j & \bar{t}_j \\ & \bar{\tau}_j \end{bmatrix}, \quad \bar{U}_j \bar{y}_j = \bar{t}_j, \quad \bar{y}_j = \arg\min_y \|\bar{R}_j \begin{bmatrix} -y \\ 1 \end{bmatrix}\|_2, \quad |\bar{\tau}_j| = \|\bar{R}_j \begin{bmatrix} -\bar{y}_j \\ 1 \end{bmatrix}\|_2. \tag{7.17}$$

In theory $\bar{y}_j$ minimizes $\|b_j - B_{j-1} y\|_2$, but we would like to know that loss of orthogonality caused by rounding errors in MGS does not prevent this. One indicator of loss of orthogonality is $\tilde{V}_{j-1}^T \tilde{v}_j$. From (7.17) we see that

$$\bar{R}_j^{-1} = \begin{bmatrix} \bar{U}_j^{-1} & -\bar{U}_j^{-1} \bar{t}_j \bar{\tau}_j^{-1} \\ & \bar{\tau}_j^{-1} \end{bmatrix} = \begin{bmatrix} \bar{U}_j^{-1} & \begin{bmatrix} -\bar{y}_j \\ 1 \end{bmatrix} \bar{\tau}_j^{-1} \\ 0 & \end{bmatrix}, \quad \bar{R}_j^{-1} e_j \bar{\tau}_j = \begin{bmatrix} -\bar{y}_j \\ 1 \end{bmatrix}, \tag{7.18}$$

so that with (5.5) we have with $\bar{r}_j \equiv b_j - B_{j-1} \bar{y}_j$, (see (7.14) and (7.15) but now using E'_j and its bound in (3.14) rather than $\hat{E}_j$ and its bound in (7.10)),

$$(I - \tilde{S}_j) \begin{bmatrix} \tilde{V}_{j-1}^T \tilde{v}_j \\ 0 \end{bmatrix} = \tilde{S}_j e_j = E'_j \bar{R}_j^{-1} e_j = E'_j \begin{bmatrix} -\bar{y}_j \\ 1 \end{bmatrix} \bar{\tau}_j^{-1}, \ |\|\bar{r}_j\|_2 - |\bar{\tau}_j|| \le j^{\frac{1}{2}} \tilde{\gamma}_n \psi_m(\bar{y}_j). \tag{7.19}$$

Now define a *normwise relative backward error* (in the terminology of [13, Thm.7.1])

$$\beta_F(b,A,y) \equiv \beta_F^{A,b}(b,A,y) \quad \text{where} \quad \beta_F^{G,f}(b,A,y) \equiv \frac{\|b-Ay\|_2}{\|f\|_2 + \|G\|_F\|y\|_2}. \tag{7.20}$$

REMARK 7.1. *[13, Thm.7.1] assumes a vector norm with its subordinate matrix norm, but with the Frobenius norm in the denominator Rigal and Gaches' theory still works, so this is a possibly new, useful (and usually smaller) construct that is easier to compute than the usual one. A proof similar to that in [13, Thm.7.1] shows that*

$$\beta_F^{G,f}(b,A,y) = \min_{\delta A,\delta b}\{\eta : (A+\delta A)y = b+\delta b,\ \|\delta A\|_F \le \eta\|G\|_F,\ \|\delta b\|_2 \le \eta\|f\|_2\}.$$

Using (7.20) with the bounds in (3.14), (5.6), (7.19) and the definition in (7.14) (see also (5.3)) shows that

$$|\bar\tau_j|.\|\tilde V_{j-1}^T\tilde v_j\|_2 = \|(I-\tilde S_j)^{-1}E_j'\begin{bmatrix}-\bar y_j\\ 1\end{bmatrix}\|_2 \le j^{\frac12}\tilde\gamma_n\psi_m(\bar y_j)/(1-\|\tilde S_j\|_2),$$

$$\beta_F(b_j,B_{j-1},\bar y_j)\|\tilde V_{j-1}^T\tilde v_j\|_2 \le \frac{j^{\frac12}\tilde\gamma_n}{1-\|\tilde S_j\|_2}. \tag{7.21}$$

REMARK 7.2. *The product of the loss of orthogonality $\|\tilde V_{j-1}^T\tilde v_j\|_2$ at step j and the normwise relative backward error $\beta_F(b_j,B_{j-1},\bar y_j)$ of the LVLS problem is bounded by $O(\epsilon)$ until $\|\tilde S_j\|_2 \approx 1$, that is until orthogonality of the $\tilde v_1,\dots,\tilde v_j$ is totally lost, see (5.5), also Lemma 5.1.*

This is another nice result, as it again reveals how MGS applied to B_m loses orthogonality at *each* step — see the related Section 5. These bounds on the individual $\|\tilde V_{j-1}^T\tilde v_j\|_2$ complement the bounds in (5.6), since they are essentially in terms of the individual normwise relative backward errors $\beta_F(b_j,B_{j-1},\bar y_j)$, rather than $\tilde\kappa_F(B_j)$. However it is important to note that the "last vector" least squares (LVLS) problem considered in this section (see the line after (7.17)) is *not* the least squares problem solved for MGS-GMRES, which has the form of (7.6) instead. The two can give very different results in the general case, but in the problems we have solved via MGS-GMRES, these normwise relative backward errors seem to be of similar magnitudes for both problems, and this led to the conjecture in the first place. The similarity in behavior of the two problems is apparently related to the fact that B_m in MGS-GMRES is a Krylov basis. In this case it appears that the normwise relative backward errors of both least squares problems will converge (numerically) as the columns of B_j approach numerical linear dependence, see [17, 18]. Thus we have neither proven nor disproven the conjecture, but we have added weight to it.

8. Numerical behavior of the MGS-GMRES algorithm. We now only consider MGS-GMRES, and use k instead of $m-1$ to avoid many indices of the form $m-1$. In Section 4 we saw that k steps of the Arnoldi algorithm is in theory just $k+1$ steps of the MGS algorithm applied to $B_{k+1} \equiv [b, AV_k]$ to give $[b, AV_k] = V_{k+1}R_{k+1} = V_{k+1}[e_1\rho, H_{k+1,k}]$. And in practice the only difference in the rounding error analysis is that we apply ordinary MGS to $\bar B_{k+1} \equiv [b, fl(A\bar V_k)] = [b, A\tilde V_k] + [0, \Delta V_k]$, see (4.3). In Section 8.1 we combine this fact with the results of Section 7.1 to prove backward stability of the MGS-GMRES least squares solution $\bar y_k$ *at every step.*

In theory MGS-GMRES *must* solve $Ax = b$ for nonsingular $n \times n$ A in n steps since we cannot have more than n orthonormal vectors in $\mathbf{R}^n$. But in practice the vectors in MGS-GMRES lose orthogonality, so we need another way to prove that we reach a solution to (1.1). In Section 8.2 we will show that the MGS-GMRES algorithm for any problem satisfying (1.1) must, for some k, produce $\bar{V}_{k+1}$ so that numerically b lies in the range of $A\bar{V}_k$, and that MGS-GMRES must give a backward stable solution to (1.1). This k is $\hat{m}-1$, which is $\leq n$, see (6.1).

8.1. Backward stability of the MGS-GMRES least squares solutions. The equivalent of the MGS result (7.13) for MGS-GMRES is obtained by replacing $[b, C]$ by $\bar{B}_{k+1} \equiv [b, A\tilde{V}_k + \Delta V_k]$ throughout (7.13), see Theorem 4.1. Thus the computed $\bar{y}_k$ at step k in MGS-GMRES satisfies (with (4.3) and Section 6)

$$\bar{y}_k = \arg\min_y \|\tilde{r}_k(y)\|_2, \quad \tilde{r}_k(y) \equiv b + \Delta b_k(y) - [A\tilde{V}_k + \Delta V_k + \Delta C_k(y)]y \tag{8.1}$$

$$\|[\Delta b_k(y), \Delta C_k(y)]e_j\|_2 \leq \tilde{\gamma}_{kn}\|\bar{B}_{k+1}e_j\|_2, \quad j = 1, \ldots, k+1; \quad \|\Delta V_k\|_F \leq k^{\frac{1}{2}}\gamma_n\|A\|_F,$$

$$\|\Delta b_k(y)\|_2 \leq \tilde{\gamma}_{kn}\|b\|_2, \ \|\Delta V_k + \Delta C_k(y)\|_F \leq \tilde{\gamma}_{kn}[\|A\|_F + \|A\tilde{V}_k\|_F] \leq \tilde{\gamma}'_{kn}\|A\|_F \text{ if } k < \hat{m}.$$

This has proven the MGS-GMRES least squares solution $\bar{y}_k$ is backward stable for

$$\min_y \|b - A\tilde{V}_k y\|_2, \quad \text{for all } k < \hat{m},$$

which is all we need for this least squares problem. But even if $k \geq \hat{m}$, it is straightforward to show that it still gives a backward stable least squares solution.

8.2. Backward stability of MGS-GMRES for $Ax = b$ in (1.1). Even though MGS-GMRES always computes a backward stable solution $\bar{y}_k$ for the least squares problem (7.3), see Section 8.1, we still have to prove that $\bar{V}_k\bar{y}_k$ will be a a backward stable solution for the original system (1.1) for some k (we take this k to be $\hat{m}-1$ in (6.1)), and this is exceptionally difficult. Usually we want to show we have a backward stable solution when we *know* we have a small residual. The analysis here is different in that we will first prove that $\bar{B}_{\hat{m}}$ is numerically rank deficient, see (8.4), but to prove backward stability, we will then have to *prove* that our residual will be small, amongst other things, and this is far from obvious. Fortunately two little known researchers have studied this arcane area, and we will take ideas from [17], see Theorem 2.4. To simplify the development and expressions we will absorb all small constants into the $\tilde{\gamma}_{kn}$ terms below.

In (8.1) set $k \equiv \hat{m}-1 \leq n$ from (6.1), and write

$$\tilde{r}_k(\bar{y}_k) = b_k - A_k\bar{y}_k, \quad b_k \equiv b + \Delta b_k(\bar{y}_k), \quad A_k \equiv A\tilde{V}_k + \Delta\tilde{V}_k(\bar{y}_k), \tag{8.2}$$

$$\|\Delta b_k(\bar{y}_k)\|_2 \leq \tilde{\gamma}_{kn}\|b\|_2, \quad \Delta\tilde{V}_k(y) \equiv \Delta V_k + \Delta C_k(y), \quad \|\Delta\tilde{V}_k(y)\|_F \leq \tilde{\gamma}_{kn}\|A\|_F.$$

We need to take advantage of the scaling invariance of MGS in order to obtain our results. Here we need only scale b, so write $D \equiv \text{diag}(\phi, I_k)$ for any scalar $\phi > 0$. Since $\bar{B}_{k+1} \equiv [b, fl(A\bar{V}_k)] = [b, A\tilde{V}_k + \Delta V_k]$, from (8.2) with the bounds in (8.1) we have

$$[b_k\phi, A_k] = \bar{B}_{k+1}D + \Delta B_k D, \quad \Delta B_k \equiv [\Delta b_k(\bar{y}_k), \Delta C_k(\bar{y}_k)], \tag{8.3}$$

$$\|\Delta B_k D\|_F \leq \tilde{\gamma}_{kn}\|\bar{B}_{k+1}D\|_F \leq \tilde{\gamma}'_{kn}\|[b_k\phi, A_k]\|_F,$$

$$\|\bar{B}_{k+1}D\|_F \leq (1-\tilde{\gamma}_{kn})^{-1}\|[b_k\phi, A_k]\|_F, \quad \|b_k\|_2 \leq (1+\tilde{\gamma}_{kn})\|b\|_2.$$

In addition, $k+1$ is the first integer such that $\kappa_2(\tilde{V}_{k+1}) > 4/3$, so Section 6 gives

$$\begin{aligned} &\sigma_{min}(\bar{B}_{k+1}D) < 8(k+1)\hat{\gamma}_n\|\bar{B}_{k+1}D\|_F \le \tilde{\gamma}_{kn}\|[b_k\phi, A_k]\|_F,\ \forall\ \phi > 0; \\ &\kappa_2(\tilde{V}_k),\ \sigma_{min}^{-1}(\tilde{V}_k),\ \sigma_{max}(\tilde{V}_k) \le 4/3; \\ &\text{and similarly}\quad \|A_k\|_F \le \|A\tilde{V}_k\|_F + \tilde{\gamma}_{kn}\|A\|_F \le (4/3+\tilde{\gamma}_{kn})\|A\|_F. \end{aligned} \tag{8.4}$$

We can combine (8.2), (8.3) and (8.4) to give under the condition in (1.1)

$$\begin{aligned} &\sigma_{min}(A_k) \ge \sigma_{min}(A\tilde{V}_k) - \|\Delta\tilde{V}_k(\bar{y}_k)\|_2 \ge 3\sigma_{min}(A)/4 - \tilde{\gamma}_{kn}\|A\|_F > 0, \\ &\sigma_{min}([b_k\phi, A_k]) \le \sigma_{min}(\bar{B}_{k+1}D) + \|\Delta B_k D\|_2 \le \tilde{\gamma}_{kn}\|[b_k\phi, A_k]\|_F. \end{aligned} \tag{8.5}$$

The above allows us to define and analyze an important scalar, see Theorem 2.4,

$$\delta_k(\phi) \equiv \frac{\sigma_{min}([b_k\phi, A_k])}{\sigma_{min}(A_k)} \le 1, \tag{8.6}$$

where from (8.5) A_k has full column rank. Now $\bar{y}_k$ and $\tilde{r}_k(\bar{y}_k)$ solve the linear least squares problem $A_k y \approx b_k$ in (8.2), see (8.1). If $[b_k, A_k]$ does not have full column rank then $\tilde{r}_k(\bar{y}_k) = 0$, so $\tilde{x}_k \equiv \tilde{V}_k\bar{y}_k$ is a backward stable solution for (1.1), which we wanted to show. Next suppose $[b_k, A_k]$ has full column rank. We will not seek to minimize with respect to ϕ the upper bound on $\|\hat{r}\|_2^2$ in Theorem 2.4, which would be unnecessarily complicated, but instead prove that there exists a value $\hat{\phi}$ of ϕ satisfying (8.7) below, and use this value:

$$\hat{\phi} > 0, \qquad \sigma_{min}^2(A_k) - \sigma_{min}^2([b_k\hat{\phi}, A_k]) = \sigma_{min}^2(A_k)\|\bar{y}_k\hat{\phi}\|_2^2. \tag{8.7}$$

Writing LHS $\equiv \sigma_{min}^2(A_k) - \sigma_{min}^2([b_k\phi, A_k])$, RHS $\equiv \sigma_{min}^2(A_k)\|\bar{y}_k\phi\|_2^2$ we want to find ϕ so that LHS=RHS. But $\phi{=}0 \Rightarrow$ LHS $>$ RHS, while $\phi{=}\|\bar{y}_k\|_2^{-1} \Rightarrow$ LHS $<$ RHS, so from continuity $\exists\ \hat{\phi} \in (0, \|\bar{y}_k\|_2^{-1})$ satisfying (8.7). With (8.6) this shows that

$$\delta_k(\hat{\phi}) < 1, \quad \hat{\phi}^{-2} = \|\bar{y}_k\|_2^2/[1{-}\delta_k(\hat{\phi})^2], \quad 0 < \hat{\phi} < \|\bar{y}_k\|_2^{-1}. \tag{8.8}$$

It then follows from Theorem 2.4 that with (8.5), (8.8) and (8.4),

$$\begin{aligned} \|\tilde{r}_k(\bar{y}_k)\|_2^2 &\le \sigma_{min}^2([b_k\hat{\phi}, A_k])(\hat{\phi}^{-2} + \|\bar{y}_k\|_2^2/[1{-}\delta_k(\hat{\phi})^2]) \\ &\le \tilde{\gamma}_{kn}^2(\|b_k\hat{\phi}\|_2^2 + \|A_k\|_F^2)2\hat{\phi}^{-2}. \end{aligned} \tag{8.9}$$

But from (8.1) and (8.2) since $\tilde{r}_k(\bar{y}_k) = b_k - A_k\bar{y}_k$, $A_k^T\tilde{r}_k(\bar{y}_k) = 0$, and from (8.8),

$$\begin{aligned} \|b_k\hat{\phi}\|_2^2 &= \|\tilde{r}_k(\bar{y}_k)\hat{\phi}\|_2^2 + \|A_k\bar{y}_k\hat{\phi}\|_2^2, \\ &\le 2\tilde{\gamma}_{kn}^2(\|b_k\hat{\phi}\|_2^2 + \|A_k\|_F^2) + \|A_k\|_2^2(1{-}\delta_k(\hat{\phi})^2) \\ &\le 2\tilde{\gamma}_{kn}^2\|b_k\hat{\phi}\|_2^2 + (1 + 2\tilde{\gamma}_{kn}^2)\|A_k\|_F^2, \\ \|b_k\hat{\phi}\|_2^2 &\le \frac{1 + 2\tilde{\gamma}_{kn}^2}{1 - 2\tilde{\gamma}_{kn}^2}\|A_k\|_F^2. \end{aligned} \tag{8.10}$$

This with (8.4) and (8.5) shows that

$$\begin{aligned} \delta_k(\hat{\phi}) \equiv \frac{\sigma_{min}([b_k\hat{\phi}, A_k])}{\sigma_{min}(A_k)} &\le \frac{\tilde{\gamma}'_{kn}\|[b_k\hat{\phi}, A_k]\|_F}{\sigma_{min}(A){-}\tilde{\gamma}_{kn}\|A\|_F} \\ &\le \frac{\tilde{\gamma}''_{kn}\|A_k\|_F}{\sigma_{min}(A){-}\tilde{\gamma}_{kn}\|A\|_F} \le \frac{\tilde{\gamma}'''_{kn}\|A\|_F}{\sigma_{min}(A){-}\tilde{\gamma}_{kn}\|A\|_F} \le \frac{1}{2}\ \text{ under (1.1)}, \end{aligned} \tag{8.11}$$

since this last bound can be rewritten as $\sigma_{min}(A) \geq (2\tilde{\gamma}'''_{kn}+\tilde{\gamma}_{kn})\|A\|_F$, which we see will hold if A satisfies (1.1). This bound on $\delta_k(\hat{\phi})$ shows that $\hat{\phi}^{-2} \leq 4\|\bar{y}_k\|_2^2/3$ in (8.8), and using this in (8.9) gives the desired bound:

$$\|\tilde{r}_k(\bar{y}_k)\|_2 \leq \tilde{\gamma}_{kn}(\|b\|_2^2 + \|A\|_F^2\|\bar{y}_k\|_2^2)^{\frac{1}{2}} \leq \tilde{\gamma}_{kn}(\|b\|_2 + \|A\|_F\|\bar{y}_k\|_2). \tag{8.12}$$

But we compute $\bar{x}_j = fl(\bar{V}_j\bar{y}_j)$, not $\tilde{V}_j\bar{y}_j$, so to complete this analysis, we have to show that $\bar{x}_k$ is a backward stable solution for (1.1). Now, see (4.3), $\bar{x}_k = fl(\bar{V}_k\bar{y}_k) = (\bar{V}_k + \Delta V'_k)\bar{y}_k$ with $|\Delta V'_k| \leq \gamma_k|\bar{V}_k|$. With $\Delta\tilde{V}_k(y)$ in (8.2) define

$$\Delta A_k \equiv [\Delta\tilde{V}_k(\bar{y}_k) - A(\Delta V'_k + \bar{V}_k - \tilde{V}_k)]\bar{y}_k\|\bar{x}_k\|_2^{-2}\bar{x}_k^T,$$

so that $(A + \Delta A_k)\bar{x}_k = (A\tilde{V}_k + \Delta\tilde{V}_k(\bar{y}_k))\bar{y}_k$, and, see (8.1), (8.2), (2.2),

$$\|b+\Delta b_k(\bar{y}_k)-(A + \Delta A_k)\bar{x}_k\|_2 = \min_y \|b+\Delta b_k(y)-[A\tilde{V}_k+\Delta\tilde{V}_k(y)]y\|_2, \tag{8.13}$$

$$\|\Delta b_k(\bar{y}_k)\|_2 \leq \tilde{\gamma}_{kn}\|b\|_2,$$

$$\|\Delta A_k\|_F \leq [\|\Delta\tilde{V}_k(\bar{y}_k)\|_F+\|A(\Delta V'_k+\tilde{V}_k\Delta_k)\|_F]\|\bar{y}_k\|_2/\|\bar{x}_k\|_2, \tag{8.14}$$

where we know from (8.12) that (8.13) is bounded by $\tilde{\gamma}_{kn}(\|b\|_2 + \|A\|_F\|\bar{y}_k\|_2)$. But $\|\Delta V'_k\|_F \leq k^{\frac{1}{2}}\gamma_k$, so from (2.2) $\|A(\Delta V'_k+\tilde{V}_k\Delta_k)\|_F \leq k^{\frac{1}{2}}\tilde{\gamma}_n\|A\|_2$, and from (8.2) $\|\Delta\tilde{V}_k(\bar{y}_k)\|_F \leq \tilde{\gamma}_{kn}\|A\|_F$, so with (2.2) and (8.4)

$$\|\bar{x}_k\|_2 = \|(\bar{V}_k + \Delta V'_k)\bar{y}_k\|_2 \geq \|\bar{V}_k\bar{y}_k\|_2 - \|\Delta V'_k\|_F\|\bar{y}_k\|_2 \geq \|\bar{y}_k\|_2(3/4 - k^{\frac{1}{2}}\gamma_n).$$

Combining these with (8.1) shows that $\|\Delta A_k\|_F \leq \tilde{\gamma}_{kn}\|A\|_F$ in (8.14). Summarizing:

$$\begin{aligned} &\tilde{r}_k(\bar{y}_k) = b+\Delta b_k(\bar{y}_k)-(A + \Delta A_k)\bar{x}_k, \quad \|\tilde{r}_k(\bar{y}_k)\|_2 \leq \tilde{\gamma}_{kn}(\|b\|_2 + \|A\|_F\|\bar{x}_k\|_2), \\ &\|\Delta b_k(\bar{y}_k)\|_2 \leq \tilde{\gamma}_{kn}\|b\|_2, \qquad \|\Delta A_k\|_F \leq \tilde{\gamma}_{kn}\|A\|_F. \end{aligned} \tag{8.15}$$

Using the usual approach of combining (8.15) with the definitions

$$\begin{aligned} &\Delta b'_k \equiv -\frac{\|b\|_2}{\|b\|_2 + \|A\|_F\|\bar{x}_k\|_2}\tilde{r}_k(\bar{y}_k), \quad \Delta A'_k \equiv \frac{\|A\|_F\|\bar{x}_k\|_2}{\|b\|_2 + \|A\|_F\|\bar{x}_k\|_2}\frac{\tilde{r}_k(\bar{y}_k)\bar{x}_k^T}{\|\bar{x}_k\|_2^2}, \\ \text{shows} \quad &(A + \Delta A_k + \Delta A'_k)\bar{x}_k = b + \Delta b_k(\bar{y}_k) + \Delta b'_k, \\ &\|\Delta A_k + \Delta A'_k\|_F \leq \tilde{\gamma}_{kn}\|A\|_F, \quad \|\Delta b_k(\bar{y}_k) + \Delta b'_k\|_2 \leq \tilde{\gamma}_{kn}\|b\|_2, \end{aligned}$$

proving that the MGS-GMRES solution $\bar{x}_k$ is backward stable for (1.1).

9. Comments and conclusions. The form of the restriction in (1.1) suggests that we might be able to ease this restriction somewhat by using $\tilde{\kappa}_F(A)$ as defined in (2.1), instead of $\|A\|_F/\sigma_{min}(A)$ in (1.1). However $\tilde{\kappa}_F(B_j)$ was useful when we applied MGS to B_j, see for example (5.7), while in MGS-GMRES we apply MGS to $[b, AV_{j-1}]$, so it looks like we cannot get an *a priori* restriction involving $\tilde{\kappa}_F(A)$ this way. See also Remark 4.1. Appendix A discusses a possibly superior way of meeting the restriction in (1.1) for difficult problems.

Now to conclude this. Among many other things, we showed that MGS-GMRES

- gives a backward stable least squares solution at every step, (Section 8.1);
- obtains a backward stable solution to the problem (1.1), (Section 8.2);
- and up until this point $\kappa_2(\tilde{V}_m) \leq 4/3$, (Section 6).

Thus we can say that the MGS–GMRES method is backward stable for computing the solution x to $Ax = b$ for sufficiently nonsingular A, answering an important open question. Despite loss of orthogonality, it provides an acceptable solution within $n+1$ MGS steps (n steps of MGS-GMRES). The loss of orthogonality is usually inversely proportional to the level of convergence. Complete loss of orthogonality implies a solution exists, and MGS-GMRES necessarily finds this under reasonable restrictions (1.1) (or more practically but less rigorously (1.2)) on the problem. From this we see that the numerical behavior is far better than was often thought. This means we do not have to do anything special to ameliorate the effect of rounding errors — we certainly do not need reorthogonalization — and need only concentrate on finding solutions more quickly, mainly by seeking better preconditioning techniques.

The final proof was seen to require an instance of a more general result on the backward stability of a variant of the MGS algorithm applied to a matrix B in order to solve a linear least squares problem, see Section 7.1. In Section 5 we showed more precisely than before how orthogonality could be lost in the MGS algorithm, in particular by using the condition number $\tilde{\kappa}_F(B)$ defined in (2.1).

Acknowledgments. The main approach here was to base the analysis on the surprising relationship between MGS and the Householder reduction of an augmented matrix that was discovered by Charles Sheffield and proven and developed by Björck and Paige in [5], and combine this with the elegant result discovered by Giraud and Langou in [10] (responding to a request by Mario Arioli). Once we had made that choice the task was still extremely difficult, and we had to draw on many other works as well — among these the excellent book by Higham [13] facilitated our work greatly.

This paper is the end result of a long term collaboration of its three authors aimed at producing a rounding error analysis of the MGS-GMRES method. And although this is unusual, the second and third authors (alphabetically) would like to thank the first author for carrying this project to such a satisfying conclusion.

Two referees' comments added nicely to the history and precision of the paper.

Appendix A. Condition numbers. If $\kappa_F(A) \equiv \|A\|_F/\sigma_{min}(A)$, then (2.1) is

$$\tilde{\kappa}_F(A) \equiv \min_{\text{diagonal } D > 0} \kappa_F(AD).$$

For $m \times n$ A, if positive diagonal $\widetilde{D}$ is such that in $A\widetilde{D}$ all columns have equal 2-norm, then van der Sluis [21, Thm. 3.5, (b)] showed that $\kappa_F(A\widetilde{D})$ is no more than a factor $\sqrt{n}$ away from its minimum (here $\tilde{\kappa}_F(A)$), and this is the first mention of the condition number $\kappa_F(A)$ (and at least by implication, of $\tilde{\kappa}_F(A)$) that we have seen so far. He also stated in [22, §3.9] that if $\|\delta A e_j\| < \|Ae_j\|/\kappa_F(A)$ for $j = 1, \ldots, n \le m$, then $A + \delta A$ has full rank n. This is easy to see since it ensures that $\|\delta A\|_F < \sigma_{min}(A)$. He also points out that this is in some sense tight, in that if $\|\delta A e_j\| = \|Ae_j\|/\kappa_F(A)$ for $j = 1, \ldots, n \le m$ is allowed, then for any prescribed value of $\kappa_F(A) \ge \sqrt{n}$ there exist A and δA such that $A + \delta A$ is rank deficient. Since the backward error bounds in this paper were obtained column by column, see Lemma 3.2 and for example the column bounds in (8.1), this suggests that the form of the restriction in (1.1) is optimal, even down to the factor $n^2\epsilon$. See also the first paragraph of Section 4.

Moreover, instead of solving (1.1) we can solve $(AD)y = b$ for some positive diagonal D, and then form $x = Dy$. By taking $D = \widetilde{D}$ above we see from van der Sluis's theory that we can approach the value of $\tilde{\kappa}_F(A)$ with $\kappa_F(A\widetilde{D})$, and perhaps alter a problem with an ill-conditioned A so that it meets the restriction (1.1). This

is another justification for using such a $\widetilde{D}$ as a basic simple preconditioner when MGS-GMRES is applied to ill-conditioned problems.

REFERENCES

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

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

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

[4] Å. Björck, *Numerical Methods for Least Squares Problems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1996.

[5] Å. 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.

[6] P. Brown and H. Walker, *GMRES on (Nearly) Singular Systems*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 37–51.

[7] J. Demmel, Y. Hida, W. Kahan, X. S. Li, S. Mukherjee and E. J. Riedy, *Error Bounds from Extra Precise Iterative Refinement, ACM Transactions on Mathematical Software*, to appear. U. C. Berkeley Technical Report UCB/CSD-04-1344, (2005).

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

[9] 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.

[10] L. Giraud and J. Langou, *When modified Gram-Schmidt generates a well-conditioned set of vectors*, IMA Journal on Numerical Analysis, 22 (2002), pp. 521–528.

[11] G. H. Golub and C. F. Van Loan, *Matrix Computations,* 3rd Edn., The Johns Hopkins University Press, Baltimore, Maryland, 1996.

[12] A. Greenbaum, M. Rozložník, and Z. Strakoš, *Numerical behavior of the modified Gram-Schmidt GMRES implementation*, BIT, 37 (1997), pp. 706–719.

[13] N. J. Higham, *Accuracy and Stability of Numerical Algorithms, 2nd Edn.*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2002.

[14] J. Langou, *Iterative methods for solving linear systems with multiple right-hand sides*, Ph.D. Thesis, Institut Nationales des Sciences Appliquées de Toulouse, 2003.

[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 and M. A. Saunders, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software, 8 (1982), pp. 43–71.

[17] C. C. Paige and Z. Strakoš, *Bounds for the least squares distance using scaled total least squares*, Numerische Mathematik, 91 (2002), pp. 93–115.

[18] C. C. Paige and Z. Strakoš, *Residual and Backward Error Bounds in Minimum Residual Krylov Subspace Methods*, SIAM J. Sci. Comput., 23, (2002), pp. 1899-1924.

[19] M. Rozložník, *Numerical Stability of the GMRES Method*, Ph.D. Thesis, Institute of Computer Science, Academy of Sciences, Prague, 1997.

[20] 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.

[21] A. van der Sluis, *Condition numbers and equilibration matrices*, Numer. Math., 14 (1969), pp. 14–23.

[22] A. van der Sluis, *Stability of the solutions of linear least squares problems*, Numer. Math., 23 (1975), pp. 241–254.

[23] L. Smoch, *Some Results about GMRES in the Singular Case*, Numerical Algorithms, 22 (1999), pp. 193–212.

[24] 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.

[25] H. F. Walker, *Implementation of the GMRES method*, J. Comput. Phys., 53 (1989), pp. 311–320.

[26] J. H. Wilkinson, *The Algebraic Eigenvalue Problem,* Clarendon Press, Oxford, 1965.