BIT 35 (1995), 309–330

NUMERICAL STABILITY OF GMRES *

J. DRKOŠOVÁ,[1] A. GREENBAUM,[2] M. ROZLOŽNÍK[1] and Z. STRAKOŠ[1] †

[1] *Institute of Computer Science, Academy of Sciences of the Czech Republic*
Pod vodárenskou věží 2, 182 07 Praha 8, Czech Republic
e-mail: jitka@uivt.cas.cz, miro@uivt.cas.cz, strakos@uivt.cas.cz

[2] *Courant Institute of Mathematical Sciences, 251 Mercer Street, New York*
NY 10012, U.S.A. e-mail: greenbau@greenbaum.cims.nyu.edu

Abstract.

The Generalized Minimal Residual Method (GMRES) is one of the significant methods for solving linear algebraic systems with nonsymmetric matrices. It minimizes the norm of the residual on the linear variety determined by the initial residual and the n-th Krylov residual subspace and is therefore optimal, with respect to the size of the residual, in the class of Krylov subspace methods. One possible way of computing the GMRES approximations is based on constructing the orthonormal basis of the Krylov subspaces (Arnoldi basis) and then solving the transformed least squares problem. This paper studies the numerical stability of such formulations of GMRES. Our approach is based on the Arnoldi recurrence for the actually, i.e., in finite precision arithmetic, computed quantities. We consider the Householder (HHA), iterated modified Gram-Schmidt (IMGSA), and iterated classical Gram-Schmidt (ICGSA) implementations. Under the obvious assumption on the numerical nonsingularity of the system matrix, the HHA implementation of GMRES is proved backward stable in the normwise sense. That is, the backward error $\|b - A\tilde{x}\|/(\|A\|\|\tilde{x}\| + \|b\|)$ for the approximation $\tilde{x}$ is proportional to machine precision ε. Additionally, it is shown that in most cases the norm of the residual computed from the transformed least squares problem (Arnoldi residual) gives a good estimate of the true residual norm, until the true residual norm has reached the level $\varepsilon\|A\|\|x\|$.

AMS subject classification: 65F05.

Key words: Linear algebraic systems, numerical stability, backward error, GMRES method.

1 Introduction.

The Generalized Minimal Residual method (GMRES) is an important iterative method for solving nonsymmetric linear systems

$$Ax = b. \tag{1.1}$$

Assume that A is a real nonsingular N by N matrix and b a real vector. Given an initial guess x^0 for the solution, the algorithm generates approximate solutions

*Received February 1994. Revised June 1995.
†This work was supported by NSF contract Int921824.

x^n, $n = 1, 2, \ldots, N$ from the linear variety

$$x^0 + K_n(A, r^0) \tag{1.2}$$

minimizing the Euclidean norm of the residual,

$$\|b - Ax^n\| = \min_{u \in x^0 + K_n(A,r^0)} \|b - Au\|, \tag{1.3}$$

where $r^0 = b - Ax^0$ is the initial residual and $K_n(A, r^0)$ is the n-th Krylov subspace generated by A, r^0,

$$K_n(A, r^0) = span\{r^0, Ar^0, \ldots, A^{n-1}r^0\}. \tag{1.4}$$

Clearly, $r^n \in r^0 + AK_n(A, r^0)$. We call $AK_n(A, r^0)$ a Krylov residual subspace. Such an approximation always exists and is unique. It can be computed in many different ways which are reviewed, e.g., in [11]. For another description of GMRES variants and their relations to other methods we refer to the forthcoming paper [23].

Most methods for computing the approximation x^n satisfying (1.3) start by constructing an orthonormal basis, called an Arnoldi basis, for the Krylov subspaces (1.4). The recurrence for the basis vectors can be written in matrix form as

$$AV_n = V_{n+1}H_{n+1,n}, \tag{1.5}$$

where V_{n+1} is the N by $(n + 1)$ matrix with the orthonormal basis vectors $v_1, v_2, \ldots, v_{n+1}$ as its columns and $H_{n+1,n}$ is the $(n+1)$ by n upper Hessenberg matrix of the orthogonalization (and normalization) coefficients, $n < N$. The initial vector v_1 is r^0/ϱ, where $\varrho = \|r^0\|$. The approximate solution x^n is then taken to be of the form $x^n = x^0 + V_n y^n$, where y^n is chosen to minimize

$$\begin{aligned} \|b - Ax^n\| &= \|r^0 - AV_n y^n\| \\ &= \|V_{n+1}(\varrho e_1 - H_{n+1,n}y^n)\| \\ &= \|\varrho e_1 - H_{n+1,n}y^n\|. \end{aligned}$$

The problem of solving approximately the original N-dimensional system $Ax = b$ is thus transformed to the n-dimensional least squares problem

$$\|\varrho e_1 - H_{n+1,n}y^n\| = \min_y \|\varrho e_1 - H_{n+1,n}y\|, \quad x^n = x^0 + V_n y^n. \tag{1.6}$$

We call $b - Ax^n$ a true residual, $\rho e_1 - H_{n+1,n}y^n$ an Arnoldi residual. While the norms of these two vectors are the same in exact arithmetic, we will see later that they may differ in finite precision arithmetic.

Several algorithms for computing the Arnoldi basis exist. They differ in the way in which the orthogonalization in the Arnoldi process is carried out. Utilizing [24], [27], [15] and [3], we will consider the column by column modified Gram-Schmidt (MGS), Householder (HH) and iterated classical, resp. modified, Gram-Schmidt (ICGS, IMGS) orthogonalizations.

In the variants based on the Gram-Schmidt method, a new basis vector v_{n+1} is the normalized result of the orthogonalization of the vector Av_n against the previously computed vectors $v_1, \cdots, v_n$. A detailed description of the MGS, ICGS and IMGS algorithms can be found in the thorough survey paper [3]. For the construction of the Arnoldi basis, the algorithms can be written in the following form

MGS-Arnoldi algorithm (MGSA)

$\varrho = \|r^0\|$
$v_1 = r^0/\varrho$
for $i = 1, 2, \ldots, n$
  $w = Av_i$
  for $k = 1, 2, \ldots, i$
  $\varrho_{k,i} = v_k^T w$
  $w = w - \varrho_{k,i} v_k$
  end
  $\varrho_{i+1,i} = \|w\|$
  $v_{i+1} = w/\varrho_{i+1,i}$
end

IMGS-Arnoldi algorithm (IMGSA)

$\varrho = \|r^0\|$
$v_1 = r^0/\varrho$
for $i = 1, 2, \ldots, n$
  for $k = 1, 2, \ldots, i$
    $\varrho_{k,i} = 0$
  end
  $w = Av_i$
  repeat
    $\omega = \|w\|$
    for $k = 1, 2, \ldots, i$
      $\nu_{k,i} = v_k^T w$
      $w = w - \nu_{k,i} v_k$
      $\varrho_{k,i} = \varrho_{k,i} + \nu_{k,i}$
    end
  until $\|w\| > \omega/\gamma$
  {for some $\gamma > (0.83 - \varepsilon)^{-1}$}
  $\varrho_{i+1,i} = \|w\|$
  $v_{i+1} = w/\varrho_{i+1,i}$
end

ICGS-Arnoldi algorithm (ICGSA)

$\varrho = \|r^0\|$
$v_1 = r^0/\varrho$
for $i = 1, 2, \ldots, n$
  for $k = 1, 2, \ldots, i$
    $\varrho_{k,i} = 0$
  end
  $w = Av_i$
  repeat
    $\omega = \|w\|$
    $h = V_i^T w = (\eta_1, \ldots, \eta_i)^T$
    $w = w - V_i h$
    for $k = 1, \ldots, i$
      $\varrho_{k,i} = \varrho_{k,i} + \eta_k$
    end
  until $\|w\| > \omega/\gamma$
  {for some $\gamma > (0.83 - \varepsilon)^{-1}$}
  $\varrho_{i+1,i} = \|w\|$
  $v_{i+1} = w/\varrho_{i+1,i}$
end

where ε denotes the machine precision unit. For a discussion of the choice of the parameter γ, see [15] and the papers referred to there. Ignoring rounding errors, the recurrence for the resulting basis V_{n+1} can be written in matrix form as

$$(v_1, AV_n) = V_{n+1} R_{n+1}, \tag{1.7}$$

where R_{n+1} is the $(n+1)$ by $(n+1)$ upper triangular matrix with the orthogonalization coefficients in its strict upper triangular part, and the normalization coefficients on the diagonal.

An implementation based on HH orthogonalization was proposed by Walker [27]. In this variant, a new basis vector v_{n+1} is produced as the $(n+1)$-th column of the product of the $(N \times N)$ elementary Householder matrices,

$$v_{n+1} = P_1 P_2 \cdots P_{n+1} e_{n+1}; \quad v_1 = r^0/\varrho. \tag{1.8}$$

The matrices P_i are determined by $P_1 = I - 2s_1 s_1^T$, $P_1 v_1 = e_1$, and

$$\begin{aligned} P_i = I - 2s_i s_i^T, \qquad & s_i = (0, \cdots, 0, \omega_{i,i}, \omega_{i,i+1}, \cdots, \omega_{i,N})^T, \\ P_i \left(P_{i-1} P_{i-2} \cdots P_1 A v_{i-1}\right) = & \ (\varrho_{1,i-1}, \varrho_{2,i-1}, \cdots, \varrho_{i,i-1}, 0, \cdots, 0)^T, \\ & i = 2, \cdots, n+1. \end{aligned} \tag{1.9}$$

where $\|s_i\| = 1$, $i = 1, \cdots, n+1$. In matrix form,

$$P_{n+1} P_n \cdots P_1 (v_1, AV_n) = \begin{pmatrix} R_{n+1} \\ 0 \end{pmatrix}, \tag{1.10}$$

$$(v_1, AV_n) = V_{n+1} R_{n+1}, \quad V_{n+1} = P_1 P_2 \cdots P_{n+1} \begin{pmatrix} I_{n+1} \\ 0 \end{pmatrix}, \tag{1.11}$$

where R_{n+1} is, as before, an $(n+1)$ by $(n+1)$ upper triangular matrix, I_{n+1} is the $(n+1)$ by $(n+1)$ identity matrix. The elementary Householder matrix P_i operates only on the last $N-(i-1)$ elements of any vector. We will refer to (1.8) – (1.11) as the algorithm HHA.

For all the implementations described above, the Arnoldi process in steps 1 through $n+1$ can be viewed as a recursive QR decomposition of the matrix (v_1, AV_n). The upper Hessenberg matrix $H_{n+1,n}$ satisfies

$$R_{n+1} = (e_1 \, , H_{n+1,n}), \tag{1.12}$$

and its columns h_i are given by

$$h_i = (\varrho_{1,i}, \varrho_{2,i}, \cdots, \varrho_{i,i}, \varrho_{i+1,i}, 0, \cdots, 0)^T, \quad i = 1, \cdots, n.$$

Though all the orthogonalizations described above are mathematically equivalent, they may have different numerical properties.

This paper studies the numerical stability of the GMRES formulations described above. Some aspects of this problem have already been studied by Karlson [16]. His work is based on forward error estimates and contains several interesting ideas and experimental observations, but, unfortunately, it does not give clear, well justified conclusions. Backward stability of several related iterative methods for both solution of linear systems and computing eigenvalues was studied experimentally by Chatelin, Frayssé, Godet-Thobie and Braconnier [9], [10], [6], [7]. Our work was partially motivated by the open questions related

to these papers. For concepts of stability and many other relevant results we refer to a recent paper by Higham and Knight [14].

Our paper is organized as follows. In Section 2, the Arnoldi recurrence for the quantities actually computed in finite precision arithmetic is described. In Section 3, the relation between the true and Arnoldi residuals in the presence of rounding errors is analyzed. It is shown that the difference between the norms of the true and Arnoldi residuals can become large only if the solution y^n of the transformed least squares problem (1.6) is large in norm. Assuming that the orthogonality of the Arnoldi basis is well preserved, this can happen only when the Arnoldi residual is fairly large compared to its final value. If the relative norm of the Arnoldi residual reaches the level of machine precision ε it is shown that the norm of y^n is of the same order as the norm of x and it follows that the relative norm of the true residual is also of order ε. This motivates questions about the ultimate (final) accuracy of GMRES measured by the norm of the true residual and about the backward stability of GMRES. Backward stability is examined in Section 4. It is shown that the HHA implementation of GMRES is normwise backward stable in the usual sense (see e.g., [14]). Section 5 contains concluding remarks and indications for further work.

Throughout this paper, we assume that for a given matrix A and initial residual r^0 none of the algorithms HHA, MGSA, IMGSA, ICGSA terminates until step N. In other words, we assume that the steps from 1 to N of all these algorithms are well defined, both mathematically (i.e., in exact arithmetic) and numerically (i.e., in the presence of rounding errors). This assumption is rather technical and does not cause any loss of generality. For a more general situation the statements can be modified accordingly. For convenience, most of the statements will be formulated for $n < N$ throughout the text. It is obvious that for the final step $n = N$ these statements can be easily reformulated. We will return to this point in Section 4.

A remark on the notation. Studying effects of rounding errors, we will follow the notation used by Paige in his papers on the symmetric Lanczos method [19], [20], [21]. By $r^0, \varrho, P_n, R_n, V_n, H_{n+1,n}, H_{N,N}, S_n, Z_n, y^n$ and x^n we denote from now on the *actually computed quantities*. No confusion is possible because nowhere will the computed quantities be mixed together with the corresponding exact precision counterparts. (One must, of course, distinguish between the results of floating point operations and the mathematical expressions used in developing error bounds. For example, the computed initial residual

$$r^0 = fl(b - fl(Ax^0)),$$

where $fl(.)$ denotes the floating point result of the operation (.), is not equal to the mathematical expression $b - Ax^0$).

Throughout the paper, $\|Z\|$ denotes the 2-norm and $\|Z\|_F$ the Frobenius norm of the matrix Z; $\|z\|$ denotes the Euclidean norm of a vector z. Let X be an m by l matrix, $m \geq l$, with full column rank. Then $\sigma_k(X)$, $k = 1, 2, \ldots, l$ denote the singular values of X in descending order,

$$\sigma_1(X) \geq \sigma_2(X) \geq \ldots \geq \sigma_l(X) > 0.$$

The condition number of X is denoted by $\kappa(X) \equiv \sigma_1(X)/\sigma_l(X)$.

In the bounds we present only those terms which are linear in ε and do not account for the terms proportional to higher powers of ε. Constant factors are introduced in a number of places. These are all universal constants, independent of the problem parameters (e.g., N, $\kappa(A)$, etc.) and the machine precision ε, but dependent on the precise details of the machine arithmetic.

2 Finite precision Arnoldi recurrence.

Because of rounding errors, the computed quantities do not satisfy the exact recurrence (1.5). Assuming that a variant based on the Gram-Schmidt orthogonalization (MGSA, ICGSA, IMGSA) or the Householder orthogonalization (HHA) is used for computing the vectors $v_1, v_2, \cdots, v_n$ of the Arnoldi basis, we find the analogue of (1.5) for the finite precision results. It is an extension of Paige's formula for the 3-term recurrences in the symmetric case [20].

In the Introduction, the computation of the Arnoldi basis was described as a column-oriented recursive QR decomposition of the matrix (v_1, AV_n). As a consequence, the rounding error analysis developed by Wilkinson and Björck directly applies to the case of HHA and MGSA. Using [30], pp. 152 - 161, 236, 382 - 388 and [1], we have for both HHA and MGSA the recurrence formula for the computed quantities

$$(v_1, fl(Av_1), \ldots, fl(Av_n)) = V_{n+1}R_{n+1} + F_{0,n}, \tag{2.1}$$

i.e., rewriting it for the Arnoldi recurrence,

$$(fl(Av_1), \ldots, fl(Av_n)) = AV_n + F_{A,n},$$

$$AV_n = V_{n+1}H_{n+1,n} + F_n. \tag{2.2}$$

For $\|F_n\|$ we have the bound

$$\|F_n\| \leq \|F_{0,n}\| + \|F_{A,n}\| \leq O(n^{3/2}N)\varepsilon\|A\| + O(n^{1/2}lN^{1/2})\varepsilon\|A\|$$

where l is the maximal number of nonzero entries per row of the matrix A and $O(m)$ stands for a quantity proportional to the argument m ([30], [1]). We denote $\eta(n,l,N) = (n^{3/2}N + n^{1/2}lN^{1/2})$. Then

$$\|F_n\| \leq \zeta_1(n^{3/2}N + n^{1/2}lN^{1/2})\varepsilon\|A\| = \zeta_1\eta(n,l,N)\varepsilon\|A\| \leq \zeta_1' nN^{3/2}\varepsilon\|A\| \tag{2.3}$$

for some positive constants ζ_1 and ζ_1', independent of the problem parameters. It is easy to see that the equations (2.1)-(2.3) hold true even for the IMGSA and ICGSA algorithms, assuming that for every index i the number of reorthogonalization steps is bounded by a small constant independent of n and i.

Loss of orthogonality for the HHA implementation is bounded independently of the properties of the matrix (v_1, AV_n)

$$(HHA): \quad \|I - V_n^T V_n\| \leq \zeta_2 n^{3/2}N\varepsilon, \tag{2.4}$$

while for the MGSA implementation the loss of orthogonality is bounded in terms of the condition number $\kappa(v_1, AV_n)$. Assuming that $nN^2\varepsilon\,\kappa(v_1, AV_n) \ll 1$, we can write

$$(2.5)\,(MGSA): \quad \|I - V_n^T V_n\| \le O(n^2N)\varepsilon\kappa(v_1, AV_n) \le \zeta_3 n^2 N\varepsilon\kappa(v_1, AV_n).$$

Moreover, it follows from [30] and [4], see also [3], that for both the HHA and MGSA implementations there exists an exactly orthonormal matrix $\hat{V}_{n+1}$ such that

$$AV_n = \hat{V}_{n+1}H_{n+1,n} + \tilde{F}_n, \tag{2.6}$$

$$\|\tilde{F}_n\| \le \zeta_4\eta(n,l,N)\varepsilon\|A\| \le \zeta_4' nN^{3/2}\varepsilon\|A\|. \tag{2.7}$$

For the HHA implementation, the matrix V_{n+1} is close to $\hat{V}_{n+1}$,

$$(HHA): \quad \|V_{n+1} - \hat{V}_{n+1}\| \le \zeta_5 n^{3/2}N\varepsilon. \tag{2.8}$$

For the MGSA implementation, the analogous bound depends again on the quantity $\kappa(v_1, AV_n)$. Supposing that $nN^{3/2}\varepsilon\,\kappa(v_1, AV_n) \ll 1$, we have

$$(MGSA): \quad \|V_{n+1} - \hat{V}_{n+1}\| \le \zeta_6\eta(n,l,N)\varepsilon\kappa(v_1, AV_n). \tag{2.9}$$

We defer the discussion of the IMGSA, ICGSA analogues of (2.5)-(2.9) to Section 3. The positive constants $\zeta_1, \zeta_2, \ldots, \zeta_6$ depend on the details of the arithmetic ([30], [1], [4]). We assume, for simplicity, $\zeta_j \ge 1$, $j = 1, 2, \ldots, 6$.

Formulas (2.2) and (2.3) suggest an explanation for the deterioration effects of rounding errors to the Krylov subspace methods which were observed for very ill-conditioned matrices in [10], [7], for example. Suppose that for some actual finite precision run the size of $\|F_n\|$ in (2.2) is close to the given bound (2.3). If, at the same time, the entries in $H_{n+1,n}$ are much less in magnitude than $\|A\|$, then any method, for which the construction of the approximate solution is based on the matrix $H_{n+1,n}$, will suffer from loss of accuracy of the computed result due to relatively large elements in F_n. It is clear that any Krylov subspace method may suffer from this difficulty. For the GMRES method, the situation will be quantified in detail in the next section.

3 Arnoldi and true residuals.

Due to rounding errors, the actual norm of the n-th GMRES residual $\|b - Ax^n\|$ can differ from the norm of the Arnoldi residual $\|\varrho e_1 - H_{n+1,n}y^n\|$. We will look for an expression bounding this difference. The n-th approximate solution x^n is given as a finite precision solution of (1.6), i.e.,

$$x^n = x^0 + V_n y^n + d_n, \tag{3.1}$$

where d_n substitutes for local rounding errors in computing x^n from x^0, V_n and y^n. Using (2.2),

$$\begin{aligned} b - Ax^n &= b - Ax^0 - AV_n y^n - Ad_n \\ &= b - Ax^0 - V_{n+1}H_{n+1,n}y^n - F_n y^n - Ad_n \\ &= (b - Ax^0 - \varrho v_1) + (\varrho v_1 - V_{n+1}H_{n+1,n}y^n) - F_n y^n - Ad_n \\ &= (b - Ax^0 - \varrho v_1) + V_{n+1}(\varrho e_1 - H_{n+1,n}y^n) - F_n y^n - Ad_n. \end{aligned}$$

Consequently,

$$(3.2)\quad (b - Ax^n) - V_{n+1}(\varrho e_1 - H_{n+1,n}y^n) = (b - Ax^0 - \varrho v_1) - F_n y^n - Ad_n.$$

Taking norms in (3.2),

$$(3.3)\quad \begin{aligned} \|(b - Ax^n) - V_{n+1}(\varrho e_1 - H_{n+1,n}y^n)\| \;&\leq\; \|b - Ax^0 - \varrho v_1\| \\ &+\; \|F_n y^n\| + \|Ad_n\|. \end{aligned}$$

We give bounds for the terms on the right hand side. Following the elementary rounding error relations in [20], [21] and using (2.3),

$$\begin{aligned} \|b - Ax^0 - \varrho v_1\| &\leq O(lN^{1/2} + N)\varepsilon\|A\|\|x^0\| + O(N)\varepsilon\|b\|, \\ \|Ad_n\| &\leq O(n^{3/2})\varepsilon\|A\|\|y^n\| + \varepsilon\|A\|\|x^0\|, \\ \|F_n y^n\| &\leq \zeta_1\eta(n,l,N)\varepsilon\|A\|\|y^n\|, \end{aligned}$$

which gives

$$(3.4)\quad \begin{aligned} &\|(b - Ax^n) - V_{n+1}(\varrho e_1 - H_{n+1,n}y^n)\| \\ &\qquad \leq \zeta_7\eta(n,l,N)\varepsilon\|A\|\|y^n\| + \nu(A,b,x^0), \end{aligned}$$

where

$$\nu(A,b,x^0) = O(lN^{1/2} + N)\varepsilon\|A\|\|x^0\| + O(N)\varepsilon\|b\|.$$

Assuming that $\|x^0\|$ is not extremely large, $\nu(A,b,x^0)$ is not significant; the bound in (3.4) can become large if and only if $\|y^n\|$ becomes large.

It follows that we need to find a bound for $\|y^n\|$. There are two possible approaches for bounding $\|y^n\|$, one that is useful in general and one that is more appropriate when the algorithm is near convergence. We first obtain a bound on $\|y^n\|$, valid at all steps n, using the fact that y^n is the computed solution to the least squares problem (1.6).

The vector y^n is computed from the transformed least squares problem using the QR decomposition of the matrix $H_{n+1,n}$ via Givens rotations [24]. From the rounding error analysis of the Givens transformation ([30], pp. 131 - 139), there exists an exactly orthogonal matrix $\hat{Z}_{n+1}$ of dimension $(n+1)$ by $(n+1)$ such that

$$(3.5)\quad \hat{Z}_{n+1}H_{n+1,n} = \begin{pmatrix} S_n \\ 0 \end{pmatrix} + G_n,$$

$$(3.6)\quad \|G_n\| \leq \zeta_8 n^{3/2}\varepsilon\|H_{n+1,n}\|,$$

where the n by n matrix S_n is the upper triangular matrix actually computed in finite precision arithmetic and the constant $\zeta_8 \geq 1$ depends on the details of the arithmetic. From (3.5), (3.6) and the perturbation theory of singular values ([12], p. 428) it follows that

$$(3.7)\quad |\sigma_k(H_{n+1,n}) - \sigma_k(S_n)| \leq \zeta_8 n^{3/2}\varepsilon\|H_{n+1,n}\|, \qquad k = 1, 2, \ldots, n.$$

In the following text we assume that

$$\sigma_n(H_{n+1,n}) > 2\zeta_8 n^{3/2}\varepsilon\|H_{n+1,n}\|. \tag{3.8}$$

Then (3.7) implies

$$\sigma_n(S_n) = (\|S_n^{-1}\|)^{-1} \geq \sigma_n(H_{n+1,n}) - \zeta_8 n^{3/2}\varepsilon\sigma_1(H_{n+1,n}), \tag{3.9}$$

$$\begin{aligned}\sigma_1(S_n) &= \|S_n\| \leq (1+\zeta_8 n^{3/2}\varepsilon)\sigma_1(H_{n+1,n}) \\ &= (1+\zeta_8 n^{3/2}\varepsilon)\|H_{n+1,n}\|.\end{aligned} \tag{3.10}$$

Using the backward error analysis for back substitution (see [12], pp. 88 - 89), we have

$$(S_n + \tilde{G}_n)y^n = \varrho\hat{Z}_{n+1}e_1 + \bar{g}_n,$$

where $\|\tilde{G}_n\| \leq n\varepsilon\|S_n\|$, $\|\bar{g}_n\| \leq \varrho\zeta_9 n\varepsilon$, which gives

$$\begin{aligned}\|y^n\| &\leq \varrho(1+\zeta_9 n\varepsilon)[\sigma_n(S_n+\tilde{G}_n)]^{-1} \\ &\leq \varrho(1+\zeta_9 n\varepsilon)[\sigma_n(H_{n+1,n}) - 2\zeta_8 n^{3/2}\varepsilon\sigma_1(H_{n+1,n})]^{-1} + O(\varepsilon^2).\end{aligned} \tag{3.11}$$

Summarizing, we can state the following Lemma:

LEMMA 3.1. *Assuming (3.8), the following inequality holds*

$$\begin{aligned}&\|(b-Ax^n) - V_{n+1}(\varrho e_1 - H_{n+1,n}y^n)\| \\ &\quad\leq \zeta_{10}\eta(n,l,N)\varepsilon\varrho\|A\|[\sigma_n(H_{n+1,n})]^{-1}[1-2\zeta_8 n^{3/2}\varepsilon\kappa(H_{n+1,n})]^{-1} \\ &\qquad + \nu(A,b,x^0).\end{aligned} \tag{3.12}$$

PROOF. Combining (3.4) with (3.11) gives the statement. □

Lemma 3.1 still does not offer satisfactory insight into the problem of monitoring the convergence of GMRES by the norm of the Arnoldi residual. We need an estimate for the difference between the norms of the true and Arnoldi residuals

$$|\,\|b-Ax^n\| - \|\varrho e_1 - H_{n+1,n}y^n\|\,|.$$

Considering

$$|\,\|b-Ax^n\| - \|V_{n+1}(\varrho e_1 - H_{n+1,n}y^n)\|\,| \leq \|(b-Ax^n) - V_{n+1}(\varrho e_1 - H_{n+1,n}y^n)\|,$$

we must ask for the difference between $\|V_{n+1}(\varrho e_1 - H_{n+1,n}y^n)\|$ and $\|\varrho e_1 - H_{n+1,n}y^n\|$. In the general case, we are not able to say more than is given in the following corollary.

COROLLARY 3.2. *Assuming (3.8), the norm of the true GMRES residual is bounded by*

$$\begin{aligned}\|\,b-Ax^n\,\| &\leq (n+1)^{1/2}(1+(N+4)\varepsilon)\,\|\,\varrho e_1 - H_{n+1,n}y^n\,\| \\ &+ \zeta_{10}\eta(n,l,N)\varepsilon\varrho\frac{\|\,A\,\|}{\sigma_n(H_{n+1,n})}[1-2\zeta_8 n^{3/2}\varepsilon\kappa(H_{n+1,n})]^{-1} \\ &+ \nu(A,b,x^0).\end{aligned} \tag{3.13}$$

PROOF. Using $\| V_{n+1}(\varrho e_1 - H_{n+1,n}y^n) \| \leq \| V_{n+1} \| \| \varrho e_1 - H_{n+1,n}y^n \|$ and $\|V_{n+1}\| \leq (n+1)^{1/2}(1+(N+4)\varepsilon)$, the statement follows from (3.12). □

For a vanishing Arnoldi residual, (3.13) gives an upper bound for the final accuracy of GMRES. If the matrix $H_{n+1,n}$ is ill-conditioned, however, the bound on the norm of the true residual may be large. We do not have an upper bound on the Arnoldi residual in terms of the true residual, but experiments suggest that the Arnoldi residual is never significantly larger than the true residual.

We also wish to express the size of

$$\sigma_n(H_{n+1,n})[1 - 2\zeta_8 n^{3/2}\varepsilon\kappa(H_{n+1,n})]$$

in terms of the singular values of the original matrix A. For that we need an additional restriction on the computed Arnoldi basis V_n. More specifically, we will assume that, for any n, V_n is not too far from an orthonormal matrix. In Lemma 3.3, a general form of this assumption is used; the statements for the specific GMRES implementations will follow.

LEMMA 3.3. *Consider the Arnoldi recurrence (2.2). Suppose there exists an N by $(n+1)$ matrix $\bar{V}_{n+1}$ with orthonormal columns such that*

$$\|V_{n+1} - \bar{V}_{n+1}\| \leq \delta < 1. \tag{3.14}$$

Then

$$\begin{aligned}(1-\delta)\|\varrho e_1 - H_{n+1,n}y^n\| &\leq \|V_{n+1}(\varrho e_1 - H_{n+1,n}y^n)\| \\ &\leq (1+\delta)\|\varrho e_1 - H_{n+1,n}y^n\|,\end{aligned} \tag{3.15}$$

$$\sigma_1(H_{n+1,n}) \leq \delta_1\sigma_1(A), \qquad \delta_1 = \frac{1+\delta+\zeta_1\eta(n,l,N)\varepsilon}{1-\delta}, \tag{3.16}$$

$$\sigma_n(H_{n+1,n}) \geq \sigma_N(A) - \delta_2\sigma_1(A)\ , \quad \delta_2 = \delta(1+\delta_1) + \zeta_1\eta(n,l,N)\varepsilon. \tag{3.17}$$

PROOF. For any $(n+1)$ by n matrix X

$$\sigma_{n+1}(V_{n+1})\|X\| \leq \|V_{n+1}X\| \leq \sigma_1(V_{n+1})\|X\|.$$

Using the perturbation result for singular values ([12], p. 428), (3.15) follows immediately from (3.14). From (2.2), (3.14)

$$A\bar{V}_n + A(V_n - \bar{V}_n) + (\bar{V}_{n+1} - V_{n+1})H_{n+1,n} - F_n = \bar{V}_{n+1}H_{n+1,n}.$$

Taking norms of both sides we obtain, after a simple manipulation, (3.16). Moreover,

$$A\bar{V}_n - \bar{F}_n = \bar{V}_{n+1}H_{n+1,n},$$

where

$$\begin{aligned}\|\bar{F}_n\| &\leq (\delta + \delta\delta_1)\|A\| + \zeta_1\eta(n,l,N)\varepsilon\|A\| \\ &= (\delta(1+\delta_1) + \zeta_1\eta(n,l,N)\varepsilon)\|A\| = \delta_2\sigma_1(A).\end{aligned}$$

Using again the perturbation result for singular values,

$$\sigma_n(H_{n+1,n}) = \sigma_n(\bar{V}_{n+1}H_{n+1,n}) \geq \sigma_n(A\bar{V}_n) - \|\bar{F}_n\| \geq \sigma_n(A\bar{V}_n) - \delta_2\sigma_1(A).$$

Considering

$$\sigma_n(A\bar{V}_n) = \min_{\|u\|=1} \|A\bar{V}_n u\| \geq \min_{\|v\|=1} \|Av\| = \sigma_N(A)$$

we have

$$\sigma_n(A\bar{V}_n) \geq \sigma_N(A),$$

which finishes the proof. □

For the HHA implementation, (2.8) suggests the substitution $\bar{V}_{n+1} \equiv \hat{V}_{n+1}$, $\delta = \zeta_5 n^{3/2} N\varepsilon$. We will assume $\zeta_5 n^{3/2} N\varepsilon \ll 1$. Then $\delta_1 \leq 1 + (2\zeta_5 n^{3/2}N + \zeta_1\eta(n,l,N))\varepsilon + O(\varepsilon^2)$ and

$$\begin{aligned}\sigma_n(H_{n+1,n}) &- 2\zeta_8 n^{3/2}\varepsilon\sigma_1(H_{n+1,n}) \\ &\geq \sigma_N(A) - (\zeta_5 n^{3/2}N + \zeta_1\eta(n,l,N) + 2\zeta_8 n^{3/2})\varepsilon\sigma_1(A) + O(\varepsilon^2).\end{aligned}$$

We denote $\zeta_5 n^{3/2}N + \zeta_1\eta(n,l,N) + 2\zeta_8 n^{3/2} = \eta'(n,l,N) \leq \zeta_{11} nN^{3/2}$. Consequently,

$$\sigma_n(H_{n+1,n}) - 2\zeta_8 n^{3/2}\varepsilon\sigma_1(H_{n+1,n}) \geq \sigma_N(A) - \zeta_{12}\eta'(n,l,N)\varepsilon\sigma_1(A).$$

Summarizing, we have the following theorem.

THEOREM 3.4. *Consider the Householder implementation of GMRES. Assuming that*

$$\zeta_5 n^{3/2}N\varepsilon \ll 1, \qquad 1 - \zeta_{12}\eta'(n,l,N)\varepsilon\kappa(A) > 0,$$

the norms of the true and Arnoldi residuals satisfy

$$\begin{aligned}(1 - \zeta_5 n^{3/2}N\varepsilon) \,\|\, \varrho e_1 - H_{n+1,n}y^n \,\|\, - \varrho\omega(A) - \nu(A,b,x^0) \\ \leq \|\, b - Ax^n \,\| \\ \leq (1 + \zeta_5 n^{3/2}N\varepsilon) \,\|\, \varrho e_1 - H_{n+1,n}y^n \,\|\, + \varrho\omega(A) + \nu(A,b,x^0),\end{aligned} \tag{3.18}$$

where

$$\omega(A) = \frac{\zeta_{10}\eta(n,l,N)\varepsilon\kappa(A)}{1 - \zeta_{12}\eta'(n,l,N)\varepsilon\kappa(A)} \tag{3.19}$$

and $\nu(A,b,x^0)$ is defined by (3.4).

PROOF. Combining Lemma 3.1 and Lemma 3.3, the statement follows after some manipulations. □

If the Arnoldi residual norm is less than $\rho\omega(A)$, then (3.18) suggests that the Arnoldi residual norm may no longer provide a reasonable estimate of the true residual norm. Recall, however, that the term $\rho\omega(A)$ was derived from our bound on $\|y^n\|$. We will now show that when the Arnoldi residual norm, divided by $\|H_{n+1,n}\|\|y^n\| + \rho$, becomes much smaller than $\kappa(A)^{-1}$ and $\kappa(A)\eta(n,l,N)\varepsilon \ll 1$, a better bound on $\|y^n\|$ can be derived by using its relation to x^n in (3.1).

LEMMA 3.5. *Consider the Householder implementation of GMRES, and assume*

$$1 - \zeta_5 n^{3/2} N\varepsilon - O(n^{3/2})\varepsilon \gg 0.$$

Suppose

$$\|\rho e_1 - H_{n+1,n} y^n\| / (\|H_{n+1,n}\| \|y^n\| + \rho) \leq \psi. \tag{3.20}$$

Then

$$\frac{\|b - Ax^n\|}{\|A\|(\|x\| + \|x^n\|)} \leq \psi(2 + 2\mu) + \varepsilon \zeta_7' \eta(n, l, N)(1 + \mu) + O(\varepsilon^2 + \psi\varepsilon), \tag{3.21}$$

where $\mu = \|x^0\|/\|x\|$, and ζ_7' is a constant to be defined below.

PROOF. From (3.4), we have

$$\begin{aligned} &\|(b - Ax^n) - V_{n+1}(\rho e_1 - H_{n+1,n} y^n)\| \\ &\qquad \leq \zeta_7 \eta(n, l, N)\varepsilon \|A\| \|y^n\| + O(lN^{1/2} + N)\varepsilon \|A\| \|x^0\| + O(N)\varepsilon \|b\|, \end{aligned}$$

and (3.15) and (2.8) imply

$$\|V_{n+1}(\rho e_1 - H_{n+1,n} y^n)\| \leq (1 + \delta)\|\rho e_1 - H_{n+1,n} y^n\|, \quad \delta = \zeta_5 n^{3/2} N\varepsilon.$$

Combining these results we have

$$\begin{aligned} \|b - Ax^n\| \;&\leq\; (1 + \delta)\|\rho e_1 - H_{n+1,n} y^n\| + \zeta_7 \eta(n, l, N)\varepsilon \|A\| \|y^n\| \\ &+\; O(lN^{1/2} + N)\varepsilon \|A\| \|x^0\| + O(N)\varepsilon \|b\|, \end{aligned}$$

and, with the assumption (3.20), this becomes

$$\begin{aligned} \frac{\|b - Ax^n\|}{\|A\|(\|x\| + \|x^n\|)} \;&\leq\; (1 + \delta)\psi \frac{\|H_{n+1,n}\| \|y^n\| + \rho}{\|A\|(\|x\| + \|x^n\|)} \\ &+\; \zeta_7 \eta(n, l, N)\varepsilon (\|y^n\| / (\|x\| + \|x^n\|)) \\ &+\; O(lN^{1/2} + N)\varepsilon\mu + O(N)\varepsilon. \end{aligned} \tag{3.22}$$

We therefore need to bound $\|y^n\|/(\|x\| + \|x^n\|)$ and

$$(\|H_{n+1,n}\| \|y^n\| + \rho)/(\|A\|(\|x\| + \|x^n\|)).$$

From (3.1) we have

$$\|V_n y^n\| \leq \|x^n\| + \|x^0\| + O(n^{3/2})\varepsilon \|y^n\| + \varepsilon \|x^0\|,$$

and using (2.8) with this gives

$$\begin{aligned} \|y^n\| \;&=\; \|\hat{V}_n y^n\| \leq \|V_n y^n\| + \|\hat{V}_n - V_n\| \|y^n\| \\ &\leq\; \|x^n\| + (1 + \varepsilon)\|x^0\| + [O(n^{3/2})\varepsilon + \zeta_5 n^{3/2} N\varepsilon] \|y^n\|. \end{aligned}$$

Assuming $1 - \zeta_5 n^{3/2} N\varepsilon - O(n^{3/2})\varepsilon \gg 0$, this implies

$$\|y^n\| \leq \frac{1}{1 - \zeta_5 n^{3/2} N\varepsilon - O(n^{3/2})\varepsilon} \left(\|x^n\| + (1 + \varepsilon)\|x^0\| \right),$$

and we will write this in the form

$$\|y^n\| \le (1+\zeta_5' n^{3/2} N\varepsilon)\left(\|x^n\| + (1+\varepsilon)\|x^0\|\right). \tag{3.23}$$

It follows that

$$\|y^n\|/(\|x\| + \|x^n\|) \le (1+\zeta_5' n^{3/2} N\varepsilon)\,(1+(1+\varepsilon)\mu). \tag{3.24}$$

We also know that

$$\rho \le \|b - Ax^0\| + O(lN^{1/2}+N)\varepsilon\|A\|\|x^0\| + O(N)\varepsilon\|b\|, \tag{3.25}$$

and from (3.16),

$$\|H_{n+1,n}\| \le \delta_1\|A\|, \quad \delta_1 = \frac{1+\delta+\zeta_1\eta(n,l,N)\varepsilon}{1-\delta}. \tag{3.26}$$

Combining (3.23), (3.25), and (3.26) gives

$$\begin{aligned} \frac{\|H_{n+1,n}\|\|y^n\| + \rho}{\|A\|(\|x\| + \|x^n\|)} &\le \delta_1(1+\zeta_5' n^{3/2}N\varepsilon)(1+(1+\varepsilon)\mu) \\ &+ 1+\mu+O(lN^{1/2}+N)\varepsilon\mu + O(N)\varepsilon. \end{aligned} \tag{3.27}$$

Substituting (3.24) and (3.27) into (3.22) we find

$$\begin{aligned} \frac{\|b-Ax^n\|}{\|A\|(\|x\|+\|x^n\|)} &\le (1+\delta)\psi[\delta_1(1+\zeta_5' n^{3/2}N\varepsilon)(1+(1+\varepsilon)\mu) \\ &+ 1+\mu+O(lN^{1/2}+N)\varepsilon\mu + O(N)\varepsilon] \\ &+ \zeta_7\eta(n,l,N)\varepsilon(1+\zeta_5' n^{3/2}N\varepsilon)(1+(1+\varepsilon)\mu) \\ &+ O(lN^{1/2}+N)\varepsilon\mu + O(N)\varepsilon. \end{aligned}$$

Omitting terms of order ε^2 or $\psi\varepsilon$, this becomes

$$\begin{aligned} \frac{\|b-Ax^n\|}{\|A\|(\|x\|+\|x^n\|)} &\le \psi(2+2\mu) \\ &+ \varepsilon[\zeta_7\eta(n,l,N)(1+\mu) + O(lN^{1/2}+N)\mu + O(N)] \\ &+ O(\varepsilon^2+\psi\varepsilon). \end{aligned}$$

Since the largest term inside the brackets is of order $\eta(n,l,N)(1+\mu)$, if we define ζ_7' so that

$$\zeta_7'\eta(n,l,N) \ge \zeta_7\eta(n,l,N) + O(lN^{1/2}+N) + O(N),$$

then (3.21) is proved. □

It is reasonable to assume that $\mu = \|x^0\|/\|x\|$ is of moderate size. If it is extremely large, then the user has chosen an inappropriate initial guess and none of the GMRES implementations discussed will guarantee an accurate solution. If the initial guess is the zero vector, then $\mu = 0$.

Let τ represent the expression on the right-hand side of (3.21). Then (3.21) implies

$$\frac{\|x - x^n\|}{\|x\| + \|x^n\|} \leq \kappa(A)\tau.$$

If $\kappa(A)\tau < 1$, then this implies

$$\max\{\|x\|, \|x^n\|\} \leq \left(\frac{1 + \kappa(A)\tau}{1 - \kappa(A)\tau}\right) \min\{\|x\|, \|x^n\|\}. \tag{3.28}$$

Suppose $(1 + \kappa(A)\tau)/(1 - \kappa(A)\tau) = \theta$ is of moderate size. This will be the case if $\psi(2 + 2\mu) \ll \kappa(A)^{-1}$ and if $\varepsilon\zeta_7'\eta(n, l, N)$ is also much less than $\kappa(A)^{-1}$; i.e., $\kappa(A) \ll \varepsilon^{-1}/(\zeta_7'\eta(n, l, N))$. We can then use (3.28) in (3.23) to obtain a bound on $\|y^n\|$ in terms of $\|x\|$ and $\|x^0\|$,

$$\|y^n\| \leq (1 + \zeta_5' n^{3/2} N\varepsilon)(\theta\|x\| + (1 + \varepsilon)\|x^0\|).$$

With this estimate for $\|y^n\|$, (3.4) becomes

$$\begin{aligned} &\|(b - Ax^n) - V_{n+1}(\rho e_1 - H_{n+1,n}y^n)\| \\ &\quad \leq \zeta_7\eta(n, l, N)\varepsilon\|A\|(\theta\|x\| + \|x^0\|) + \nu(A, b, x^0) + O(\varepsilon^2), \end{aligned} \tag{3.29}$$

and we have the following theorem.

THEOREM 3.6. *Consider the Householder implementation of GMRES. Assume that*

$$\zeta_5 n^{3/2} N\varepsilon \ll 1, \quad 1 - \zeta_{12}\eta'(n, l, N)\varepsilon\kappa(A) > 0,$$

and

$$1 - \kappa(A)\tau > 0,$$

where τ is the expression on the right-hand side of (3.21). *Define $\theta \equiv (1 + \kappa(A)\tau)/(1 - \kappa(A)\tau)$. Then the norms of the true and Arnoldi residuals satisfy*

$$\begin{aligned} &(1 - \zeta_5 n^{3/2} N\varepsilon)\|\rho e_1 - H_{n+1,n}y^n\| - \zeta_7\eta(n, l, N)\varepsilon\|A\|(\theta\|x\| + \|x^0\|) - \nu(A, b, x^0) \\ &\qquad \leq \|b - Ax^n\| \leq \\ &(1 + \zeta_5 n^{3/2} N\varepsilon)\|\rho e_1 - H_{n+1,n}y^n\| + \zeta_7\eta(n, l, N)\varepsilon\|A\|(\theta\|x\| + \|x^0\|) + \nu(A, b, x^0). \end{aligned} \tag{3.30}$$

PROOF. Combining Lemma 3.3 with inequality (3.29) gives the desired result. □

Theorem 3.6 gives a more realistic estimate of the ultimately attainable accuracy of the Householder implementation of GMRES. If the Arnoldi residual becomes tiny, the true residual reaches a level of order $\varepsilon\|A\|\|x\|$ instead of $\kappa(A)\varepsilon\rho$, as suggested in Theorem 3.4. The upper bound in Theorem 3.4 appears to be realistic, however, when the Arnoldi residual is not small enough to satisfy the assumptions of Theorem 3.6. In those cases, the difference between the true and Arnoldi residual norms may indeed depend on $\kappa(A)$. The constants in both theorems are large overestimates, especially for the "average case". For some related

results see [13] and also [5] and [31]. Similar dependence of accuracy of iterative algorithms for computing eigenvalues and solving linear systems on the properties of the problem was described experimentally by Chatelin, Godet-Thobie, Frayssé and Braconnier in [9], [10], [6], [7].

REMARK: In the previous text we compared the norms of the true and Arnoldi residuals. In practice, however, the norm of the Arnoldi residual is computed as an absolute value ξ_{n+1} of the bottom element of the right hand side vector ϱe_1 transformed by application of the Givens rotations reducing the upper Hessenberg matrix $H_{n+1,n}$ into the upper triangular form (cf. (3.5), (3.6))

$$\xi_{n+1} = |e_{n+1}^T fl(Z_{n+1}(\varrho e_1))|.$$

We should therefore compare $\|b - Ax^n\|$ with the computed value ξ_{n+1} rather than $\|\varrho e_1 - H_{n+1,n}y^n\|$. As the difference $|\xi_{n+1} - \|\varrho e_1 - H_{n+1,n}y^n\||$ is less than the bound for $|\|b - Ax^n\| - \|\varrho e_1 - H_{n+1,n}y^n\||$, this would make no difference in our bounds except changing the multiplicative constant. Indeed

$$\begin{aligned}
\|\varrho e_1 - H_{n+1,n}y^n\| &= \|\hat{Z}_{n+1}(\varrho e_1 - H_{n+1,n}y^n)\| \\
&\leq \left\| \begin{pmatrix} S_n \\ 0 \end{pmatrix} y^n + G_n y^n - fl(Z_{n+1}(\varrho e_1)) \right\| + O(n)\varrho\varepsilon \\
&\leq \|S_n y^n - (\, I_n \quad 0\,)\, fl(Z_{n+1}(\varrho e_1))\| \\
&+ \quad \xi_{n+1} + \|G_n\|\|y^n\| + O(n)\varrho\varepsilon.
\end{aligned}$$

The first term is the residual of the back substitution, which is bounded by

$$\|S_n y^n - (\, I_n \quad 0\,)\, fl(Z_{n+1}(\varrho e_1))\| \leq \|\tilde{G}_n\|\|y^n\| \leq O(n)\varepsilon\|S_n\|\|y^n\|.$$

Because $\|G_n\| \leq O(n^{3/2})\varepsilon\|H_{n+1,n}\|$, we have

$$\begin{aligned}
&|\,\|\varrho e_1 - H_{n+1,n}y^n\| - \xi_{n+1}| \\
&\qquad \leq O(n)\varepsilon\|S_n\|\|y^n\| + O(n^{3/2})\varepsilon\|H_{n+1,n}\|\|y^n\| + O(n)\varrho\varepsilon \\
&\qquad \leq O(n^{3/2})\varepsilon\|A\|\|y^n\| + O(n)\varrho\varepsilon,
\end{aligned}$$

and this is smaller than the difference in (3.4).

It should be noted that in practice, the number ξ_{n+1} may become *much* smaller than either the true or the Arnoldi residual norm, but only after the true and Arnoldi residual norms have become about as small as one could reasonably expect. We will show in Section 4, for instance, that for the Householder implementation, the true residual at step N is of order $\varepsilon\|A\|\|x\|$, which is the same order as the residual from a direct solution using QR decomposition, but the number ξ_{n+1} may still be orders of magnitude smaller than this.

Up to now, we considered in this section the HHA implementation of GMRES. Can similar statements be established for the GMRES implementations based on the Gram-Schmidt algorithm? The loss of orthogonality among the basis vectors computed via the MGSA implementation seems to be a serious

difficulty. Though (2.6) and (2.7) look formally the same for both the HHA and MGSA implementations, for MGSA the loss of orthogonality (described by (2.9)) depends on the condition number of the matrix (v_1, AV_n), which may become large. Finite precision analysis of this situation needs further work; we will return to this point elsewhere.

For the IMGS, the Hoffmann conjecture in [15] states that for the QR decomposition of any full column rank matrix, the resulting matrix of the basis vectors satisfies

$$\|I - Q^T Q\| \leq \zeta \gamma N^{1/2} \varepsilon \tag{3.31}$$

for any value of the parameter γ (see the description of IMGSA, ICGSA in the Introduction), where ζ is a constant independent of γ, N and ε. He observed that for small values of γ (say, $\gamma \leq 2$), the orthogonality of the resulting matrix is of the same order even for ICGS. Moreover, for any computed basis vector, one reorthogonalization step was always sufficient. A detailed discussion of the Hoffmann conjecture and its analogue for the IMGSA and ICGSA is intricate, and its theoretical justification has not been completed yet. The following assumption on IMGSA and ICGSA seems realistic for small values of γ.

ASSUMPTION 3.7. *For γ sufficiently small, the Arnoldi basis V_N computed via the IMGSA or ICGSA satisfies*

$$\|I - V_N^T V_N\|_F \leq \zeta_{13} \gamma N^{\alpha} \varepsilon, \tag{3.32}$$

while the total number of reorthogonalization steps does not exceed ξN. Here, ζ_{13} is a constant independent of γ, N and ε; ξ and α stand for small constants close to one.

As was pointed out by the referee, (3.32) implies that there exists an exactly orthonormal matrix $\bar{V}_N$ such that,

$$\|V_N - \bar{V}_N\|_F \leq \zeta_{13} \gamma N^{\alpha} \varepsilon. \tag{3.33}$$

Indeed, suppose we express (3.32) as $\|I - V^T V\|_F \leq \phi$, where $V = V_N$ and $\phi = \zeta_{13} \gamma N^{\alpha} \varepsilon$. We have the SVD of the matrix V, $V = U \Sigma W^T$, where $U^T U = W^T W = I$, and $\Sigma = diag(\sigma_1, \ldots, \sigma_N)$. Then $\|I - V^T V\|_F \leq \phi$ implies

$$\phi^2 \geq \|I - V^T V\|_F^2 = \|W(I - \Sigma^2) W^T\|_F^2 = \sum_i (1 - \sigma_i^2)^2.$$

Noting that $(1 - \sigma_i)^2 \leq (1 - \sigma_i^2)^2$, we then have

$$\|V - UW^T\|_F^2 = \|U(\Sigma - I) W^T\|_F^2 = \sum_i (1 - \sigma_i)^2 \leq \sum_i (1 - \sigma_i^2)^2 \leq \phi^2.$$

Thus $\bar{V}_N \equiv UW^T$ is an exactly orthonormal matrix satisfying $\|V_N - \bar{V}_N\|_F \leq \phi$. Consequently, the analogues of Theorems 3.4 and 3.6 for the ICGSA and IMGSA implementations can be easily obtained.

4 Backward stability.

Given an approximation $\tilde{x}$ to the solution $Ax = b$, we consider the normwise backward error

$$\zeta(\tilde{x}) = \min\{\nu : (A + \Delta A)\tilde{x} = b + \Delta b, \|\Delta A\|/\|A\| \leq \nu, \|\Delta b\|/\|b\| \leq \nu\}. \tag{4.1}$$

As shown by Rigal and Gaches [22], the backward error (4.1) can be expressed using the residual $b - A\tilde{x}$,

$$\zeta(\tilde{x}) = \|b - A\tilde{x}\|/(\|A\|\|\tilde{x}\| + \|b\|). \tag{4.2}$$

There are several concepts of stability of methods for solving $Ax = b$; for thorough discussions we refer to [14] and [8]. In this paper we call a method normwise backward stable, if it gives an approximation $\tilde{x}$ to the solution such that the corresponding backward error $\zeta(\tilde{x})$ is proportional to the machine precision ε.

First, we consider direct solution of $Ax = b$ using the QR decomposition computed via the Householder or Givens transformations. From [12] and the results by Lawson and Hanson [17], the computed solution $\tilde{x}$ represents the exact solution of a least squares problem

$$\|(A + \Delta A)\tilde{x} - (b + \Delta b)\| = \min_z \|(A + \Delta A)z - (b + \Delta b)\|,$$

where

$$\|\Delta A\|/\|A\| \leq \zeta_{14} N^{5/2}\varepsilon, \|\Delta b\|/\|b\| \leq \zeta_{14} N^2 \varepsilon.$$

Assuming that

$$\kappa(A)\|\Delta A\|/\|A\| \leq \zeta_{14} N^{5/2}\kappa(A)\varepsilon < 1, \tag{4.3}$$

we can apply results developed by Wedin [29], and conclude that

$$\|x - \tilde{x}\| \leq \frac{\zeta_{14} N^{5/2}\kappa(A)\varepsilon}{1 - \zeta_{14} N^{5/2}\kappa(A)\varepsilon}(\|x\| + \|b\|/\|A\|), \tag{4.4}$$

$$\|b - A\tilde{x}\| \leq \zeta_{14} N^{5/2}\varepsilon(\|A\|\|x\| + \|b\|). \tag{4.5}$$

Moreover, under the assumption (4.3) the matrix $A + \Delta A$ is nonsingular. Consequently,

$$(A + \Delta A)\tilde{x} = b + \Delta b,$$

and using (4.1), the backward error $\zeta(\tilde{x})$ is bounded by

$$\zeta(\tilde{x}) \leq \zeta_{14} N^{5/2}\varepsilon. \tag{4.6}$$

In this section, we will show analogous results for the final approximation x^N computed by the HHA implementation of the GMRES method.

THEOREM 4.1. *Consider the Householder implementation of GMRES. Assume that*

$$1 - O(N^{5/2})\varepsilon\kappa(A) \gg 0,$$

and set $\mu = \|x^0\|/\|x\|$. *Then*

$$\|b - Ax^N\|/(\|A\|(\|x\| + \|x^N\|)) \leq (\zeta_{14} + \zeta_7')N^{5/2}\varepsilon(2 + 2\mu) + O(\varepsilon^2). \tag{4.7}$$

PROOF. At the final step of the QR decomposition (1.8)–(1.11), the N-th column $H_{.,N}$ of the matrix $H_{N,N}$ is generated by applying the computed elementary Householder matrices $P_1, \ldots, P_N$ to the vector $fl(Av_N)$. Using the rounding error analysis of the Householder orthogonalizations (see, e.g., [30], [3]) we can write

$$H_{.,N} = fl(P_N \ldots P_1 A v_N) = \hat{V}_N^T A v_N + \tilde{F}_{.,N},$$

where

$$\|\tilde{F}_{.,N}\| \leq O(N^2)\varepsilon\|A\| + O(lN^{1/2})\varepsilon\|A\| + O(\varepsilon^2).$$

Combining with (2.6), (2.7)

$$AV_N = \hat{V}_N H_{N,N} + \tilde{F}_N, \tag{4.8}$$

$$\|\tilde{F}_N\| \leq \zeta_{15}N^{5/2}\varepsilon\|A\|. \tag{4.9}$$

Note that this bound does not depend on the sparsity of the matrix A because for $n = N$ the bound for the error in the orthogonalization process is larger than the estimate of the part of the error which is affected by the matrix-vector multiplication.

Similarly to the proof of Lemma 3.3, the relations (2.8), (4.8) and (4.9) imply

$$\begin{aligned} \sigma_1(H_{N,N}) &\leq [1 + (\zeta_{15} + \zeta_5)N^{5/2}\varepsilon]\sigma_1(A), \\ \sigma_N(H_{N,N}) &\geq \sigma_N(A) - (\zeta_{15} + \zeta_5)N^{5/2}\varepsilon\sigma_1(A), \end{aligned}$$

which, assuming that $1 - (\zeta_{15} + \zeta_5)N^{5/2}\varepsilon\kappa(A) > 0$, gives

$$\kappa(H_{N,N}) \leq \frac{1 + \zeta_{16}N^{5/2}\varepsilon}{1 - \zeta_{16}N^{5/2}\varepsilon\kappa(A)}\kappa(A), \quad \zeta_{16} = \zeta_{15} + \zeta_5. \tag{4.10}$$

Following the results by Lawson and Hanson [17], the computed solution y^N of the transformed least squares problem

$$\min_y \|\varrho e_1 - H_{N,N}y\|$$

represents the exact solution of a perturbed problem

$$\|(\varrho e_1 + \Delta e) - (H_{N,N} + \Delta H)y^N\| = \min_z \|(\varrho e_1 + \Delta e) - (H_{N,N} + \Delta H)z\|,$$

where the size of perturbation is bounded by

$$\|\Delta H\|/\|H_{N,N}\| \leq \zeta_{14}N^{5/2}\varepsilon, \quad \|\Delta e\|/\varrho \leq \zeta_{14}N^2\varepsilon.$$

Assuming that

$$\kappa(H_{N,N})\|\Delta H\|/\|H_{N,N}\| \leq \zeta_{14}N^{5/2}\varepsilon\kappa(H_{N,N}) < 1,$$

which is certainly satisfied if

$$\frac{\zeta_{14}N^{5/2}\varepsilon\kappa(A)}{1-\zeta_{16}N^{5/2}\varepsilon\kappa(A)}(1+\zeta_{16}N^{5/2}\varepsilon) < 1, \tag{4.11}$$

the matrix $H_{N,N}+\Delta H$ is nonsingular and y^N solves the linear system

$$(H_{N,N}+\Delta H)y^N = \varrho e_1 + \Delta e.$$

The backward error in solving this system must then be bounded by

$$\frac{\|\varrho e_1 - H_{N,N}y^N\|}{\|H_{N,N}\|\|y^N\|+\varrho} \leq \zeta_{14}N^{5/2}\varepsilon. \tag{4.12}$$

Using this value for ψ in Lemma 3.5 gives the desired result (4.7). □

COROLLARY 4.2. *Let the assumptions of Theorem 4.1 hold and assume also that*

$$1-\kappa(A)\tau > 0,$$

where τ represents the expression on the right-hand side of (4.7). Define

$$\theta = (1+\kappa(A)\tau)/(1-\kappa(A)\tau).$$

Then the backward error of the final GMRES approximation is bounded by

$$\frac{\|b-Ax^N\|}{\|A\|\|x^N\|+\|b\|} \leq (1+\theta)(\zeta_{14}+\zeta_7')N^{5/2}\varepsilon(2+2\mu)+O(\varepsilon^2). \tag{4.13}$$

Also

$$\max\left\{\frac{\|b-Ax^N\|}{\|A\|\|x\|}, \frac{\|b-Ax^N\|}{\|A\|\|x^N\|}\right\} \leq (1+\theta)(\zeta_{14}+\zeta_7')N^{5/2}\varepsilon(2+2\mu) + O(\varepsilon^2). \tag{4.14}$$

PROOF. Arguing as in Section 3, it follows from (4.7) and the assumption on τ that

$$\max\{\|x\|,\|x^N\|\} \leq \theta\min\{\|x\|,\|x^N\|\},$$

and from this it follows that the quantity on the left-hand side of (4.14), as well as the smaller quantity on the left-hand side of (4.13), is bounded by $(1+\theta)\|b-Ax^N\|/(\|A\|(\|x\|+\|x^N\|))$. The desired results now follow from Theorem 4.1. □

Again, using the Assumption 3.5 and relation (3.22), the analogous statement can be proved for the ICGSA and IMGSA implementations.

5 Concluding remarks.

We have concentrated on the formulation of GMRES based on constructing the Arnoldi basis of the Krylov subspaces and then solving the transformed least squares problem. We have shown that if the Arnoldi basis is computed via Householder orthogonalization and the transformed least squares problem is solved using Givens rotations, then the computed GMRES approximation x^N has a guaranteed backward error of size at worst $O(N^{5/2})\varepsilon$. This means that the backward error and the final residual norm guaranteed by the HHA implementation of GMRES are essentially the same as those guaranteed by direct solving of the system $Ax = b$ via the Householder or Givens QR decomposition. Note, however, that these results have been achieved under slightly different assumptions on the numerical nonsingularity of the system matrix. The bound for the direct application of the QR decomposition requires that $\zeta_{14} N^{5/2} \kappa(A)\varepsilon < 1$, see (4.3) and (4.5). In the case of GMRES, the assumptions in Theorem 4.1 and Corollary 4.2 are slightly stronger. They mean essentially that

$$(\zeta_{14} + \zeta_7')N^{5/2}(2 + 2\mu)\kappa(A)\varepsilon \leq c < 1,$$

so that

$$\theta \leq (1 + c)/(1 - c) = O(1), \quad \text{and} \quad \zeta_{16} N^{5/2} \kappa(A)\varepsilon \leq c' < 1.$$

Theorem 3.4 uses similar assumptions as the analysis of direct QR, but does not guarantee the same size of the residual.

We left many questions open; e.g., the Hoffman conjecture and its analogues for IMGSA and ICGSA should be proved. The stability analysis of the modified Gram-Schmidt GMRES implementation is another example. We concentrated on the most stable implementations of GMRES in which the orthogonality of the vectors in the Arnoldi basis is well preserved. For other implementations, one would intuitively expect much worse convergence behavior due to uncontrolled loss of orthogonality. In many experiments, however, we have observed a different behavior—until the *linear independence* of the Arnoldi vectors is lost, the norms of the true residuals in any GMRES implementation match those of the HHA implementation, regardless of more or less significant loss of orthogonality. Assuming that the linear independence of the Arnoldi vectors is well preserved, the loss of orthogonality seems not to be a very important issue. This phenomenon is not fully understood yet. On the other hand, a gradual loss of orthogonality among the Arnoldi vectors usually leads to the loss of linear independence—this is, of course, a good reason for preserving the orthogonality. It is, however, always a question of extra work. For the modified Gram-Schmidt algorithm, the orthogonality among the Arnoldi vectors depends considerably on the ordering in which the orthogonalization against the previously computed vectors is performed. Recently, Nour-Omid et al. [18] proposed ordering by the

size of the components of the orthogonalized vector in the direction of the previously computed vectors. In our opinion, however, this strategy does not always give satisfactory results. It may be outperformed, e.g., by the proper iterated Gram-Schmidt schemes. The problems mentioned above need further work and we will return to them elsewhere.

Acknowledgements.

The authors are indebted for stimulating discussions on the subject and for useful comments to Valérie Frayssé, Åke Björck and Homer Walker. The comments and suggestions of the anonymous referees have also significantly helped to improve the presentation of the paper.

REFERENCES

1. Å. Björck, *Solving linear least squares problems by Gram-Schmidt orthogonalization*, BIT 7 (1967), pp. 1–21.
2. Å. Björck, *Stability analysis of the method of seminormal equations for linear least squares problems*, Linear Algebra Appl. 88/89 (1987), pp. 31–48.
3. Å. Björck, *Numerics of Gram-Schmidt orthogonalization*, Linear Algebra Appl. 197 (1994), pp. 297–316.
4. Å. Björck and C. C. Paige, *Loss and recapture of orthogonality in the modified Gram-Schmidt algorithm*, SIAM J. Matrix Anal. Appl. 13, 1 (1992), pp. 176–190.
5. J. Bollen, *Round-off error analysis of descent methods for solving linear equations*, Tech. Hogeschool Eindhoven, Netherlands, Ph.D. thesis, 1980.
6. T. Braconnier, F. Chatelin and V. Frayssé, *The influence of large nonnormality on the quality of convergence of iterative methods in linear algebra*, submitted to Linear Algebra Appl.
7. T. Braconnier, *The Arnoldi-Tchebycheff algorithm for solving large nonsymmetric eigenproblems*, Technical Report TR/PA/93/25, CERFACS, Toulouse, France, 1993.
8. J. R. Bunch, *The weak and strong stability of algorithms in numerical linear algebra*, Linear Algebra Appl. 88/89 (1987), pp. 49–66.
9. F. Chatelin and S. Godet-Thobie, *Stability analysis in aeronautical industries*, in Proceedings of the 2nd Symposium on High-Performance Computing, Montpellier, France, M. Durand and F. El Dabaghi, eds., Elsevier, North-Holland, 1991, pp. 415–422.
10. F. Chatelin and V. Frayssé, *Qualitative computing: Elements of a theory for finite precision computation*, COMETT Matari Programme, Orsay, France, 1993.
11. R. W. Freund, G. H. Golub and N. M. Nachtigal, *Iterative solution of linear systems*, Acta Numerica 1 (1992), pp. 1–44.
12. G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, 1989.
13. A. Greenbaum, *Accuracy of computed solutions from conjugate-gradient-like methods*, in Advances in Numerical Methods for Large Sparse Sets of Linear

Equations, Proceedings of PCG '94, M. Natori and T. Nodera, eds., Keio University, Yokohama, Japan, 1994.

14. N. J. Higham and D. Knight, *Componentwise error analysis for stationary iterative methods*, Linear Alg. Appl. 192 (1993), pp. 165-186.
15. W. Hoffmann, *Iterative algorithms for Gram-Schmidt orthogonalization*, Computing 41 (1989), pp. 335–348.
16. R. Karlson, *A study of some roundoff effects of the GMRES method*, Technical Report Li-TH-MAT-R-1990 11, University of Linköping, 1991.
17. C. L. Lawson and R. J. Hanson, *Solving least squares problems*, Prentice-Hall, Englewood Cliffs, New Jersey, 1974.
18. B. Nour-Omid, W. S. Dunbar and A. D. Woodbury, *Ordered modified Gram-Schmidt orthogonalization*, Technical Report, 1992.
19. C. C. Paige, *Computational variants of the Lanczos method for the eigenproblem*, J. Inst. Maths. Applics. 10 (1972), pp. 373–381.
20. C. C. Paige, *Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix*, J. Inst. Maths. Applics. 18 (1976), pp. 341–349.
21. C. C. Paige, *Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem*, Linear Algebra Appl. 34 (1980), pp. 235–258.
22. J. Rigal and J. Gaches, *On the compatibility of a given solution with the data of a linear system*, J. of ACM 14 (1967), pp. 543–526.
23. M. Rozložník and Z. Strakoš, *Variants of the residual minimizing Krylov space methods*, Technical Report 592, Inst. of Computer Science, Academy of Sciences, Prague, 1994.
24. Y. Saad and M. H. Schultz, *GMRES: A Generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput. 7 (1986), pp. 856–869.
25. G. W. Stewart, *Afternotes on numerical analysis*, University of Maryland at College Park, 1993.
26. H. F. Walker and Lu Zhou, *A Simpler GMRES*, Research Rep. 10/92/54, Dept. of Mathematics and Statistics, Utah State University, Logan, 1992.
27. H. F. Walker, *Implementation of the GMRES method using Householder transformations*, SIAM J. Sci. Stat. Comput. 9, 1 (1988), pp. 152–163.
28. H. F. Walker, *Implementations of the GMRES method*, Computer Physics Communications 53, North-Holland, Amsterdam, 1989, pp. 311–320.
29. P. Å. Wedin, *Perturbation theory for pseudo-inverses*, BIT 13 (1973), pp. 217–232.
30. J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.
31. H. Wozniakowski, *Roundoff-error analysis of a new class of conjugate-gradient algorithms*, Linear Algebra Appl. 29 (1980), pp. 507–529.