BIT 37:3 (1997), 706–719.

NUMERICAL BEHAVIOUR OF THE MODIFIED GRAM-SCHMIDT GMRES IMPLEMENTATION *

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

[1] *Courant Institute of Mathematical Sciences, 251 Mercer Street, New York NY 10012, U.S.A. email: greenbau@greenbau.cims.nyu.edu*

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

Abstract.

In [6] the Generalized Minimal Residual Method (GMRES) which constructs the Arnoldi basis and then solves the transformed least squares problem was studied. It was proved that GMRES with the Householder orthogonalization-based implementation of the Arnoldi process (HHA), see [9], is backward stable. In practical computations, however, the Householder orthogonalization is too expensive, and it is usually replaced by the modified Gram-Schmidt process (MGSA). Unlike the HHA case, in the MGSA implementation the orthogonality of the Arnoldi basis vectors is not preserved near the level of machine precision. Despite this, the MGSA-GMRES performs surprisingly well, and its convergence behaviour and the ultimately attainable accuracy do not differ significantly from those of the HHA-GMRES. As it was observed, but not explained, in [6], it is the *linear independence* of the Arnoldi basis, not the orthogonality near machine precision, that is important. Until the linear independence of the basis vectors is nearly lost, the norms of the residuals in the MGSA implementation of GMRES match those of the HHA implementation despite the more significant loss of orthogonality.

In this paper we study the MGSA implementation. It is proved that the Arnoldi basis vectors begin to lose their linear independence *only after* the GMRES residual norm has been reduced to almost its final level of accuracy, which is proportional to $\kappa(A)\varepsilon$, where $\kappa(A)$ is the condition number of A and ε is the machine precision. Consequently, unless the system matrix is very ill-conditioned, the use of the modified Gram-Schmidt GMRES is theoretically well-justified.

AMS subject classification: 65F05.

Key words: linear algebraic systems, numerical stability, GMRES method, modified Gram-Schmidt Arnoldi process.

*Received August 1996. Revised January 1997.

†This work was supported by the GA AS CR under grants 230401 and A 2030706, by the NSF grant Int 921824, and by the DOE grant DEFG0288ER25053. Part of the work was performed while the third author visited Department of Mathematics and Computer Science, Emory University, Atlanta, USA.

1 Introduction.

Consider the linear algebraic system

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

where A is a real nonsingular N by N matrix and b is a real vector. Starting with an initial guess x_0, the GMRES method generates approximate solutions $x_n \in x_0 + K_n(A, r_0)$, $n = 1, 2, \dots$; $K_n(A, r_0) = \text{span}\{r_0, Ar_0, \dots, A^{n-1}r_0\}$, such that

$$\|b - Ax_n\| = \min_{u \in x_0 + K_n(A, r_0)} \|b - Au\|, \tag{1.2}$$

where $r_0 = b - Ax_0$ is the initial residual. One possible way to compute x_n involves constructing an orthonormal basis $\{v_1, \dots, v_n\}$ for the Krylov space $K_n(A, r_0)$ using the Arnoldi process. The Arnoldi recurrence is described in matrix form by

$$AV_n = V_{n+1}H_{n+1,n}, \tag{1.3}$$

where the columns of V_n are the n orthonormal basis vectors, the last column of V_{n+1} is the next orthonormal basis vector for the space $K_{n+1}(A, r_0)$, and $H_{n+1,n}$ is an $n+1$ by n upper Hessenberg matrix. To obtain the approximation x_n satisfying (1.2), one sets $x_n = x_0 + V_n y_n$, where

$$\|\varrho e_1 - H_{n+1,n}y_n\| = \min_y \|\varrho e_1 - H_{n+1,n}y\|, \quad \varrho \equiv \|r_0\|. \tag{1.4}$$

In [6] the Householder orthogonalization was considered for building up the Arnoldi basis. It was proved that the resulting (so called HHA) implementation of GMRES (proposed originally by Walker, see [9]) is backward stable. The proof relied upon the fact that in the HHA process the loss of orthogonality of the computed Arnoldi vectors is proportional to the machine precision (independent of the condition number of the matrix A). This is not true, however, for the cheaper and therefore more frequently used MGSA implementation, in which the Arnoldi basis is computed using the algorithm

```
MGS-Arnoldi algorithm:
$\varrho = \|r_0\|$;
$v_1 = r_0/\varrho$;
for $i = 1, 2, \dots, n$
    $w = Av_i$;
    for $k = 1, 2, \dots, i$
        $h_{k,i} = v_k^T w$;
        $w = w - h_{k,i}v_k$;
    end
    $h_{i+1,i} = \|w\|$;
    $v_{i+1} = w/h_{i+1,i}$;
end.
```

Here, the upper Hessenberg matrix $H_{n+1,n}$ is still computed in a backward stable way, i.e., in finite precision arithmetic we have (see [2, 4, 1, 6])

$$AV_n = V_{n+1}H_{n+1,n} + F_n, \quad AV_n = \hat{V}_{n+1}H_{n+1,n} + \hat{F}_n, \tag{1.5}$$

where

$$\|F_n\| \leq \zeta_1 \eta(n,l,N)\varepsilon\|A\|, \quad \|\hat{F}_n\| \leq \zeta_1 \eta(n,l,N)\varepsilon\|A\|,$$
$$\hat{V}_{n+1}^T\hat{V}_{n+1} = I_{n+1}, \quad \eta(n,l,N) = n^{3/2}N + n^{1/2}lN^{1/2},$$

$\hat{V}_{n+1}$ is the closest orthonormal matrix to V_{n+1} in any unitarily invariant norm, ε is the machine precision, and l is the maximal number of nonzero entries per row of the matrix A (as in [6], by r_0, ϱ, V_n, $H_{n+1,n}$, y_n and x_n we denote from now on the computed quantities). The orthogonality of V_{n+1} may gradually deteriorate, but assuming $nN^{3/2}\varepsilon\kappa([v_1, AV_n]) \ll 1$, it can be shown, see [2, 4, 6], that

$$\|I - V_{n+1}^T V_{n+1}\| \leq \zeta_2 \eta(n,l,N)\varepsilon\kappa([v_1, AV_n]), \tag{1.6}$$

and also

$$\|\hat{V}_{n+1} - V_{n+1}\| \leq \zeta_2 \eta(n,l,N)\varepsilon\kappa([v_1, AV_n]), \tag{1.7}$$

for a moderate size constant ζ_2. Constant factors ζ_1, ζ_2 (as the other constants $\zeta_j, j = 3, \ldots$, introduced in a number of places) are independent of the problem parameters (e.g. $N, \kappa(A)$, etc.) and the machine precision ε, but dependent on the details of the machine arithmetic. In a finite precision MGS-Arnoldi computation, the modified Gram-Schmidt process is applied to the matrix $[v_1, fl(Av_1), \ldots, fl(Av_n)]$, where $fl(\cdot)$ denotes the floating point result of the operation $(\cdot)$. In developing the bounds, however, it is convenient to work with the matrix $[v_1, AV_n]$. We have therefore included effects of rounding errors in the matrix-vector multiplications in the bounds (1.6), (1.7) and all the following results will be related to $[v_1, AV_n]$. Throughout this paper we will assume that $nN^{3/2}\varepsilon \ll 1$, so that (1.6) holds for $n+1 \leq N$ if $\kappa([v_1, AV_n]) \ll N^{-5/2}\varepsilon^{-1}$.

It was observed in [6] that although orthogonality of the MGSA vectors is not maintained near the machine precision, as for the HHA implementation, the norms of the computed residuals of the MGSA-GMRES are almost identical to those of the HHA-GMRES, until the smallest singular value of the matrix V_n begins to depart from the value one. At that point the MGSA-GMRES residual norm begins to stagnate close to its final precision level. This observation is demonstrated on the numerical examples for matrices STEAM1 ($N = 240$, symmetric positive definite matrix used in oil recovery simulations) and IMPCOLE ($N = 225$, unsymmetric matrix from modelling of the hydrocarbon separation problem) taken from the Harwell-Boeing collection (see Figures 1.1–1.4).

In both experiments $x = (1, \ldots, 1)^T$, $b = Ax$ and $x_0 = 0$. Experiments were performed on the SGI Crimson workstation, $\varepsilon = 2.2 \cdot 10^{-16}$, using MATLAB. The solid line represents the norm of the relative true residual $\|r_n\|/\varrho = \|b -$

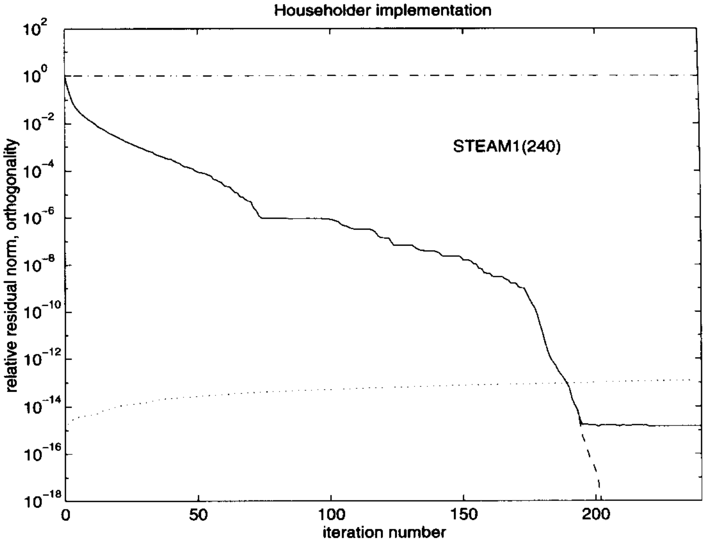

Figure 1.1: Householder implementation for STEAM1.

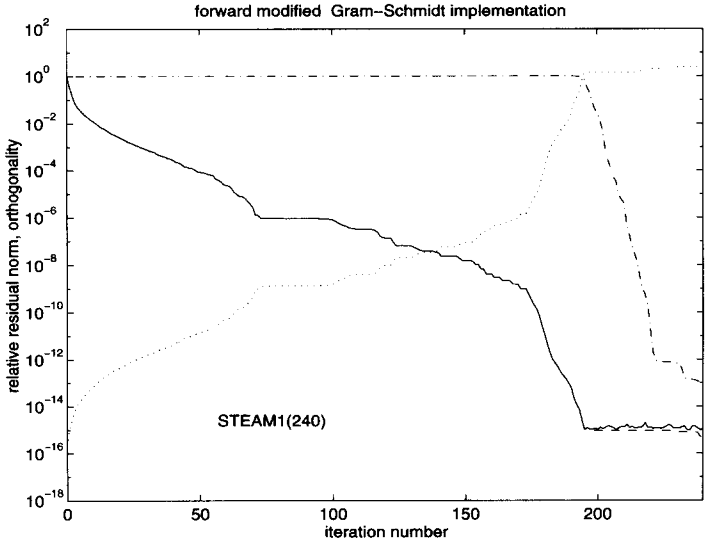

Figure 1.2: Modified Gram-Schmidt implementation for STEAM1.

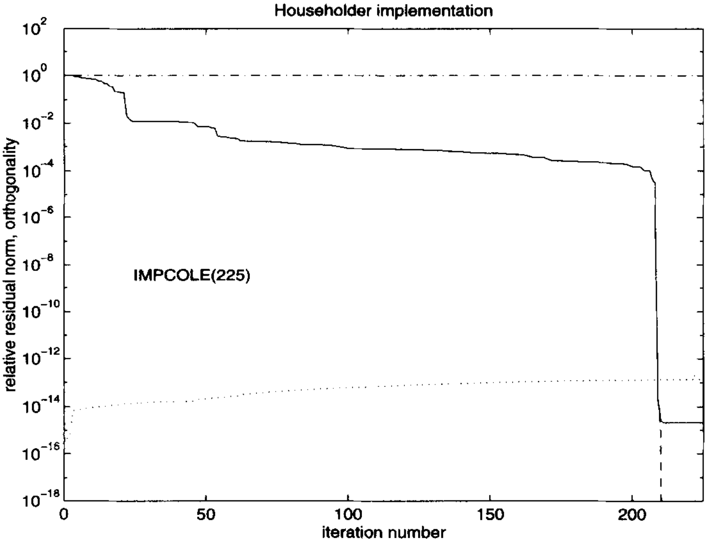

Figure 1.3: Householder implementation for IMPCOLE.

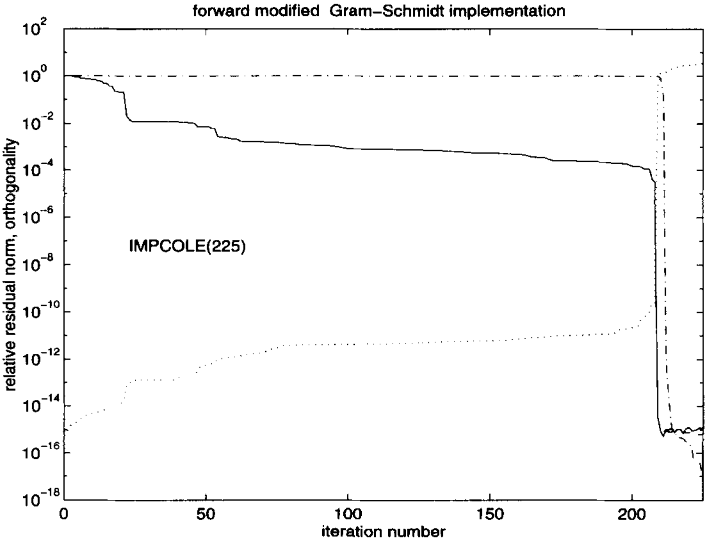

Figure 1.4: Modified Gram-Schmidt implementation for IMPCOLE.

$Ax_n\|/\varrho$, the dashed line the norm of the relative Arnoldi residual $\|t_n\|/\varrho = \|\varrho e_1 - H_{n+1,n}y_n\|/\varrho$, the dash-dotted line the smallest singular value of the matrix V_n and the dotted line the loss of orthogonality $\|I - V_n^T V_n\|_F$ measured in the Frobenius norm. Figures 1.1 and 1.2 describe the results for STEAM1 (the HHA and MGSA implementations, respectively). Similarly, Figures 1.3 and 1.4 correspond to the IMPCOLE. We emphasize that the behaviour illustrated on these figures represents typical behaviour of the MGSA and HHA-GMRES. The condition number of the system matrix was $\kappa(A) = 2.855 \times 10^7$ for STEAM1 and $\kappa(A) = 7.102 \times 10^6$ for IMPCOLE.

In Sections 2 and 3 these observations are justified theoretically. It is proved that the condition number of the matrix $[v_1, AV_n]$ in (1.6) is approximately bounded by the condition number of A divided by the relative norm of the MGSA-GMRES residual at step n. It follows that until this residual norm reaches the level $\kappa(A)\varepsilon$, which is its final theoretical level of accuracy, no serious loss of orthogonality occurs in the modified Gram-Schmidt process, and the behaviour of the MGSA-GMRES is approximately the same as in the HHA-GMRES.

2 Condition of the Arnoldi vectors.

The MGS-Arnoldi algorithm given in Section 1 applies the modified Gram-Schmidt procedure to the independent vectors $[v_1, Av_1, Av_2, \ldots, Av_n]$, where v_1 is given and $v_2, \ldots, v_n$ are computed vectors, each having norm approximately one. In exact arithmetic, the vectors $v_1, \ldots, v_n$ are orthonormal, but in finite precision arithmetic this may not be the case.

The success of the modified Gram-Schmidt procedure in constructing an orthonormal basis depends on the condition number of the matrix $[v_1, AV_n]$, as shown by inequality (1.6). It is easy to see, in addition, that the orthogonality of computed modified Gram-Schmidt vectors is independent of the norms of the original vectors. Thus, the expression $\kappa([v_1, AV_n])$ in (1.6) can be replaced by the condition number of the best conditioned matrix of the form $[v_1, AV_n]D$, where D is any diagonal matrix. In the rest of the paper we will assume that A has been multiplied by a constant so that $\|A\| = 1$ and, for convenience, we will assume also that the norms of the columns of V_n are exactly 1; that is, $D = \mathrm{diag}(1/\|v_1\|, 1/(\|A\|\|v_1\|), \ldots, 1/(\|A\|\|v_n\|))$. In this section we will show that the condition number $\kappa([v_1, AV_n])$ is less than or equal to a moderate size multiple of the condition number of A divided by the norm of the optimal approximation to v_1 by a linear combination of the columns of AV_n, $\min_y \|v_1 - AV_n y\| \equiv \|\hat{r}_n\|/\varrho$. In exact arithmetic, this would be the norm of the GMRES residual at step n divided by the norm of the initial residual. In finite precision arithmetic these two quantities may differ, but it will be shown in Section 3 that, until the relative GMRES residual drops close to the level $\kappa(A)\varepsilon$, they are approximately the same.

We begin with the following general result relating the smallest singular value of an m by n matrix of rank n, $m > n$, to the smallest singular value of the matrix with an extra column appended.

THEOREM 2.1. *Let G be an m by n matrix of rank n, $m > n$, and let h be an m-vector. Let $\sigma_n(G)$ denote the smallest singular value of G, and let $\sigma_{n+1}([G,h])$ denote the smallest singular value of the m by $n+1$ matrix consisting of G appended with the last column h. Then*

$$\sigma_n(G) \geq \sigma_{n+1}([G,h]) \geq \frac{\tau}{\sqrt{\sigma_n^2(G) + \tau^2 + \|h\|^2}}\, \sigma_n(G), \tag{2.1}$$

where $\tau = \min_y \|h - Gy\|$.

PROOF. If $\sigma_n(G) = \sigma_{n+1}([G,h])$, the statement is trivial. Assume $\sigma_n(G) > \sigma_{n+1}([G,h])$. Then the smallest eigenvalue λ of the matrix $[G,h][G,h]^T = GG^T + hh^T$ (which equals $\sigma_{n+1}^2([G,h])$) satisfies the secular equation

$$1 + \|h\|^2 \sum_{j=1}^{n} \frac{w_j^2}{\sigma_j^2 - \lambda} - \frac{\tau^2}{\lambda} = 0. \tag{2.2}$$

Here σ_j, $j = 1, 2, \ldots, n$, are the singular values of the matrix G, $w \equiv (w_1, \ldots, w_n) = U_1^T h/\|h\|$, and U_1 is the m by n matrix whose columns are the left singular vectors of G (see, for example, [5]). Considering

$$\sum_{j=1}^{n} \frac{w_j^2}{\sigma_j^2 - \lambda} \leq \frac{1}{\sigma_n^2 - \lambda},$$

we have

$$\begin{aligned} 1 + \|h\|^2 \frac{1}{\sigma_n^2 - \lambda} - \frac{\tau^2}{\lambda} &\geq 0, \\ \lambda^2 - (\sigma_n^2 + \tau^2 + \|h\|^2)\lambda + \tau^2\sigma_n^2 &\leq 0. \end{aligned}$$

Solving this quadratic for λ, we find

$$\begin{aligned} \lambda &\geq \frac{\sigma_n^2 + \tau^2 + \|h\|^2 - \sqrt{(\sigma_n^2 + \tau^2 + \|h\|^2)^2 - 4\tau^2\sigma_n^2}}{2} \\ &\geq \frac{\tau^2\sigma_n^2}{\sigma_n^2 + \tau^2 + \|h\|^2}, \end{aligned}$$

and the result (2.1) follows upon taking square roots. □

COROLLARY 2.2. *The smallest singular value of $[v_1, AV_n]$ satisfies*

$$\sigma_{n+1}([v_1, AV_n]) \geq \frac{\|\hat{r}_n\|}{\varrho} \frac{\sigma_n(AV_n)}{\sqrt{2 + \sigma_n^2(AV_n)}}, \tag{2.3}$$

where $\|\hat{r}_n\|$ is defined as $\|\hat{r}_n\| = \varrho \min_y \|v_1 - AV_n y\|$.

PROOF. If AV_n does not have rank n, then the result is trivial, so assume $\text{rank}(AV_n) = n$. The quantity $\|\hat{r}_n\|/\varrho$ plays the role of τ in Theorem 2.1. Since $\|v_1\| = 1$ and $\|\hat{r}_n\|/\varrho \leq 1$, the denominator in (2.1) is less than or equal to $\sqrt{2 + \sigma_n^2}$. □

With (2.3) and the assumption $\|A\| \equiv \sigma_1(A) = 1$, we can now bound the condition number of the matrix $[v_1, AV_n]$.

THEOREM 2.3. *Let $\|\hat{r}_n\| = \varrho \min_y \|v_1 - AV_n y\|$, V_n be the matrix of basis vectors computed in a finite precision MGS-Arnoldi computation. Assume that $\|A\| = 1$, that the columns of V_n have norm 1, that inequality* (1.6) *holds, and that $\zeta_2 N^{3/2}\varepsilon \le 1/2$, where ζ_2 is the constant in* (1.6). *Assume also that*

$$\|\hat{r}_n\|/\varrho \ge 2\sqrt{15}\zeta_2 n N^{3/2}\kappa(A)\varepsilon. \tag{2.4}$$

Then

$$\kappa([v_1, AV_n]) \le \frac{\sqrt{15}\varrho\kappa(A)}{\|\hat{r}_n\|}. \tag{2.5}$$

PROOF. The smallest singular value of AV_n satisfies

$$\sigma_n(AV_n) \ge \sigma_N(A)\sigma_n(V_n).$$

Let κ_n denote the condition number $\kappa([v_1, AV_{n-1}])$. It follows from (1.6) that

$$\sigma_n(V_n) \ge \sqrt{1 - \zeta_2 n N^{3/2}\kappa_n\varepsilon}. \tag{2.6}$$

Since the function $\sigma/\sqrt{2+\sigma^2}$ in (2.3) is an increasing function of σ, it follows from (2.6) that (2.3) can be replaced by

$$\sigma_{n+1}([v_1, AV_n]) \ge \frac{\|\hat{r}_n\|}{\varrho} \frac{\sigma_N(A)\sqrt{1 - \zeta_2 n N^{3/2}\kappa_n\varepsilon}}{\sqrt{2 + \sigma_N^2(A)(1 - \zeta_2 n N^{3/2}\kappa_n\varepsilon)}}. \tag{2.7}$$

The largest singular value of $[v_1, AV_n]$ is bounded above by the Frobenius norm of the matrix, and so we have

$$\sigma_1([v_1, AV_n]) \le \sqrt{\|v_1\|^2 + \|A\|^2\|V_n\|^2} \le \sqrt{1 + \sigma_1^2(V_n)}.$$

It also follows from (1.6) that

$$\sigma_1^2(V_n) \le 1 + \zeta_2 n N^{3/2}\kappa_n\varepsilon,$$

and so we can write

$$\sigma_1([v_1, AV_n]) \le \sqrt{2 + \zeta_2 n N^{3/2}\kappa_n\varepsilon}. \tag{2.8}$$

Combining (2.7) and (2.8) gives

$$\kappa([v_1, AV_n]) \le \frac{\varrho\kappa(A)}{\|\hat{r}_n\|}\sqrt{\frac{(2 + \zeta_2 n N^{3/2}\kappa_n\varepsilon)(2 + \sigma_N^2(A)(1 - \zeta_2 n N^{3/2}\kappa_n\varepsilon))}{1 - \zeta_2 n N^{3/2}\kappa_n\varepsilon}}. \tag{2.9}$$

First, *assume* that $1 - \zeta_2 n N^{3/2}\kappa_n\varepsilon \ge 1/2$. Using this and the bound $\sigma_N^2(A) \le 1$ in (2.9) gives the result (2.5). Now, for $n = 1$, since $\kappa_1 = 1$, we have by assumption that $1 - \zeta_2 N^{3/2}\kappa_1\varepsilon \ge 1/2$, and hence (2.5) holds for $n = 1$. With the assumption (2.4), which must also hold at steps $j < n$, since $\|\hat{r}_j\|$ decreases with j, we can now conclude that $1 - \zeta_2 2N^{3/2}\kappa_2\varepsilon \geq 1/2$, and proceeding in this way, we see, by induction, that the necessary assumption holds and (2.5) is proved. □

We end up this section with the following summary of the bounds for the loss of orthogonality among the Arnoldi basis vectors and bounds for the singular values of V_n. All of them are immediate consequences of Theorem 2.3.

COROLLARY 2.4. *Denote*

$$\delta_n = \sqrt{15}\zeta_2 n N^{3/2}\varepsilon \frac{\varrho\kappa(A)}{\|\hat{r}_n\|}.$$

Then, under the assumptions of Theorem 2.3, $\delta_n \leq 1/2$,

$$\|I - V_{n+1}^T V_{n+1}\| \leq \delta_n, \qquad \|V_{n+1} - \hat{V}_{n+1}\| \leq \delta_n,$$

$$\sqrt{1-\delta_{n-1}} \leq \sigma_n(V_n) \leq \sigma_1(V_n) \leq \sqrt{1+\delta_{n-1}},$$

where the matrix $\hat{V}_{n+1}$ *is the closest orthonormal matrix to* V_{n+1} *in any unitarily invariant norm.*

3 Comparison of the residuals.

The preceding results relate the conditioning of the Arnoldi vectors to the size of $\|\hat{r}_n\|/\varrho$. In this section we show that until the last term reaches the level proportional to $\kappa(A)\varepsilon$, the norms of the true residual, $\|r_n\| = \|b - Ax_n\|$ and the Arnoldi residual $\|t_n\| = \|\varrho e_1 - H_{n+1,n}y_n\|$, are approximately the same as $\|\hat{r}_n\|$. In the following bounds we present only those terms which are linear in ε and do not account for the terms proportional to the higher powers of ε.

THEOREM 3.1. *Under the assumptions of Theorem* 2.3, *and assuming moreover that* $\eta(n,l,N)\varepsilon\kappa(A) \ll 1/\sqrt{2}$, *the norm of the Arnoldi residual is bounded by*

$$\frac{1}{\sqrt{1+\delta_n}}\|\hat{r}_n\| - \varrho\chi_3 \leq \|\varrho e_1 - H_{n+1,n}y_n\| \leq \frac{1}{\sqrt{1-\delta_n}}\left(\|\hat{r}_n\| + \varrho\chi_3\right)$$

where

$$\chi_3 = \frac{\zeta_{11}\sqrt{1+\delta_{n-1}}\eta(n,l,N)\varepsilon\kappa(A)}{\sqrt{1-\delta_{n-1}} - \zeta_{12}\sqrt{1+\delta_{n-1}}\eta(n,l,N)\varepsilon\kappa(A)}. \tag{3.1}$$

PROOF. It is easy to see that under the assumptions of Theorem 2.3,

$$\sigma_1(AV_n) \leq \sqrt{1+\delta_{n-1}}\sigma_1(A), \quad \sigma_n(AV_n) \geq \sqrt{1-\delta_{n-1}}\sigma_N(A);$$

the matrices V_{n+1} and $AV_n = V_{n+1}H_{n+1,n} + F_n$ have full column rank. From $AV_n = \hat{V}_{n+1}H_{n+1,n} + \hat{F}_n$ it follows

$$\begin{aligned}
\sigma_1(H_{n+1,n}) &\leq \sigma_1(AV_n) + \|\hat{F}_n\| \leq \sqrt{1+\delta_{n-1}}\|A\| + \zeta_1\eta(n,l,N)\varepsilon\|A\|,\\
\sigma_n(H_{n+1,n}) &\geq \sigma_n(AV_n) - \|\hat{F}_n\| \geq \sqrt{1-\delta_{n-1}}\sigma_N(A) - \zeta_1\eta(n,l,N)\varepsilon\|A\|.
\end{aligned}$$

Similarly, from $AV_n = V_{n+1}H_{n+1,n} + F_n$,

$$\sigma_n(V_{n+1}H_{n+1,n}) = \sigma_n(AV_n - F_n) \geq \sqrt{1-\delta_{n-1}}\sigma_N(A) - \zeta_1\eta(n,l,N)\varepsilon\|A\|.$$

Assuming $\eta(n,l,N)\varepsilon\kappa(A) \ll 1/\sqrt{2}$, it follows that the matrices $H_{n+1,n}$ and $V_{n+1}H_{n+1,n}$ have also full column rank. We will use the perturbation theory for the least squares problem. Using (1.5),

$$\|\hat{r}_n\| = \varrho \min_y \|v_1 - (V_{n+1}H_{n+1,n} + F_n)y\|,$$

which can be considered as a perturbation of the following least squares problem

$$\|\tilde{r}_n\| = \varrho \min_y \|v_1 - V_{n+1}H_{n+1,n}y\|. \tag{3.2}$$

Applying the perturbation theorem proved by Wedin [10], see also [3], to the problem (3.2), we obtain

$$\|\tilde{r}_n - \hat{r}_n\| \leq \frac{\varrho\|F_n\|}{\|V_{n+1}H_{n+1,n}\|}\left\{\|V_{n+1}H_{n+1,n}\|\|\tilde{y}_n\| + \kappa(V_{n+1}H_{n+1,n})\|\tilde{r}_n\|/\varrho\right\},$$

which, considering $\|\tilde{y}_n\| \leq 1/\sigma_n(V_{n+1}H_{n+1,n})$ and $\|\tilde{r}_n\|/\varrho \leq 1$ gives

$$\|\tilde{r}_n - \hat{r}_n\| \leq \varrho\frac{\zeta_3\|F_n\|}{\sigma_n(V_{n+1}H_{n+1,n})}, \quad 1 \leq \zeta_3 \leq 2,$$

or

$$\|\tilde{r}_n - \hat{r}_n\| \leq \varrho\frac{\zeta_3\zeta_1\eta(n,l,N)\varepsilon\kappa(A)}{\sqrt{1-\delta_{n-1}} - \zeta_1\eta(n,l,N)\varepsilon\kappa(A)}. \tag{3.3}$$

Moreover, from the definition of $\tilde{r}_n$ we have

$$\|\tilde{r}_n\|/\sqrt{1+\delta_n} \leq \varrho\ \min_y\|e_1 - H_{n+1,n}y\| \leq \|\tilde{r}_n\|/\sqrt{1-\delta_n}. \tag{3.4}$$

The transformed least squares problem (1.4) is solved by using the QR decomposition of the matrix $H_{n+1,n}$ via Givens rotations [8]. Following the results by Lawson and Hanson [7], the computed solution y_n represents the exact solution of a perturbed problem

$$\|\varrho e_1 + \Delta s_n - (H_{n+1,n} + \Delta H_n)y_n\| = \min_y \|\varrho e_1 + \Delta s_n - (H_{n+1,n} + \Delta H_n)y\|,$$

where

$$\|\Delta H_n\|/\|H_{n+1,n}\| \leq \zeta_4 n^{5/2}\varepsilon, \quad \|\Delta s_n\|/\varrho \leq \zeta_4 n^{5/2}\varepsilon.$$

Using the results of the rounding error analysis of the Givens rotations (similarly to [6], rel. (3.5)–(3.11)), we have the estimate for the norm of the vector y_n

$$\|y_n\| \leq \varrho\,(1+\zeta_5 n\varepsilon)[\sigma_n(H_{n+1,n}) - 2\zeta_6 n^{3/2}\varepsilon\|H_{n+1,n}\|]^{-1}. \tag{3.5}$$

We assume here that $\eta(n,l,N)\varepsilon\kappa(A)$ is small enough, and thus, after substituting the bounds for the singular values of $H_{n+1,n}$, the vector $\|y_n\|$ is reasonably bounded. Then, for the Arnoldi residual we may write

$$\begin{aligned}(3.6)\quad & |\|\varrho e_1 - H_{n+1,n}y_n\| - \min_y\|\varrho e_1 - H_{n+1,n}y\|| \\ & \leq \left|\min_y\|\varrho e_1 + \Delta s_n - (H_{n+1,n} + \Delta H_n)y\| - \min_y\|\varrho e_1 - H_{n+1,n}y\|\right| \\ & + \|\Delta H_n\|\|y_n\| + \|\Delta s_n\|.\end{aligned}$$

Applying again the perturbation theorem for the least squares problem, we obtain the bound

$$\begin{aligned}(3.7)\quad & \left|\min_y\|\varrho e_1 + \Delta s_n - (H_{n+1,n} + \Delta H_n)y\| - \min_y\|\varrho e_1 - H_{n+1,n}y\|\right| \\ & \leq \varrho\,\zeta_4 n^{5/2}\varepsilon\left\{\frac{\|H_{n+1,n}\|}{\sigma_n(H_{n+1,n})} + \kappa(H_{n+1,n})\min_y\|e_1 - H_{n+1,n}y\|\right\} \\ & \leq \varrho\,\zeta_7 n^{5/2}\varepsilon\kappa(H_{n+1,n}).\end{aligned}$$

Substituting (3.7) and (3.5) into (3.6), we get after a simple manipulation

$$\begin{aligned}(3.8)\quad & \left|\|\varrho e_1 \quad - \quad H_{n+1,n}y_n\| - \min_y\|\varrho e_1 - H_{n+1,n}y\|\right| \\ & \leq \varrho\frac{\zeta_8 n^{5/2}\varepsilon\|H_{n+1,n}\|}{\sigma_n(H_{n+1,n}) - 2\zeta_6 n^{3/2}\varepsilon\|H_{n+1,n}\|}.\end{aligned}$$

Considering the assumption $\eta(n,l,N)\varepsilon\kappa(A) \ll 1$, and the bounds for the singular values of the matrix $H_{n+1,n}$, it follows

$$\begin{aligned}(3.9)\quad \zeta_8 n^{5/2}\varepsilon\|H_{n+1,n}\| & \leq \zeta_8 n^{5/2}\varepsilon(\sqrt{1+\delta_{n-1}} + \zeta_1\eta(n,l,N)\varepsilon)\|A\| \\ & \leq \zeta_8\sqrt{1+\delta_{n-1}}n^{5/2}\varepsilon\|A\| + O(\varepsilon^2) \\ & \leq \zeta_9\sqrt{1+\delta_{n-1}}n^{5/2}\varepsilon\|A\|\end{aligned}$$

and

$$\begin{aligned}(3.10)\quad & \sigma_n(H_{n+1,n}) - 2\zeta_6 n^{3/2}\varepsilon\|H_{n+1,n}\| \geq \sigma_N(A)\sqrt{1-\delta_{n-1}} - \zeta_1\eta(n,l,N)\varepsilon\|A\| \\ & - 2\zeta_6 n^{3/2}\varepsilon(\sqrt{1+\delta_{n-1}} + \zeta_1\eta(n,l,N)\varepsilon)\|A\| \\ & \geq \sigma_N(A)[\sqrt{1-\delta_{n-1}} - \zeta_{10}\sqrt{1+\delta_{n-1}}\eta(n,l,N)\varepsilon\kappa(A)]\end{aligned}$$

for some positive constants ζ_9 and ζ_{10}. Consequently,

$$(3.11)\quad \left|\|\varrho e_1 - H_{n+1,n}y_n\| - \min_y\|\varrho e_1 - H_{n+1,n}y\|\right| \leq \varrho\chi_1,$$

where

$$(3.12)\quad \chi_1 = \frac{\zeta_9\sqrt{1+\delta_{n-1}}n^{5/2}\varepsilon\kappa(A)}{\sqrt{1-\delta_{n-1}} - \zeta_{10}\sqrt{1+\delta_{n-1}}\eta(n,l,N)\varepsilon\kappa(A)}.$$

Combining (3.11) with (3.4) and (3.3), and considering that

$$\chi_1 + \frac{\chi_2}{\sqrt{1+\delta_n}} \leq \chi_3, \qquad \chi_1 + \frac{\chi_2}{\sqrt{1-\delta_n}} \leq \frac{\chi_3}{\sqrt{1-\delta_n}},$$

where χ_2 is defined by

$$\chi_2 = \frac{\zeta_3\zeta_1\eta(n,l,N)\varepsilon\kappa(A)}{\sqrt{1-\delta_{n-1}} - \zeta_1\eta(n,l,N)\varepsilon\kappa(A)}$$

we obtain the statement of the theorem. □

It remains to prove that the norms of the Arnoldi residual t_n and the true residual r_n are also close. It is done by the following theorem (for details we refer to [6]).

THEOREM 3.2. *Under the assumptions of Theorem 2.3, and assuming moreover that* $\eta(n,l,N)\varepsilon\kappa(A) \ll 1/\sqrt{2}$, *the norm of the true GMRES residual is bounded by*

$$\begin{gathered}\sqrt{1-\delta_n}\,\|\varrho e_1 - H_{n+1,n}y_n\| - \varrho\chi_4 - \nu(A,b,x_0)\\ \leq \|b - Ax_n\| \leq \\ \sqrt{1+\delta_n}\,\|\varrho e_1 - H_{n+1,n}y_n\| + \varrho\chi_4 + \nu(A,b,x_0),\end{gathered}$$

where

$$\chi_4 = \frac{\zeta_{14}\sqrt{1+\delta_{n-1}}\eta(n,l,N)\varepsilon\kappa(A)}{\sqrt{1-\delta_{n-1}} - \zeta_{10}\sqrt{1+\delta_{n-1}}\eta(n,l,N)\varepsilon\kappa(A)} \tag{3.13}$$

and $\nu(A,b,x_0) = O(lN^{1/2}+N)\varepsilon\|A\|\|x_0\| + O(N)\varepsilon\|b\|$.

PROOF. The computed approximation x_n can be written in the form

$$x_n = x_0 + V_n y_n + d_n, \tag{3.14}$$

where d_n substitutes the local rounding errors in the computation of the vector x_n and is bounded by

$$\|d_n\| \leq O(n^{3/2})\varepsilon\|y_n\| + \varepsilon\|x_0\|.$$

Then using (3.14), (1.5) it follows

$$\begin{aligned} r_n &= b - Ax_n = r_0 - AV_n y_n - Ad_n \\ &= (b - Ax_0 - \varrho v_1) + V_{n+1}(\varrho e_1 - H_{n+1,n}y_n) - F_n y_n - Ad_n. \end{aligned} \tag{3.15}$$

The first term in (3.15) represents the rounding errors in the computation of the initial vector v_1 from the initial residual r_0 and can be bounded as

$$\|b - Ax_0 - \varrho v_1\| \leq O(lN^{1/2}+N)\varepsilon\|A\|\|x_0\| + O(N)\varepsilon\|b\|.$$

Consequently (cf. [6], relation (3.4)),

$$\|b - Ax_n - V_{n+1}(\varrho e_1 - H_{n+1,n}y_n)\| \leq \zeta_{13}\eta(n,l,N)\varepsilon\|A\|\|y_n\| + \nu(A,b,x_0).$$

Using (3.5) and (3.10), we get

$$\|b - Ax_n - V_{n+1}(\varrho e_1 - H_{n+1,n}y_n)\| \le \varrho\chi_4 + \nu(A,b,x_0),$$

where χ_4 is defined by (3.13). Considering

$$\big|\|b - Ax_n\| - \|V_{n+1}(\varrho e_1 - H_{n+1,n}y_n)\|\big| \le \|b - Ax_n - V_{n+1}(\varrho e_1 - H_{n+1,n}y_n)\|,$$

this gives the assertion of the theorem. □

Note that, under the given assumptions, the statements of Theorems 3.1, 3.2 can be written in the form

$$\frac{1}{\sqrt{1+\delta_n}}\|\hat{r}_n\| - \varrho\omega_1 \le \|t_n\| \le \frac{1}{\sqrt{1-\delta_n}}\|\hat{r}_n\| + \varrho\omega_2, \tag{3.16}$$

$$\sqrt{1-\delta_n}\|t_n\| - \varrho\omega_3 \le \|r_n\| \le \sqrt{1+\delta_n}\|t_n\| + \varrho\omega_3, \tag{3.17}$$

where ω_j, $j = 1,2,3$ are of order $nN^{3/2}\varepsilon\kappa(A)$. Clearly, until $\|\hat{r}_n\|/\varrho$ is of the same order, the relative difference of $\|\hat{r}_n\|$, $\|r_n\|$ and $\|t_n\|$ is small. Note that the quantity $nN^{3/2}\varepsilon\kappa(A)$ which takes part in the bounds counts for the worst possible case and is usually a large overestimate, especially for the "average case".

In most cases (for most right hand sides) the attained accuracy of the HHA or MGSA-GMRES computation measured by the norm of the relative residual $\|r_n\|/\varrho$ is much below the level $\kappa(A)\varepsilon$. The linear independence of the computed Arnoldi vectors is well preserved until this final accuracy is reached (see Figures 1.1–1.4). It is known, but yet unexplained, that sometimes the norm of the Arnoldi residual converges to zero (for the HHA-GMRES this represents a typical behaviour, see Figures 1.1, 1.3) even after the true residual norm stagnates. For MGSA GMRES, however, the norms of the true and Arnoldi residuals stagnate obviously at about the same level.

4 Concluding remarks.

It was proved that the loss of orthogonality among the Arnoldi vectors computed via the modified Gram-Schmidt orthogonalization is related to the size of the residual of the corresponding GMRES run. This relation explains the reliable numerical performance of the modified Gram-Schmidt implementation of GMRES which has been known for a long time. The detailed comparison with the Householder GMRES still needs further work which should include quantitative bounds for the distance between the corresponding Krylov subspaces.

It can be shown numerically that an analogous relation between the loss of orthogonality among the Arnoldi vectors and decrease of the GMRES residual holds true even for the classical Gram-Schmidt GMRES implementation (the final level of accuracy is, however, much worse in the latter case). What makes this observation attractive is the fact that the classical Gram-Schmidt is easily parallelizable. We will work in these directions and report the results elsewhere.

REFERENCES

1. M. Arioli and C. Fassino, *Roundoff error analysis of algorithms based on Krylov subspace methods*, BIT, 36 (1996), pp. 189–206.
2. Å. Björck, *Solving linear least squares problems by Gram-Schmidt orthogonalization*, BIT, 7 (1967), pp. 1–21.
3. Å. Björck, *Stability analysis of the method of seminormal equations for linear least squares problems*, Linear Algebra Appl., 88/89 (1987), pp. 31–48.
4. 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.
5. J. R. Bunch and Ch. P. Nielsen, *Updating the singular value decomposition*, Numer. Math., 31 (1978), pp. 111–129.
6. J. Drkošová, A. Greenbaum, M. Rozložník, and Z. Strakoš, *Numerical stability of the GMRES method*, BIT, 35 (1995), pp. 309–330.
7. C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1974.
8. 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.
9. H. F. Walker, *Implementation of the GMRES method using Householder transformations*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 152–163.
10. P. Å. Wedin, *Perturbation theory for pseudoinverses*, BIT, 13 (1973), pp. 217–232.