Residuals and Singular Values in Linear Least Squares Problems

Paige, Ch.
1998

Dostupný z `http://www.nusl.cz/ntk/nusl-33834`

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Residuals and Singular Values in Linear Least Squares Problems

Christopher. C. Paige   Zdeněk Strakoš

Technical report No. 765

1998

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic
phone: (+4202) 6884244   fax: (+4202) 8585789
e-mail: strakos@uivt.cas.cz

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Residuals and Singular Values in Linear Least Squares Problems

Christopher. C. Paige[1] Zdeněk Strakoš[2]

Technical report No. 765
1998

Abstract

For any linear least squares problem where the matrix has full column rank, upper and lower bounds on the residual norm are derived in terms of a crucial singular value. This result is particularly applicable to minimum norm iterative methods, and the generalized minimum residual method (GMRES) [11] for solving the linear system $Ax = b$ is closely examined in this context. The bounds have significant implications for the finite precision behavior of the modified Gram-Schmidt version of GMRES.

Keywords

Least squares, singular values, linear equations, large sparse matrices, Krylov methods, iterative solution, GMRES, modified Gram-Schmidt

[1] Computer Science, McGill University, Montreal, Quebec, Canada, H3A 2A7,
E-mail: `chris@cs.mcgill.ca`.
Supported by NSERC of Canada Grant OGP0009236.

[2] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic. E-mail: `strakos@uivt.cas.cz`
Supported by the GA AS CR under grant A2030706 and by the GA CR under grant 205/96/0921.

1 The Basic Bounds

Consider an n by k matrix B of rank k and a nonzero n-dimensional vector c. For the matrix $[c, B]$, a relevant question is how large is that part of c lying in the column space of B — the complement of this is the residual $r = c - By$ of the linear least squares problem. One may also look at the residual norm as quantifying the linear independence of the vector c and the columns of B. Another way of quantifying that independence is to consider the smallest singular value of the matrix $[c, B]$. These ways are related to each other, and it seems interesting to determine this relation.

More specifically, this paper is devoted to the following result, where we introduce possibly non unit γ in $r = c\gamma - By$ for later flexibility.

Theorem 1.1 *Given a scalar $\gamma > 0$, and an n by $k+1$ matrix $[c, B]$ with B of rank k, use $\sigma(\cdot)$ to denote singular values and $\|\cdot\|$ to denote 2-norms. Define*

$$\delta_{\min} \equiv \sigma_{\min}([c, B])/\sigma_{\max}(B) \leq \delta_{\max} \equiv \sigma_{\min}([c, B])/\sigma_{\min}(B) \leq 1, \tag{1.1}$$

and

$$r \equiv c\gamma - By \quad \text{such that} \quad \|r\| = \min_z \|c\gamma - Bz\|. \tag{1.2}$$

If $\delta_{\max} < 1$ then

$$\begin{aligned} \mu_L \equiv \sigma_{\min}([c, B])\,\{\gamma^2 + \|y\|^2\}^{\frac{1}{2}} \quad &\leq \quad \sigma_{\min}([c, B])\,\{\gamma^2 + \frac{\|y\|^2}{1 - \delta_{\min}^2}\}^{\frac{1}{2}} \\ \leq \quad \|r\| \quad &\leq \quad \mu_U \equiv \sigma_{\min}([c, B])\,\{\gamma^2 + \frac{\|y\|^2}{1 - \delta_{\max}^2}\}^{\frac{1}{2}}, \end{aligned} \tag{1.3}$$

where the lowest bound on $\|r\|$ also holds if $\delta_{\max} = 1$.

Proof The conditions of the theorem show

$$r = 0 \;\Leftrightarrow\; c \in \mathcal{R}(B) \;\Leftrightarrow\; \sigma_{\min}([c, B]) = 0, \tag{1.4}$$

where $\mathcal{R}(B)$ denotes the range of B. Here (1.3) holds trivially, so now assume

$$\|r\| > 0, \qquad \sigma_{\min}([c, B]) > 0. \tag{1.5}$$

First consider (1.2). Let B have singular value decomposition (SVD)

$$B = U_k \Sigma V^H, \quad \text{with } \Sigma \equiv \operatorname{diag}(\sigma_1, \ldots, \sigma_k), \quad \sigma_1 \geq \ldots \geq \sigma_k > 0.$$

Choose unitary $U = [U_k, \hat{U}_k]$ so that $\rho > 0$ in

$$U^H[c, B]\begin{bmatrix} 1 & 0 \\ 0 & V \end{bmatrix} = \left[\begin{array}{cc} a & \Sigma \\ \rho & 0 \\ \hline 0 & 0 \end{array}\right] = \left[\begin{array}{c} N \\ \hline 0 \end{array}\right], \tag{1.6}$$

giving

$$U^H r = U^H(c\gamma - By) = \begin{bmatrix} a\gamma - \Sigma V^H y \\ \rho\gamma \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \rho\gamma \\ 0 \end{bmatrix}$$

at the minimum. With $a \equiv (\alpha_1, \ldots, \alpha_k)^T$,

$$\begin{aligned} y = V\Sigma^{-1}a\gamma, \qquad \|y\|^2 &= \gamma^2 \sum_{i=1}^{k} \frac{|\alpha_i|^2}{\sigma_i^2}, && (1.7) \\ \|r\|^2 &= \gamma^2\rho^2. && (1.8) \end{aligned}$$

Now examine the minimum singular value $\sigma \equiv \sigma_{\min}([c, B]) = \sigma_{\min}(N) > 0$ in (1.6). There exists $\begin{pmatrix} w \\ \nu \end{pmatrix} \neq 0$ so that

$$\begin{aligned} NN^H \begin{bmatrix} w \\ \nu \end{bmatrix} = \begin{bmatrix} \Sigma^2 + aa^H & a\rho \\ \rho a^H & \rho^2 \end{bmatrix} \begin{bmatrix} w \\ \nu \end{bmatrix} &= \begin{bmatrix} w \\ \nu \end{bmatrix} \sigma^2, \\ (\Sigma^2 - \sigma^2 I)w + a(a^H w + \rho\nu) &= 0 \\ \rho(a^H w + \rho\nu) &= \sigma^2\nu. && (1.9) \end{aligned}$$

If $\nu = 0$ then $w^H(\Sigma^2 - \sigma^2 I)w + |a^H w|^2 = 0$, which gives $w = 0$, a contradiction, since we have assumed $\delta_{\max} < 1$ and so $\sigma < \sigma_k$ from (1.1). Therefore $\nu \neq 0$ and a simple manipulation of equations (1.9) gives

$$\begin{aligned} 0 &= \rho a^H w + \rho a^H(\Sigma^2 - \sigma^2 I)^{-1} a(a^H w + \rho\nu) \\ &= (\sigma^2 - \rho^2)\nu + \sigma^2\nu\, a^H(\Sigma^2 - \sigma^2 I)^{-1} a, \\ 0 &= (\sigma^2 - \rho^2)/\sigma^2 + a^H(\Sigma^2 - \sigma^2 I)^{-1} a \\ &= 1 - \rho^2/\sigma^2 + \sum_{i=1}^{k} |\alpha_i|^2 (\sigma_i^2 - \sigma^2)^{-1}, \end{aligned}$$

which with (1.8) gives the key equality relating the residual norm to the singular values

$$1 + \sum_{i=1}^{k} \frac{|\alpha_i|^2}{\sigma_i^2 - \sigma^2} \quad = \quad \frac{\|r\|^2}{\gamma^2\sigma^2}. \qquad (1.10)$$

Using the identity

$$1 + \sum_{i=1}^{k} \frac{|\alpha_i|^2}{\sigma_i^2 - \sigma^2} \quad = \quad 1 + \sum_{i=1}^{k} \frac{|\alpha_i|^2}{\sigma_i^2(1 - \sigma^2/\sigma_i^2)}, \qquad (1.11)$$

it is easy to derive the bounds

$$1 + \sum_{i=1}^{k} \frac{|\alpha_i|^2}{\sigma_i^2} \leq 1 + \sum_{i=1}^{k} \frac{|\alpha_i|^2}{\sigma_i^2(1 - \sigma^2/\sigma_1^2)} \leq \frac{\|r\|^2}{\gamma^2\sigma^2} \quad \leq \quad 1 + \sum_{i=1}^{k} \frac{|\alpha_i|^2}{\sigma_i^2(1 - \sigma^2/\sigma_k^2)}, \qquad (1.12)$$

so with (1.7) and (1.1) we obtain (1.3). Finally the lowest bound on $\|r\|$ in (1.3) is seen to hold even when $\delta_{\max} = 1$ by setting $\beta = \gamma$ and $z = y$ in the definition

$$\sigma_{\min}([c, B]) \equiv \min_{\begin{pmatrix}\beta \\ z\end{pmatrix} \neq 0} \frac{\|c\beta - Bz\|}{\{|\beta|^2 + \|z\|^2\}^{\frac{1}{2}}} \leq \frac{\|r\|}{\{\gamma^2 + \|y\|^2\}^{\frac{1}{2}}}.$$

■

The full case where $\delta_{\max} = 1$ turns out to be subtle, and since it is unlikely and not crucial for our main discussion, it is treated in the Appendix. Note that we could have taken $\gamma = 1$ since γ has played no part in the theorem, and in fact just scales each expression in (1.3), but this scaling factor does give us useful flexibility in Section 4.

The reader might find it strange that we gave upper and lower bounds on the easily computable $\|r\|$ in terms of the somewhat difficult to compute $\sigma_{\min}([c, B])$, but we will later show why we think this is so useful. Of course (1.3) also tells us that:

Corollary 1.1 *Under the same conditions as in Theorem 1.1*

$$\hat{\mu}_L \equiv \|r\| / \left\{\gamma^2 + \frac{\|y\|^2}{1 - \delta_{\max}^2}\right\}^{\frac{1}{2}} \leq \sigma_{\min}([c, B]) \leq \|r\| / \left\{\gamma^2 + \frac{\|y\|^2}{1 - \delta_{\min}^2}\right\}^{\frac{1}{2}}$$
$$\leq \hat{\mu}_U \equiv \|r\| / \{\gamma^2 + \|y\|^2\}^{\frac{1}{2}}, \quad (1.13)$$

where it is now the weakest upper bound which also holds when $\delta_{\max} = 1$. ■

A crucial aspect of the theorem is that it gives both an upper and a lower bound on the residual, or singular value. The weaker lower bound in (1.3), or upper bound in (1.13), is sufficient for many uses, and is relatively easy to derive, but the upper bound in (1.3), or lower bound in (1.13), is what gives the theorem its power.

The ratios of singular values $\delta_{\max} \geq \delta_{\min} \geq 0$ appear in the bounds, which would appear to limit their usefulness. Fortunately there are cases where we know these must be small, see for example Section 4. The gaps between upper and lower bounds in (1.3) and (1.13) both depend on $\delta_{\max}$ in an interesting way, as we now show.

Corollary 1.2 *Under the conditions of Theorem 1.1 and Corollary 1.1, and assuming $\|r\| > 0, \sigma_{\min}([c, B]) > 0$, whether in (1.3) we define the relative measures of tightness*

$$\eta \equiv \frac{\|r\| - \mu_L}{\|r\|}, \qquad \zeta \equiv \frac{\mu_U - \mu_L}{\|r\|},$$

or in (1.13) we define the corresponding relative measures of tightness

$$\eta \equiv \frac{\sigma_{\min}([c, B]) - \hat{\mu}_L}{\sigma_{\min}([c, B])}, \qquad \zeta \equiv \frac{\hat{\mu}_U - \hat{\mu}_L}{\sigma_{\min}([c, B])},$$

we have the following bounds on η and ζ

$$\begin{aligned} 0 \leq \ \eta \ &\leq \delta_{\max}^2, \\ 0 \leq \ \zeta \ &\leq \frac{\delta_{\max}^2}{2(1 - \delta_{\max}^2)} \quad \left(\leq \delta_{\max}^2 \ \text{ when } \ \delta_{\max} \leq \frac{\sqrt{2}}{2}\right). \end{aligned} \quad (1.14)$$

Proof For $\eta = (\|r\| - \mu_L)/\|r\|$ we see with (1.3)

$$0 \le \eta \; = \; 1 - \frac{\mu_L}{\|r\|} \le 1 - \frac{\mu_L^2}{\|r\|^2} \le 1 - \frac{\sigma^2_{\min}([c,B])(\gamma^2 + \|y\|^2)}{\sigma^2_{\min}([c,B])(\gamma^2 + \frac{\|y\|^2}{1-\delta^2_{\max}})} \le \delta^2_{\max},$$

while for $\eta = (\sigma_{\min}([c,B]) - \hat{\mu}_L)/\sigma_{\min}([c,B])$, with (1.13) and $\sigma \equiv \sigma_{\min}([c,B])$

$$0 \le \eta \; = \; 1 - \frac{\hat{\mu}_L}{\sigma} \le 1 - \frac{\hat{\mu}_L^2}{\sigma^2} = 1 - \frac{\|r\|^2}{\sigma^2(\gamma^2 + \frac{\|y\|^2}{1-\delta^2_{\max}})} \le 1 - \frac{\sigma^2(\gamma^2 + \|y\|^2)}{\sigma^2(\gamma^2 + \frac{\|y\|^2}{1-\delta^2_{\max}})} \le \delta^2_{\max}.$$

For $\zeta = (\mu_U - \mu_L)/\|r\|$ we see with (1.3)

$$\begin{aligned} 0 \le \zeta \; &= \; \frac{\mu_U - \mu_L}{\|r\|} = \frac{\mu_U^2 - \mu_L^2}{\|r\|(\mu_U + \mu_L)} = \frac{\sigma^2_{\min}([c,B])\|y\|^2}{\|r\|(\mu_U + \mu_L)} \frac{\delta^2_{\max}}{(1-\delta^2_{\max})} \\ &\le \; \frac{\sigma^2_{\min}([c,B])\|y\|^2}{2\mu_L^2} \frac{\delta^2_{\max}}{(1-\delta^2_{\max})} \le \frac{\delta^2_{\max}}{2(1-\delta^2_{\max})}. \end{aligned}$$

For $\zeta = (\hat{\mu}_U - \hat{\mu}_L)/\sigma_{\min}([c,B])$ we see with (1.13) and the above

$$\begin{aligned} 0 \le \zeta \; &= \; \|r\| \left(\frac{1}{\mu_L} - \frac{1}{\mu_U} \right) = \|r\| \frac{\mu_U - \mu_L}{\mu_L \mu_U} = \frac{\|r\|}{\mu_L \mu_U} \frac{\sigma^2_{\min}([c,B])\|y\|^2}{(\mu_U + \mu_L)} \frac{\delta^2_{\max}}{(1-\delta^2_{\max})} \\ &\le \; \frac{\sigma^2_{\min}([c,B])\|y\|^2}{2\mu_L^2} \frac{\delta^2_{\max}}{(1-\delta^2_{\max})} \le \frac{\delta^2_{\max}}{2(1-\delta^2_{\max})}. \end{aligned}$$

■

Thus when $\delta_{\max} \ll 1$ the upper and lower bounds in (1.3) and (1.13) are not only very good, but very good in a relative sense, which is important for small $\|r\|$ or $\sigma_{\min}([c,B])$, (note (1.4)). In fact when $\delta_{\max}$ is small, these relative measures of tightness ζ are even smaller, since $\zeta \le \delta^2_{\max}$ when $\delta_{\max} < .707$. Clearly $\delta_{\max}$ is very interesting.

For small $\delta_{\max}$, (1.3) with Corollary 1.2 gives the useful information

$$\|r\| \; \sim \; \sigma_{\min}([c,B]) \{\gamma^2 + \|y\|^2\}^{\frac{1}{2}}. \tag{1.15}$$

The theorem seems very important in iterative solutions of nonsingular linear systems $Ax = b$, where $B = B_k$, $y = y_k$, and $r = r_k$ will be changing each step. In reasonable iterations (see for example Section 4) with B_k increasing in dimension with k, we will usually have $\sigma_{\min}(B_k) \rightarrow$ constant > 0, while $\sigma_{\min}([c,B_k])$ ultimately converges to zero. Consequently $0 \le \delta_{\min} \le \delta_{\max} \rightarrow 0$, and in (1.3) (with $\gamma \equiv \|r_0\|$)

$$0 \; \le \; \frac{\|r_k\| - \sigma_{\min}([c,B_k]) \{\|r_0\|^2 + \|y_k\|^2\}^{\frac{1}{2}}}{\|r_k\|} = \eta_k \le \delta^2_{\max} \; \rightarrow \; 0, \tag{1.16}$$

which is a strong “asymptotic” relation between minimum residual and minimum singular value.

In Section 2 we will briefly discuss Krylov subspace methods and recall a particular method of this class, the generalized minimum residual method (GMRES)[11], since this is what first motivated the bounds in this paper. In Section 3 we will give the necessary mathematics of GMRES. In Section 4 we will show just why Theorem 1.1 is, in our opinion, so important to our understanding of GMRES and related methods. In Section 5 we will give an extreme example which shows that the unusual case corresponding to $y = 0$ in (1.2), and possibly $\delta_{\max} = 1$ in (1.1), can occur up until the very last step of the GMRES iteration. Section 6 will suggest that for more realistic GMRES iterations we often have at each step $\delta_{\max} \ll 1$ in (1.1), so that the GMRES equivalent of (1.16) holds. The Appendix develops the fully general version of Theorem 1.1, including the case of $\delta_{\max} = 1$ in (1.1).

2 Krylov Subspace Methods

Krylov subspace methods are useful for solving some 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 [1] is a Krylov subspace method designed for solving the eigenproblem of unsymmetric matrices. The generalized minimum residual method (GMRES) [11] uses the Arnoldi iteration and adapts it for solving the linear system $Ax = b$.

Because of its computational expense, GMRES has only limited practical use for solving linear equations, and there are methods which are often far more efficient, see for example Bi-CGSTAB [12], QMR [4, 5] for unsymmetric A, and LSQR [9, 8] for unsymmetric or even rectangular A. But GMRES is interesting to study, simply because it *does* in theory minimize the 2-norm of the residual at each step for the solution approximation x_k from the linear variety $x_0 + \text{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\}$, where x_0 is an initial approximation to the solution x, and $r_0 = b - Ax_0$ is the initial residual. Thus theoretical results on GMRES can for example provide lower bounds for the residuals of other methods using this Krylov subspace.

GMRES is also interesting to study computationally, especially since there appears to be a strong relationship between convergence of GMRES to a small residual, and loss of orthogonality of the Arnoldi vectors computed via finite precision modified Gram-Schmidt (MGS) orthogonalization, see [6]. An understanding of this will be just as important for the practical use of the Arnoldi method as it will be for GMRES itself. The bounds in Theorem 1.1 appear to give the key to understanding the strong relationship between convergence and loss of orthogonality in the finite precision application of MGS-GMRES, see Section 4.

In the remainder of the paper we will use $\sigma_k(X)$ to denote the kth largest singular value of X, $\kappa(X)$ to be the ratio of the largest to the smallest singular value of X, and refer to $\kappa(X)$ briefly as the condition number of X. When it will be helpful, we will use the word "ideally" to refer to a result that would hold using exact arithmetic, and "computationally" or "numerically" to a result of a finite precision computation. The vector of elements i to j of a vector y will be denoted $y_{i:j}$, and e_j denotes the j-th column of the unit matrix I.

3 The GMRES Method

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

$$r_0 = b - Ax_0, \qquad \rho_0 = \|r_0\|, \qquad v_1 = r_0/\rho_0, \tag{3.1}$$

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

$$AV_k = V_{k+1}H_{k+1,k}, \quad V_{k+1}^T V_{k+1} = I_{k+1}, \quad V_{k+1} = [v_1, v_2, \ldots, v_{k+1}]. \tag{3.2}$$

Here $H_{k+1,k}$ is a $k+1$ by k upper Hessenberg matrix with elements h_{ij} where $h_{j+1,j} \neq 0$, $j = 1, 2, \ldots, k-1$. If at any stage $h_{k+1,k} = 0$ we would stop with $AV_k = V_k H_{k,k}$. Computationally we are unlikely to reach such a k, and we stop when we assess the norm of the residual (usually computed as below in (3.6)) is small enough.

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

$$\begin{aligned} r_k &= b - Ax_k = r_0 - AV_k y_k = v_1\rho_0 - V_{k+1}H_{k+1,k}\, y_k \\ &= V_{k+1}(e_1\rho_0 - H_{k+1,k}\, y_k), \end{aligned} \tag{3.3}$$

where y_k solves the linear least squares problem

$$\|r_k\| = \min_y \|e_1\rho_0 - H_{k+1,k}\, y\|. \tag{3.4}$$

To solve (3.4) we apply orthogonal rotations (J_i through the angle θ_i) sequentially to $H_{k+1,k}$ to bring it to upper triangular form S_k:

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

The vectors y_k and r_k ideally then satisfy

$$S_k y_k = (Q_k^T e_1 \rho_0)_{1:k}, \tag{3.5}$$

$$\begin{aligned} \|r_k\| &= |e_{k+1}^T Q_k^T e_1 \rho_0| \\ &= |s_1 s_2 \cdots s_k|\, \|r_0\|, \qquad s_i \equiv \sin\theta_i. \end{aligned} \tag{3.6}$$

The measure (3.6) of the (nonincreasing) residual norm is available without computing y_k, and since y_{k+1} will usually differ in every element from y_k, we do not compute y_k or x_k until we decide the residual is small enough to stop. Mathematically equivalent variants of the GMRES method are described in [10].

4 Application of the New Bounds

Note that Theorem 1.1 says nothing about an iterative method, or where B or c come from, and so is a general result. In applying it to any iterative method we will be interested in the cases when $\delta_{\max} < 1$, so that the upper bound is meaningful in (1.3), and when $\delta_{\max} \ll 1$ so the upper and lower bounds are very close.

With $c = v_1$, $B = AV_k$ and $\gamma = \|r_0\|$, Theorem 1.1 can be applied to the k-th step of GMRES as follows. Note that, by construction, B has full column rank.

Theorem 4.1 Residuals and singular values in GMRES.
If $\sigma_k(X)$ denotes the kth largest singular value of X, and n by n A, r_0, y_k, r_k and V_k are as in the GMRES algorithm (3.1)–(3.4) using exact arithmetic, then when $\delta_k \equiv \sigma_{k+1}([v_1, AV_k])/\sigma_k(AV_k) < 1$,

$$\begin{aligned}\sigma_{k+1}([v_1, AV_k])\,\{\|r_0\|^2 + \|y_k\|^2\}^{\frac{1}{2}} \quad &\le \quad \|r_k\| \\ \le \quad \sigma_{k+1}([v_1, AV_k])\,\{\|r_0\|^2 + (1-\delta_k^2)^{-1}\|y_k\|^2\}^{\frac{1}{2}}, & \qquad (4.1)\end{aligned}$$

and

$$\begin{aligned}\frac{\|r_k\|}{\{\|r_0\|^2 + (1-\delta_k^2)^{-1}\|y_k\|^2\}^{\frac{1}{2}}\sigma_k(AV_k)} \quad &\le \quad \delta_k \\ \le \quad \frac{\|r_k\|}{\{\|r_0\|^2 + \|y_k\|^2\}^{\frac{1}{2}}\sigma_k(AV_k)} & \qquad (4.2) \\ \le \quad \frac{\|r_k\|}{\{\|r_0\|^2 + \|y_k\|^2\}^{\frac{1}{2}}\sigma_n(A)} \quad \le \quad \frac{\|r_k\|}{\|r_0\|}\frac{1}{\sigma_n(A)}, & \qquad (4.3)\end{aligned}$$

where the lower bound in (4.1) and upper bounds in (4.2) and (4.3) also hold if $\delta_k = 1$.

Proof We see $c = v_1$, $B = AV_k$ and $\gamma = \|r_0\|$ satisfy the conditions in Theorem 1.1, and from (3.3) and (3.4) we see that r_k and y_k correspond to r and y in (1.2), so the theorem holds with (1.3) giving (4.1), while (1.13) gives (4.2), and (4.3) follows. ■

We did not include the tighter lower bound in (4.1) as it is not needed in what follows. Note that the result does not depend on orthogonality of the columns of V_k, since Theorem 1.1 says nothing about $B = AV_k$ here except that it has full column rank, but it is necessary for $\|r_k\|$ to be a minimum at each step. It should also be pointed out that due to monotonicity of $\|r_k\|$, possible oscillations in the upper bound (4.1) can be eliminated by taking the minimum

$$\|r_k\| \quad \le \quad \min_{j=1,\ldots,k}\{\sigma_{j+1}([v_1, AV_j])\,\{\|r_0\|^2 + (1-\delta_j^2)^{-1}\|y_j\|^2\}^{\frac{1}{2}}\}. \qquad (4.4)$$

A different upper bound for $\|r_k\|$ in terms of δ_k has already been derived in [6], relation (2.3), but no lower bound was given there. Extending and slightly modifying the approach from [6], we obtain the following upper and lower bounds

$$\begin{aligned}(1/\sqrt{2})\,\sigma_{k+1}([v_1, AV_k])\,\|r_0\|\,(1 + 1/\sigma_1^2(AV_k))^{\frac{1}{2}} \quad &\le \quad \|r_k\| \\ \le \quad \sigma_{k+1}([v_1, AV_k])\,\|r_0\|\,(1 + 1/\sigma_k^2(AV_k))^{\frac{1}{2}} & \qquad (4.5)\end{aligned}$$

and

$$\frac{\|r_k\|}{\|r_0\|\{1 + \sigma_k^2(AV_k)\}^{\frac{1}{2}}} \quad \le \quad \delta_k \quad \le \quad \frac{\sqrt{2}\,\|r_k\|\,\kappa(AV_k)}{\|r_0\|\{1 + \sigma_1^2(AV_k)\}^{\frac{1}{2}}}, \qquad (4.6)$$

where the upper bound in (4.5) and the lower bound in (4.6) hold trivially also for $\delta_k = 1$. The bounds from Theorem 4.1 offer, in general, much deeper insight into the problem and the following discussion is based on them.

Define the relative measure of closeness of $\|r_k\|$ to its lower bound

$$\eta_k \equiv \frac{\|r_k\| - \sigma_{k+1}([v_1, AV_k])\,\{\|r_0\|^2 + \|y_k\|^2\}^{\frac{1}{2}}}{\|r_k\|}. \tag{4.7}$$

From Corollary 1.2 we have $0 \le \eta_k \le \delta_k^2$. This implies that whenever $\delta_k \ll 1$, we have the equivalent of (1.15)

$$\|r_k\| \sim \sigma_{k+1}([v_1, AV_k])\,\{\|r_0\|^2 + \|y_k\|^2\}^{\frac{1}{2}}, \tag{4.8}$$

but perhaps more importantly (since in iterative solution of equations with nonsingular A we expect $\|r_k\| \to 0$), (4.3) shows that if $\|r_k\| \to 0$ then $\delta_k \to 0$, so in (4.7) the *relative* precision η_k of the lower bound goes to zero as the square of δ_k.

For nonsingular A it is easy to see that $\delta_k \ll 1$ is necessary eventually. Indeed,

$$\delta_k \equiv \sigma_{k+1}([v_1, AV_k])/\sigma_k(AV_k) \le \sigma_{k+1}([v_1, AV_k])/\sigma_n(A), \tag{4.9}$$

and when $\sigma_{k+1}([v_1, AV_k])$ is sufficiently small, the upper bound on δ_k becomes smaller than unity. For a general (finite dimensional) problem this seems trivial, but there are extreme possibilities: δ_k may, for example, be close to 1 (or $\delta_k = 1$ in some special cases) for $k = 1, 2, \ldots, n-1$ and $\delta_n = 0$. However, in many practical problems there exists k_0 much smaller than n such that $\delta_k \ll 1$ for $k = k_0, k_0 + 1, \ldots$ and

$$0 \le \eta_k \sim 0 \tag{4.10}$$

holds for $k > k_0$. In other problems $\delta_k \ll 1$ for a number of steps but then suddenly δ_k appears very close to 1. In these cases the smoothed upper bound (4.4) should be considered - it is usually very close to $\|r_k\|$ for all iteration steps k. Typical examples are shown in Section 6.

Now we show why we consider the bounds from Theorem 4.1 so important. As noticed in [3], the Arnoldi process (3.2) ideally gives the QR factorization of $[v_1, AV_k]$, since on defining upper triangular $R_{k+1} \equiv [e_1, H_{k+1,k}]$, we see

$$[v_1, AV_k] = V_{k+1}[e_1, H_{k+1,k}] = V_{k+1}R_{k+1}, \qquad V_{k+1}^T V_{k+1} = I_{k+1}. \tag{4.11}$$

What is more, if the orthogonalization in (3.2) is carried out by the modified Gram-Schmidt technique, then this is easily seen to be *numerically* identical to the QR factorization of $[v_1, \widetilde{AV}_k]$ by MGS, where $\widetilde{AV}_k$ indicates the multiplications Av_j, $j = 1, \ldots, k$, are computed numerically. A parallel statement holds when classical Gram-Schmidt orthogonalization is used in (3.2).

With a computer using finite precision with unit round-off ϵ, the computed vectors $v_1, v_2, \ldots$ tend to lose orthogonality. It was shown by Björck [2] that using MGS in the numerical QR factorization $C = QR$ leads to Q such that

$$\|I - Q^T Q\| \le \kappa(C)\,O(\epsilon),$$

so, from the discussion following (4.11), for the finite precision version of (3.2) we have

$$\|I - V_{k+1}^T V_{k+1}\| \leq \kappa([v_1, AV_k])\, O(\epsilon). \tag{4.12}$$

Note that $\kappa([v_1, AV_k])$ is used here instead of $\kappa([v_1, \widetilde{AV}_k])$, for the justification see [3], [6].

It has been observed that when MGS is used in (3.2), leading to the MGS-GMRES method, loss of orthogonality in V_{k+1} is accompanied by small $\|r_k\|$, see [6]. That is, loss of orthogonality in MGS-GMRES apparently cannot occur before convergence occurs. This fortuitous behavior was analyzed numerically in [6] and a partial explanation was offered there. A much stronger and more complete theoretical explanation of the observed behaviour can be derived from the bounds (4.1)-(4.3).

For this purpose we need to combine our approach described above with the rounding error properties of MGS-GMRES to prove the equivalent of Theorem 4.1 for the quantities computed numerically. We will not attempt to prove it rigorously here, but point out that such a proof would lead to the bound

$$\begin{aligned} \|I - V_{k+1}^T V_{k+1}\| &\lesssim \kappa([v_1, AV_k])\, O(\epsilon) \\ &\sim \|[v_1, AV_k]\|\, (\|r_0\|^2 + \|y_k\|^2)^{\frac{1}{2}} \frac{O(\epsilon)}{\|r_k\|}, \end{aligned} \tag{4.13}$$

which would imply that total loss of orthogonality can only occur if we have a residual norm approaching $\|[v_1, AV_k]\|\, (\|r_0\|^2 + \|y_k\|^2)^{\frac{1}{2}}\, O(\epsilon)$ in size, that is when the residual is effectively negligible. Moreover, for a given A, b, and x_0, the value $\kappa([v_1, AV_k])$ determining in practical computation loss of orthogonality (extensive experimental evidence suggests that the bound (4.12) is sharp) is inversely proportional to the norm of the MGS-GMRES residual, and so this analysis supports what has been observed in practice, see [6].

We point out that the analysis suggests we have obtained important relationships that will be useful not only for analyzing GMRES [11], but also for MINRES [7], as in theory GMRES with symmetric A behaves identically to MINRES. The approach here can also be applied to Krylov subspace methods which minimize other norms, such as minimum error methods, as we now show.

With $V_k = [v_1, \ldots, v_k]$ generated some way, and $r_0 = b - Ax_0 = v_1\|r_0\|$, $x_k = x_0 + V_k y_k$, $r_k = b - Ax_k = r_0 - AV_k y_k$, A nonsingular, if we have for example a method that minimizes $\|A^{-1}r_k\| = \|x - x_k\|$, then taking $\gamma = \rho_0 = \|r_0\|$ and $[c, B] = A^{-1}[v_1, AV_k] = [(x - x_0)/\rho_0, V_k]$ in Theorem 1.1 gives with $\delta_k = \sigma_{k+1}([(x - x_0)/\rho_0, V_k])/\sigma_k(V_k)$ the bounds

$$\begin{aligned} \sigma_{k+1}([(x - x_0)/\rho_0, V_k])\, \{\|r_0\|^2 + \|y_k\|^2\}^{\frac{1}{2}} &\leq \|x - x_k\| \\ \leq \sigma_{k+1}([(x - x_0)/\rho_0, V_k])\, \{\|r_0\|^2 + (1 - \delta_k^2)^{-1}\|y_k\|^2\}^{\frac{1}{2}}&, \end{aligned} \tag{4.14}$$

so the theory holds for more general minimum norm methods than just GMRES. Of course if $V_k^T V_k = I$ then $\sigma_k(V_k) = 1$.

The approach can also be applied to methods which minimize some norm with respect to other Krylov subspaces, such as LSQR [9, 8] for solution of equations with unsymmetric A, or least squares solutions with rectangular A. It may also be useful for methods which are not based on Krylov subspaces.

In Theorem 1.1 the scalar γ is arbitrary to the extent that if we scale it, then both $\|y\|$ and $\|r\|$ in (1.2) will be scaled proportionally, but $[c, B]$ will be unchanged, so the result (1.3) will be the same. Thus it would simplify the theorem to set $\gamma = 1$, but we left γ in (and used $c\gamma$ rather than just c) to parallel the GMRES case more obviously and facilitate our discussion above. However γ can be used for further analysis. Suppose we hold $d \equiv c\gamma$ constant while increasing $\gamma \to \infty$, then $\|y\|$ and $\|r\| = \|d - By\| = \min_z \|d - Bz\|$ will be unchanged, but $\sigma_{\min}([d/\gamma, B]) \to 0$ so $\delta_{\max} \to 0$, and the upper and lower bounds in (1.3) approach each other, and

$$0 \;\leq\; \frac{\|r\| - \sigma_{\min}([d/\gamma, B])\,\{\gamma^2 + \|y\|^2\}^{\frac{1}{2}}}{\|r\|} \;\to\; 0 \qquad \text{as } \gamma \to \infty, \tag{4.15}$$

which shows how this smallest singular value behaves for large γ.

5 Delayed Convergence of GMRES

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

$$A = [e_2\gamma_2, e_3\gamma_3, \ldots, e_n\gamma_n, e_1\gamma_1], \qquad b = e_1\beta_1, \qquad x_0 = 0,$$

with $\gamma_i > 0$, $i = 1, \ldots, n$, then in (3.1) and (3.2) for $k < n$

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

and in (3.3) and (3.4)

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

so any convergence at all is delayed until step $k = n$.

In the application of Theorem 1.1 to GMRES we took $[c, B] = [v_1, AV_k]$, so $y_k = 0$ in GMRES gives $y = 0$ in Theorem 1.1. But here $\sigma_{k+1}([v_1, AV_k]) = \sigma_k(AV_k)$ if any $\gamma_i \leq 1$, $i = 2, 3, \ldots, k+1$, since $\sigma_k(AV_k) = \min\{\gamma_2, \ldots, \gamma_{k+1}\}$, while $\sigma_{k+1}([v_1, AV_k]) = \min\{1, \gamma_2, \ldots, \gamma_{k+1}\}$. This would mean $\delta_{\max} = 1$ in (1.1), and so Theorem 1.1 would not apply, and the fully general version in the Appendix would be required.

6 Behavior of $\delta_k = \sigma_{k+1}([v_1, AV_k])/\sigma_k(AV_k)$ in GMRES

We saw that δ_k plays an important part in the analysis here, so in an attempt to understand it further, we will focus again on GMRES. Section 5 showed it is possible to have $\delta_k = 1$ for all but the last step, and in that example the residual stagnated at $\|r_0\|$ until the final step. If $\delta_k \sim 1$ then there can be a large gap between the upper and lower bounds in (4.1). This does not negate the argument that orthogonality is effectively maintained until convergence in finite precision MGS-GMRES ($\delta_k \ll 1$ is necessary eventually), but it does make us question the tightness of the bounds in (4.1).

Fortunately, experiments suggest that δ_k is frequently quite small during the computation. As k increases δ_k can decrease, then increase, but it must eventually become small, for from (4.3) we see the upper bound on δ_k must decrease as $\|r_k\|$ becomes sufficiently small. The surprising observation was that we often found $\delta_k \ll 1$ from the start, so that

$$\|r_k\| \;\sim\; \sigma_{k+1}([v_1, AV_k])\,\{\|r_0\|^2 + \|y_k\|^2\}^{\frac{1}{2}}$$

throughout such computations. Thus we often have this unexpectedly very close relationship between $\|r_k\|$ and the smallest singular value of $[v_1, AV_k]$. Another interesting experience was that even if $\delta_k \sim 1$ and there was a large gap between the upper and lower bounds in (4.1), the smoothed upper bound (4.4) was always tight. We will illustrate that by presenting results of three numerical experiments showing different types of behaviour of δ_k.

In all experiments matrices from the Harwell-Boeing collection are used. Results for the matrix STEAM1(240), $n = 240$, $\kappa(A) \sim 10^7$, $b = (1, \ldots, 1)^T$ represent the case $\delta_k \ll 1$ from the start to the end. For the matrix IMPCOLE(225), $n = 225$, $\kappa(A) \sim 10^7$, $b = (1, \ldots, 1)^T$ the residual norm decreases very slowly for many steps and then suddenly drops very sharply to its final accuracy level. The value of δ_k is close to 1 for most iteration steps and then follows the sharp drop of the residual norm. Results for the matrix WEST(132), $n = 132$, $\kappa(A) \sim 10^{12}$, $b = Ax$, $x = (1, \ldots, 1)^T$ illustrate oscillations of δ_k.

We have chosen $x_0 = 0$, $r_0 = b$ in all experiments. It is worth to mention that, with this choice, for a given matrix from the Harwell-Boeing collection the results computed for $b = (1, \ldots, 1)^T$ typically differ in both the rate of convergence and the final accuracy from those computed for $x = (1, \ldots, 1)^T$ and that this difference is significant. This fact does not play a role here (we looked for some nontrivial examples illustrating our theoretical results), but the choice of the right hand side and the initial approximation should always be examined while testing numerical software.

Experiments were performed on an SGI Indigo Workstation using MATLAB 5.0, $\epsilon = 1.11 \times 10^{-16}$. Figures 6.1–6.3 give results for STEAM1(240), figures 6.4–6.6 results for IMPCOLE(225), while Figures 6.7–6.9 results for WEST(132).

In Fig. 6.1 solid line shows the relative norm of the directly computed residual $\|b - Ax_k\|/\|r_0\|$, dashed line gives the iteratively computed residual norm (3.6) divided by $\|r_0\|$, dashed-dotted line the normalized norm of the error $\|x - x_k\|/\|x - x_0\|$ and dotted line the loss of orthogonality among the Arnoldi vectors measured in the Frobenius norm $\|I - V_k^T V_k\|_F$. Note the correspondence between the loss of orthogonality and the decrease of the residual norm (similarly in Figures 6.4 and 6.7).

Fig. 6.2 is devoted to the tightness parameters; solid line shows the values of ${\delta_k}^2$, dashed line the values of ζ_k and dotted line the values of η_k. For the graphical reasons the values of ζ_k and η_k less than 10^{-15} were set to this level. Please note that for $k \leq 65$ and $k \geq 170$ the computed values of η_k exhibit large oscillations caused by numerical errors in GMRES calcullations. The difference between the computed upper and lower bounds is during these iteration smaller than accuracy of the computed approximate solution and residual (which is, due to ill-conditioning of the linear least squares problem (3.4) from the start not better than 9 decimal digits). Additionally, loss of orthogonality of the computed Arnoldi vectors will significantly come into effect for $k \geq 170$.

Because of the small values of η_k and ζ_k throughout the computation, the lower (dashed-dotted line) and upper (dashed line) bounds from (4.1) (divided by $\|r_0\|$) are on Fig. 6.3 completely covered by the smoothed upper bound (4.4) (solid line).

On Figures 6.4, 6.5 and 6.6; 6.7, 6.8 and 6.9 the notation is analogous. For IMP-COLE(225) the bounds are tight for the first 50 steps, but $\delta_k \sim 1$ after that and the lower and upper bounds are not close to each other until the sharp drop for $k \sim 205$. Please note that the lower bound is for $50 < k < 205$ much less accurate than the upper bound and that the smoothed upper bound almost coincide with the computed residual norm for all iteration steps k. For $k \geq 205$ the results are heavily affected by the loss of orthogonality (and linear independence) among the computed Arnoldi vectors.

In the last experiment using the matrix WEST(132) the lower and upper bounds significantly differ for $5 \leq k \leq 30$ and $105 \leq k \leq 125$. Moreover, for $k \sim 25$ and $k \sim 115$ the upper bound gives a large overestimate (for $k \sim 25$ the upper bound (4.1) is significantly worse that that of (4.5). The last bound is, however, trivial, because it gives $1 \lesssim \|r_k\|/\|r_0\|$ whenever $\delta_k \sim 1$). Note the corresponding behavior of the tightness parameters on Fig. 6.8. The smoothed upper bounds are again visually indistinguishable from the computed residual norms.

Note that our experiments suggest that the equivalent of Theorem 4.1 for the *numerically computed quantities* holds. However, the statement must be slightly modified to account for the effect of rounding errors, especially for the influence of the loss of orthogonality on the size of the directly computed residuals $\|b - Ax_k\|$. A rigorous proof will require further work and will be given elsewhere.

7 Appendix

Theorem 1.1 would have been logically cleaner if we had assumed $\alpha_k \neq 0$ in a in (1.6), rather than $\delta_{\max} < 1$, since (see Corollary 7.1 here) $\alpha_k \neq 0 \Rightarrow \delta_{\max} < 1$, but not vice versa (note that when $\gamma_i > 1, i = 1, \ldots, k+1$ in the example from Section 5, then $\delta_{\max} < 1$ and $\alpha_k = \alpha_{k-1} = \cdots = \alpha_1 = 0$). However since that would have required Propositions 7.1 and 7.2 here to prove $\delta_{\max} < 1$, we chose the simpler presentation. The much longer, but full and cleaner version is given now.

Denote, as above, the singular values of B by σ_i in nonincreasing order. In (1.6) we unitarily transformed $[c, B]$ to obtain N with the same singular values, see (7.1) below. First we give propositions that will simplify the general analysis.

Proposition 7.1 *Let N be nonsingular, with all $\sigma_i > 0$, in*

$$N = \begin{bmatrix} a & \Sigma \\ \rho & 0 \end{bmatrix}, \qquad \Sigma \equiv \mathrm{diag}(\sigma_1, \ldots, \sigma_k), \qquad a \equiv (\alpha_1, \ldots, \alpha_k)^T. \tag{7.1}$$

If σ_i is a singleton singular value of Σ, that is $\sigma_i \neq \sigma_j$ for $j \neq i$, then σ_i is a singular value of N if and only if $\alpha_i = 0$.

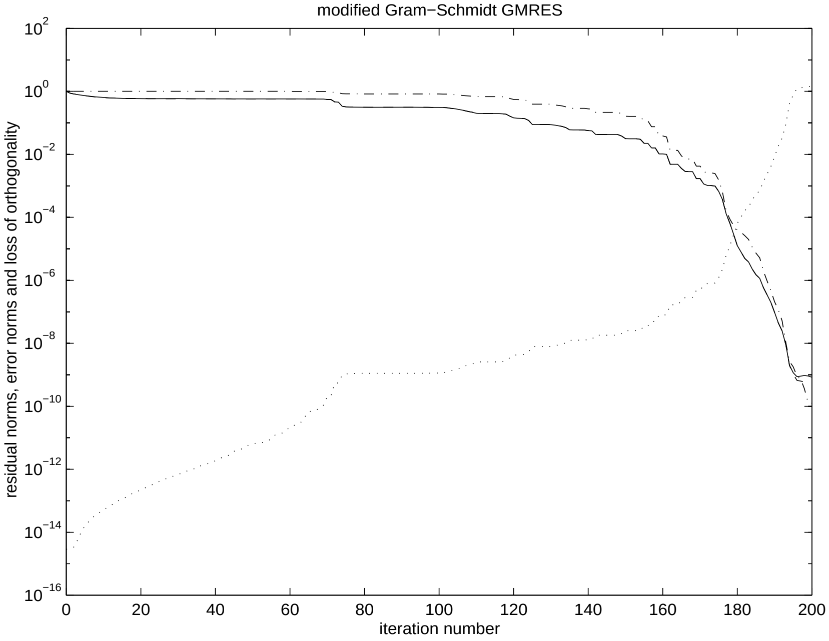

Figure 6.1: Norm of the directly computed relative residual (solid line), iteratively computed relative residual (dashed line), relative error (dashed-dotted line), and loss of orthogonality among the Arnoldi vectors measured in the Frobenius norm (dotted line) for MGS-GMRES applied to STEAM1(240).

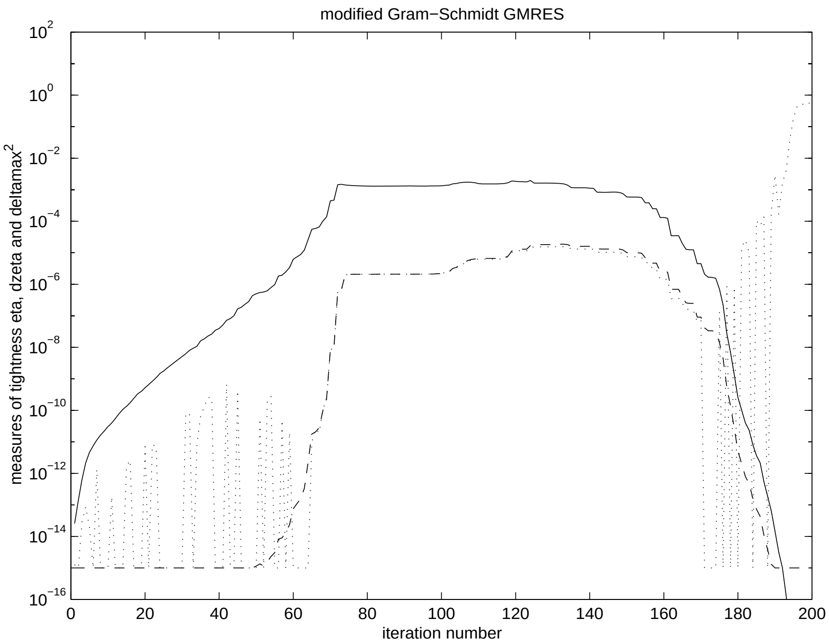

Figure 6.2: Values of the tightness parameters ${\delta_k}^2$ (solid line), η_k (dotted line) and ζ_k (dashed line) for MGS-GMRES applied to STEAM1(240).

Proof If $\alpha_i = 0$ then σ_i is clearly a singular value of N.

Now suppose σ_i is a singular value of N. From (1.9) there exists $\begin{pmatrix} w \\ \nu \end{pmatrix} \neq 0$ so that

$$\begin{aligned} (\Sigma^2 - \sigma_i^2 I)w + a(a^H w + \rho\nu) &= 0 \\ \rho(a^H w + \rho\nu) &= \sigma_i^2 \nu. \end{aligned}$$

Thus $\alpha_i(a^H w + \rho\nu) = 0$. If $(a^H w + \rho\nu) \neq 0$ then $\alpha_i = 0$ as desired. If $(a^H w + \rho\nu) = 0$ then $\nu = 0$, $a^H w = 0$. But with $w = (\omega_1, \ldots, \omega_k)^T$, we also see $\omega_j = 0$ for all $j \neq i$, so $\bar{\alpha}_i \omega_i = 0$. But $\begin{pmatrix} w \\ \nu \end{pmatrix} \neq 0$ so $\omega_i \neq 0$ and $\alpha_i = 0$ as desired. ■

We must also consider the case where the smallest singular value σ_k is repeated in Σ.

Proposition 7.2 *Let N be nonsingular with $\sigma_1 \geq \ldots \geq \sigma_k > 0$ in (7.1). Let $\sigma \equiv \sigma_{k+1}(N)$ be the smallest singular value of N, and let Σ have exactly $s \geq 1$ singular values equal to σ, so $\sigma = \sigma_k$. Then N has the form*

$$\begin{bmatrix} a_{k-s} & \Sigma_{k-s} & \\ 0 & & \sigma I_s \\ \rho & 0 & 0 \end{bmatrix}. \tag{7.2}$$

Proof Take $a_{k-s} \equiv (\alpha_1, \ldots, \alpha_{k-s})^T$, $a_k \equiv (\alpha_{k-s+1}, \ldots, \alpha_k)^T$, $\Sigma_{k-s} \equiv \mathrm{diag}(\sigma_1, \ldots, \sigma_{k-s})$ and assume that some of the elements in a_k, denoted as α

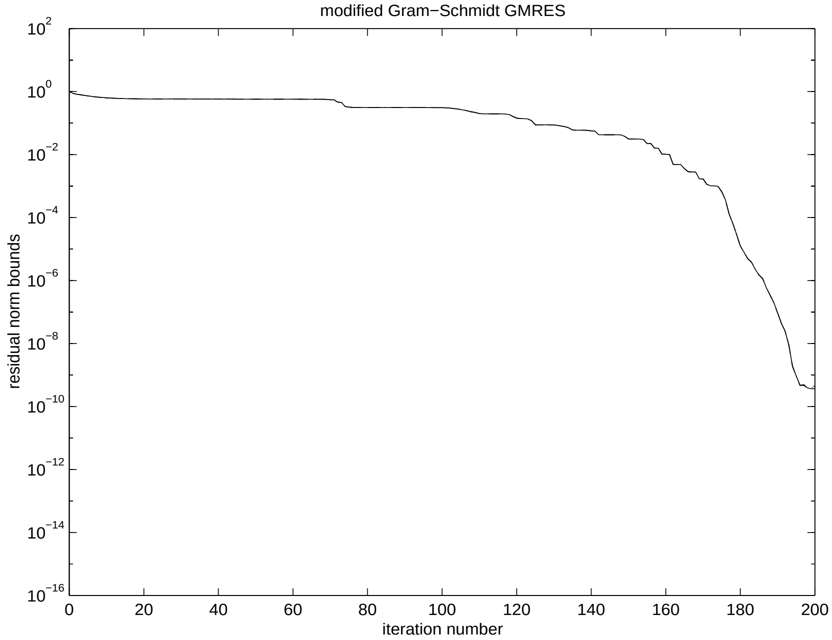

Figure 6.3: Lower bound (dashed-dotted line), upper bound (dashed line) and smoothed upper bound (solid line) for the normalized residual norm computed by MGS-GMRES applied to STEAM1(240). Note that the bounds can not be visually distinguished.

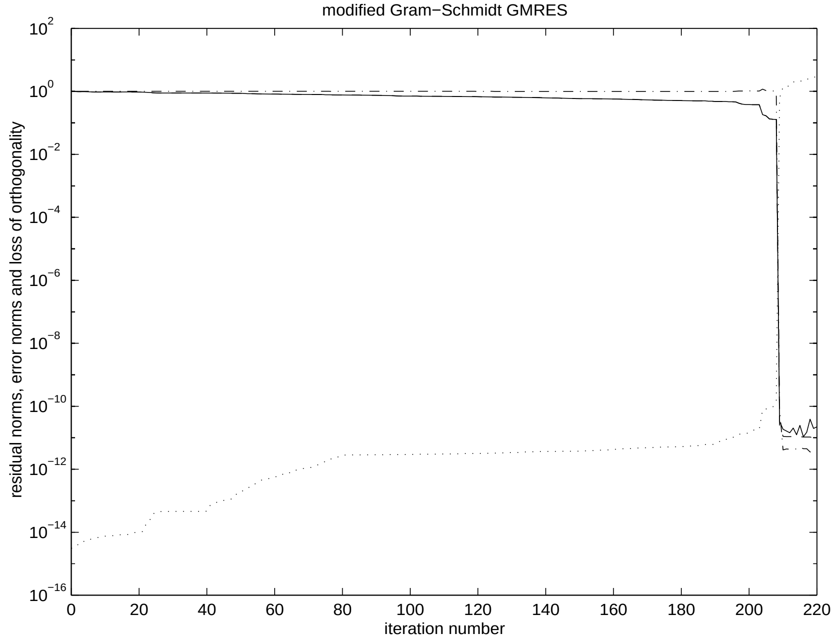

Figure 6.4: Norm of the directly computed relative residual (solid line), iteratively computed relative residual (dashed line) relative error (dashed-dotted line), and loss of orthogonality among the Arnoldi vectors measured in the Frobenius norm (dotted line) for MGS-GMRES applied to IMPCOLE(225).

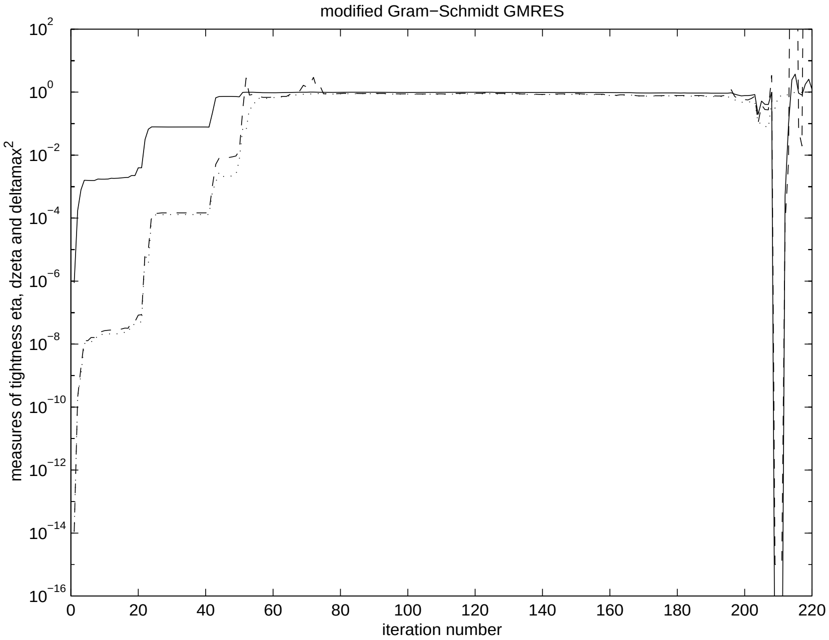

Figure 6.5: Values of the tightness parameters ${\delta_k}^2$ (solid line), η_k (dotted line) and ζ_k (dashed line) for MGS-GMRES applied to IMPCOLE(225).

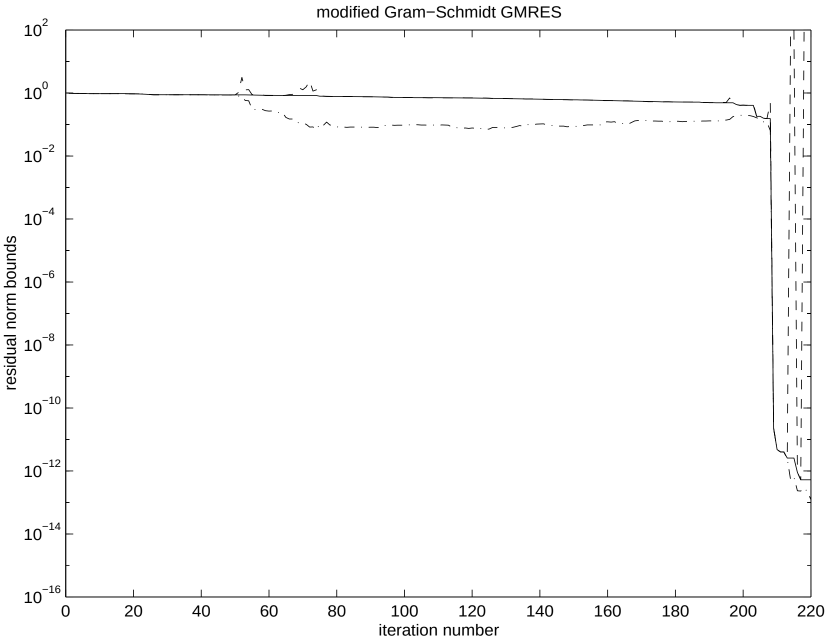

Figure 6.6: Lower bound (dashed-dotted line), upper bound (dashed line) and smoothed upper bound (solid line) for the normalized residual norm computed by MGS-GMRES applied to IMPCOLE(225).

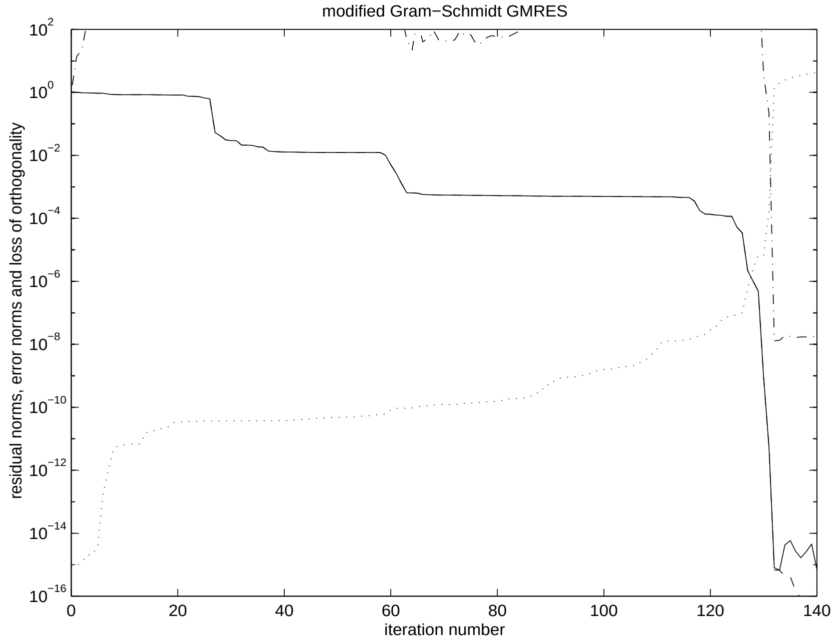

Figure 6.7: Norm of the directly computed relative residual (solid line), iteratively computed relative residual (dashed line), relative error (dashed-dotted line), and the loss of orthogonality among the Arnoldi vectors measured in the Frobenius norm (dotted line) for MGS-GMRES applied to WEST(132).

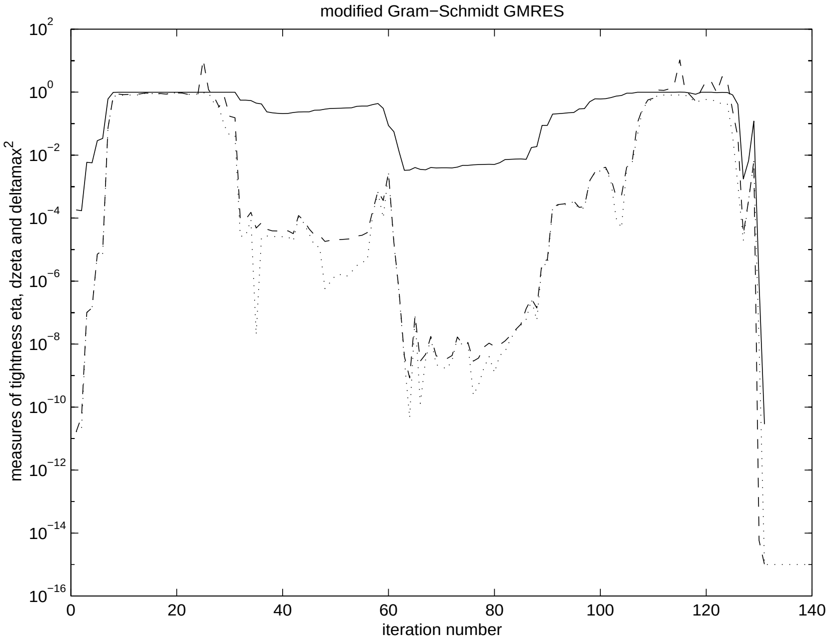

Figure 6.8: Values of the tightness parameters ${\delta_k}^2$ (solid line), η_k (dotted line) and ζ_k (dashed line) for MGS-GMRES applied to WEST(132).

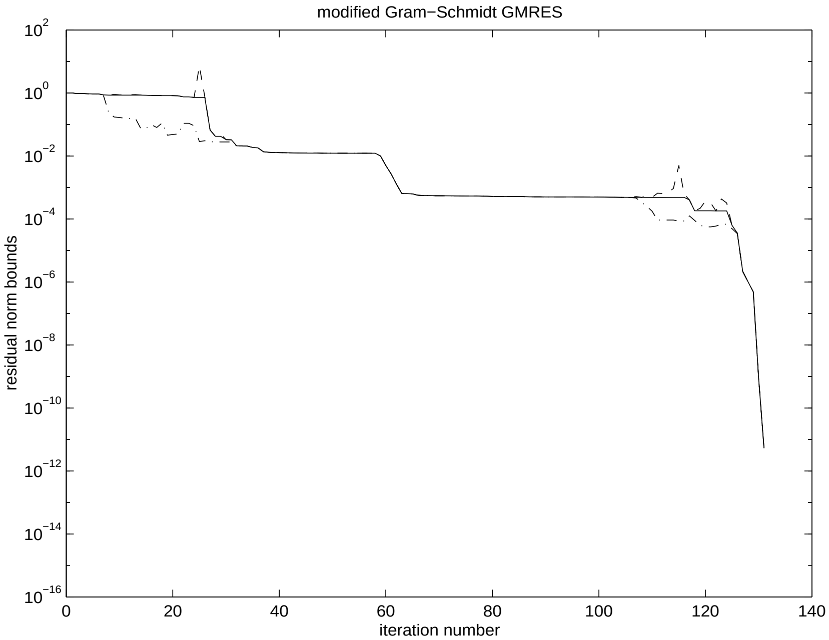

Figure 6.9: Lower bound (dashed-dotted line), upper bound norm (dashed line) and smoothed upper bound (solid line) for the normalized residual norm computed by MGS-GMRES applied to WEST(132).

is nonzero, $\alpha \neq 0$. Then by Proposition 7.1 σ is not a singular value of

$$\tilde{N} = \begin{bmatrix} a_{k-s} & \Sigma_{k-s} & \\ \alpha & & \sigma \\ \rho & 0 & 0 \end{bmatrix}. \tag{7.3}$$

Clearly, $\sigma_{\min}(\tilde{N}) < \sigma$. But by interlacing, $\sigma_{\min}(\tilde{N}) \geq \sigma$, which gives a contradiction. ■

This shows that if the *smallest* singular value of N is equal to exactly s singular values of Σ, then these s singular values must be decoupled from the rest of N, in that for each of these σ_i, $\alpha_i = 0$ in N, see (7.1). We rephrase this for our use.

Corollary 7.1 *Let N be nonsingular in (7.1) with $\sigma_1 \geq \ldots \geq \sigma_k$, and $\delta_k \equiv \sigma_{\min}(N)/\sigma_k$. Then*

$$\sigma_k = \sigma_{\min}(N) \Rightarrow \alpha_k = 0, \tag{7.4}$$

$$\delta_k = 1 \Rightarrow \alpha_k = 0, \tag{7.5}$$

$$\alpha_k \neq 0 \Rightarrow \delta_k < 1. \tag{7.6}$$

Proof (7.4) follows directly from Proposition 7.2; (7.5) is equivalent to (7.4); and (7.6) follows from (7.5) by contraposition and knowing $\delta_k \leq 1$. ■

This leads to the complete version of Theorem 1.1.

Theorem 7.3 *Given a scalar $\gamma > 0$, and an n by $k+1$ matrix $[c, B]$ with B of rank k, use $\sigma_i(\cdot)$ to denote singular values in nonincreasing order and $\|\cdot\|$ to denote 2-norms. Let B have singular value decomposition (SVD)*

$$B = U_k \Sigma V^H, \quad U_k \equiv [u_1, \ldots, u_k], \quad \Sigma \equiv \mathrm{diag}(\sigma_1, \ldots, \sigma_k), \quad \sigma_1 \geq \ldots \geq \sigma_k,$$

and define

$$r \equiv c\gamma - By \quad \text{such that } \rho \equiv \|r\|/\gamma = \min_z \|c\gamma - Bz\|/\gamma. \tag{7.7}$$

If $U_k^H c = 0 = B^H c$ then $y = 0$ and $\rho = \|r\|/\gamma = \|c\|$ is a singular value of $[c, B]$, not necessarily the smallest, and there is no further relation between $\|r\|$ and the singular values of $[c, B]$. Otherwise let j be the largest index such that $u_j^H c \neq 0$, and define $U_j \equiv [u_1, \ldots, u_j]$. Then there is an integer $s \geq 1$ such that

$$\sigma_{j+s}([c, B]) = \sigma_{\min} \begin{pmatrix} U_j^H c & U_j^H B \\ \rho & 0 \end{pmatrix} < \sigma_j \equiv \sigma_j(B), \tag{7.8}$$

giving with the definitions

$$\delta_{\min} \equiv \sigma_{j+s}([c, B])/\sigma_1(B) \leq \delta_{\max} \equiv \sigma_{j+s}([c, B])/\sigma_j(B) < 1. \tag{7.9}$$

The fully general bounds are then

$$\begin{aligned} \sigma_{j+s}([c, B]) \{\gamma^2 + \|y\|^2\}^{\frac{1}{2}} \quad &\leq \quad \sigma_{j+s}([c, B]) \left\{\gamma^2 + \frac{\|y\|^2}{1 - \delta_{\min}^2}\right\}^{\frac{1}{2}} \\ \leq \quad \|r\| \quad &\leq \quad \sigma_{j+s}([c, B]) \left\{\gamma^2 + \frac{\|y\|^2}{1 - \delta_{\max}^2}\right\}^{\frac{1}{2}}. \end{aligned} \tag{7.10}$$

Proof Choose unitary $U = [U_k, \hat{U}_k]$ so that $\hat{\rho} > 0$ in

$$U^H[c, B]\begin{bmatrix} 1 & 0 \\ 0 & V \end{bmatrix} = \left[\begin{array}{cc} a & \Sigma \\ \hat{\rho} & 0 \\ \hline 0 & 0 \end{array}\right] = \left[\begin{array}{c} N \\ \hline 0 \end{array}\right], \quad \text{say.} \tag{7.11}$$

Applying U^H to r gives at the minimum in (7.7)

$$U^H r = U^H(c\gamma - By) = \begin{bmatrix} a\gamma - \Sigma V^H y \\ \hat{\rho}\gamma \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \hat{\rho}\gamma \\ 0 \end{bmatrix}, \quad y = V\Sigma^{-1}a\gamma,$$

¿from which it is clear that $\hat{\rho} = \rho$. If $a = U_k^H c = 0$ then $B^H c = 0$ and ρ is clearly a singular value of $[c, B]$. Otherwise with j the largest index such that $u_j^H c \neq 0$, let $a_j = (\alpha_1, \ldots, \alpha_j)^T$ be the vector of the first j elements of a, and Σ_j the leading j by j submatrix of Σ, so in (7.11)

$$N = \begin{bmatrix} a_j & \Sigma_j & & & \\ 0 & & \sigma_{j+1} & & \\ \cdot & & & \cdot & \\ 0 & & & & \sigma_k \\ \rho & 0 & 0 & \cdot & 0 \end{bmatrix}, \qquad N_j \equiv \begin{bmatrix} a_j & \Sigma_j \\ \rho & 0 \end{bmatrix}. \tag{7.12}$$

Now $\alpha_j \neq 0$ in N_j, so from Corollary 7.1 $\sigma_{\min}(N_j) < \sigma_j$. It is clear that N_j has the same singular values as the middle matrix in (7.8), and from (7.11) and (7.12) that these are also singular values of $[c, B]$. It follows from the ordering of the singular values in Σ that $\sigma_{\min}(N_j) = \sigma_{j+s}([c, B])$ for some integer $s \geq 1$, and (7.8) and so (7.9) hold.

Finally, from the form of a we see that

$$\|y\|^2 = \gamma^2 \sum_{i=1}^{j} \frac{|\alpha_i|^2}{\sigma_i^2}, \qquad \|r\|^2 = \gamma^2\rho^2. \tag{7.13}$$

But since $\sigma_{j+s}([c, B]) = \sigma_{\min}(N_j) < \sigma_j$, we can apply Theorem 1.1 to γ and N_j (rather than γ and $[c, B]$), to obtain (with $\delta_{\min}$ and $\delta_{\max}$ as given in (7.9), see (7.8))

$$\begin{aligned} \sigma_{j+s}([c, B])\,\{\gamma^2 + \|\hat{y}\|^2\}^{\frac{1}{2}} \;&\leq\; \sigma_{j+s}([c, B])\left\{\gamma^2 + \frac{\|\hat{y}\|^2}{1 - \delta_{\min}^2}\right\}^{\frac{1}{2}} \\ &\leq\; \|\hat{r}\| \;\leq\; \sigma_{j+s}([c, B])\left\{\gamma^2 + \frac{\|\hat{y}\|^2}{1 - \delta_{\max}^2}\right\}^{\frac{1}{2}}. \end{aligned} \tag{7.14}$$

where

$$\hat{r} = \begin{pmatrix} a_j\gamma - \Sigma_j\hat{y} \\ \rho\gamma \end{pmatrix} = \begin{pmatrix} 0 \\ \rho\gamma \end{pmatrix}$$

at the minimum, so $\|\hat{r}\| = \|r\|$ and $\|\hat{y}\| = \|y\|$ in (7.13), and (7.14) becomes (7.10), proving our general theorem. ■

The simple idea is that the singular values $\sigma_{j+1}, \ldots, \sigma_k$ of B are decoupled from c, see (7.12), so make no contribution to the residual, and we need only consider N_j. For completeness, we show when the bounds in (7.10) are tight.

Corollary 7.2 *Let the conditions of Theorem 7.3 hold. The bounds in (7.10) are tight if* $\sigma_{\min}([c, B]) = 0$*. Otherwise, let us assume* $B^H c \neq 0$*, so that (7.10) holds. The upper bound in (7.10) is tight if and only if* c *is orthogonal to all left singular vectors of* B *having singular values greater than* $\sigma_j(B)$*, while the stronger lower bound is tight if and only if* c *is orthogonal to all left singular vectors of* B *having singular values less than* $\sigma_1(B)$*.*

Proof See (7.11) with (7.12), and note that (7.10) is proven from (1.12) with k replaced by j. The upper bound is tight if and only if $\alpha_i = 0$ for all $\sigma_i > \sigma_j$, while the stronger lower bound is tight if and only if $\alpha_i = 0$ for all $\sigma_i < \sigma_1$. ■

Bibliography

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

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

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

[4] R. Freund and N. Nachtigal, *QMR: A quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.

[5] ——, *An implementation of the QMR method based on coupled two-term recurrences*, SIAM J. Sci. Statist. Comput., 15 (1994), pp. 313–337.

[6] A. Greenbaum, M. Rozložník, and Z. Strakoš, *Numerical behavior of the modified Gram-Schmidt GMRES implementation*, BIT, 37:3 (1997), pp. 706–719.

[7] C. Paige and M. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.

[8] ——, *Algorithm 583 LSQR: Sparse linear equations and least squares problems*, ACM Trans. Math. Software, 8 (1982), pp. 195–209.

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

[10] M. Rozložník and Z. Strakoš, *Variants of the residual minimizing krylov space methods*, (1996).

[11] Y. Saad and M. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[12] 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.