NUMERICAL LINEAR ALGEBRA WITH APPLICATIONS
Numer. Linear Algebra Appl., **7**, 67–93 (2000)

Further Analysis of Minimum Residual Iterations

Yousef Saad*

Department of Computer Science and Engineering, University of Minnesota, 200 Union Street, SE, Minneapolis, MN 55455, U.S.A.

The convergence behaviour of a number of algorithms based on minimizing residual norms over Krylov subspaces is not well understood. Residual or error bounds currently available are either too loose or depend on unknown constants that can be very large. In this paper we take another look at traditional as well as alternative ways of obtaining upper bounds on residual norms. In particular, we derive inequalities that utilize Chebyshev polynomials and compare them with standard inequalities. Copyright © 2000 John Wiley & Sons, Ltd.

KEY WORDS minimal residual methods; Krylov subspace techniques

1. Introduction

A number of successful algorithms for solving large sparse asymmetric linear systems are based on minimizing the residual norm $\|b - Ax\|$ over trial solutions belonging to small-dimensional subspaces. Under mild conditions on the coefficient matrix A, the approximations provided by these 'minimal residual' (Min-Res) methods is guaranteed to make some progress toward the solution, but convergence can be quite slow.

Two types of results have been developed to analyse convergence of Min-Res methods. The first type consists of inequalities, such as those established by Eisenstat, Elman and Schultz [1], who do not attempt to be sharp but to establish global convergence of the method. A second type of error or residual bounds attempt to imitate the asymptotic behaviour of the method, specifically for Min-Res methods on Krylov subspaces. The most common analysis of this type assumes that A is diagonalizable, $A = XDX^{-1}$, and that its spectrum

* Correspondence to: Yousef Saad, Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, U.S.A.

Contract/grant sponsor: NSF; contract/grant number: CCR-9618827
Contract/grant sponsor: Minnesota Supercomputer Institute

Received 14 November 1998
Revised 3 December 1999

is enclosed in an ellipse $E(c, d, a)$ of centre c, focal distance d and major semi-axis a. The following inequality is then easily shown:

$$\|r_m\| \leq \kappa_2(X) \frac{T_m(a/d)}{T_m(c/d)} \|r_0\| \tag{1.1}$$

in which T_k represents the Chebyshev polynomial of degree k of the first kind and $\kappa_2(X)$ is the spectral condition number of X. The main drawbacks of this estimate are that: (1) $\kappa_2(X)$ is not practically computable in general; and (2) that it may be extremely large. The rationale here is that this is an asymptotic result and the actual residual norm should behave like the right-hand side—apart from the multiplicative constant $\kappa_2(X)$. However, the process is finite, being optimum on a finite-dimensional space, and the above inequality though correct may become meaningless in practice.

The standard inequality (1.1) is derived by using the spectral decomposition of A and exploiting polynomials that are small on the spectrum. As is well known, there are situations where the spectrum does not give a good indication of the convergence behaviour. As an example, a triangular matrix with ones on the diagonal has a spectrum that is reduced to the value one, yet GMRES convergence for such a matrix can be very slow if some of its non-zero entries are large. Therefore, this type of analysis does not lead to an understanding of situations that involve highly non-normal matrices such as the one above and the spectrum is not a good indicator of convergence. On the other hand, experience shows that the spectrum does help understand convergence behaviour *in general*, or, to be more accurate, *for average case situations*, though these could be hard to quantify. For example, the main argument used in explaining the improved convergence of preconditioned iterations is that the eigenvalues of the preconditioned matrix tend to be clustered around one. In such situations, Krylov methods will converge faster *in general*. Another example is the success of deflation methods for solving linear systems [2,3], as well as eigenvalue problems [4,5]. For linear systems, the objective of these techniques is to remove the eigenvalues closest to zero from the spectrum of A, yielding faster convergence in later steps of the iteration.

A number of alternative theories have been proposed to analyse Min-Res methods. The spectra of A, $A + A^{\mathrm{T}}$, $A - A^{\mathrm{T}}$, $A^{\mathrm{T}}A$ and quantities related to operator trigonometry [6] can all be invoked to try to explain the convergence behaviour differently [7]. However, none of them is sufficient by itself and several discussions to this effect exist in the literature see, e.g., References [8–10]. As an alternative to the spectrum of A, the use of the ϵ-spectrum or pseudo-spectrum [11,12] has also been advocated. The pseudo-spectrum does not provide a quantative analysis of actual behaviour. Our goal in this paper is not to propose a new theory but rather to re-examine the standard ones. We start with a brief review of the common approaches that have been taken in the past and provide a few additional results.

2. Residual bounds of first type

Consider the non-singular linear system

$$Ax = b \tag{2.1}$$

Given a subspace S and an initial guess x_0 to the solution, Min-Res methods compute the (unique) approximate solution of the form $x = x_0 + s$ where $s \in S$ minimizes the 2-norm

of the residual vector

$$r(s) = b - Ax = r_0 - As$$

Here, the common notation $r_0 = b - Ax_0$ is used. This optimal approximation will be denoted by $\tilde{x}$ and the corresponding residual by $\tilde{r}$. Hence,

$$\|\tilde{r}\| = \min_{x \in x_0+S} \|b - Ax\| = \min_{s \in S} \|r_0 - As\| \equiv \|b - A\tilde{x}\| \tag{2.2}$$

The optimal As is the orthogonal projection of r_0 into AS and, in particular

$$\|\tilde{r}\| = \sin \angle(r_0, AS)\ \|r_0\| \tag{2.3}$$

The sine of the angle between the initial residual r_0 and the subspace AS gives the *reduction factor* in the residual norm achieved by the projection process.

2.1. An overview of standard residual bounds

In order to obtain a bound for $\|\tilde{r}\|$, it is sufficient to find an upper bound for the angle $\angle(r_0, AS)$, which represents the smallest possible angle between r_0 and arbitrary vectors in AS. It is found by maximizing the normalized inner product

$$\frac{|(r_0, As)|}{\|r_0\|\|As\|}$$

which represents the cosine of the angle between the vectors r_0 and As. In other words,

$$\cos \angle(r_0, AS) = \max_{s \in S} \frac{|(r_0, As)|}{\|r_0\|\|As\|} \tag{2.4}$$

In the remainder of the paper it is assumed that S contains the initial residual vector r_0. To prove convergence of the MR and the restarted 'generalized conjugate residual' (GCR) methods when A is positive definite, Eisenstat, Elman and Schultz used the lower bound for $\cos \angle(r_0, AS)$ provided by simply taking $s \equiv r_0$, yielding [1]

$$\cos \angle(r_0, AS) \geq \frac{|(r_0, Ar_0)|}{\|r_0\|\|Ar_0\|} \tag{2.5}$$

Gustafson [6] defines the cosine of the angle of a linear operator as

$$\cos(A) \equiv \min_{x \neq 0} \frac{|(x, Ax)|}{\|x\|\|Ax\|} \tag{2.6}$$

and gives several inequalities that exploit this definition. When $r_0 \in S$, then

$$\cos \angle(r_0, AS) \geq \cos(r_0, Ar_0) \geq \cos(A) \tag{2.7}$$

When A is indefinite (not positive definite or negative definite) then $\cos(A) = 0$ so the above inequality is not useful. If A is either positive definite or negative definite, then $\cos(A) > 0$ and the relations (2.7) and (2.3) yield the well-known result, see References [6] and [13, p. 135],

$$\|\tilde{r}\| \leq \sin(A)\ \|r_0\| \tag{2.8}$$

Consider again the case when A is positive definite and define

$$\mu(A) \equiv \min_{x \neq 0} \frac{(Ax, x)}{(x, x)} \tag{2.9}$$

This is the smallest eigenvalue of $(A + A^{\mathrm{T}})/2$. From the relation,

$$\frac{(x, Ax)}{\|x\|\|Ax\|} = \frac{(x, Ax)}{\|x\|^2} \times \frac{\|x\|}{\|Ax\|}$$

it is immediately seen that

$$\cos(A) \geq \frac{\mu(A)}{\|A\|}$$

This results in the following inequality, which has been established in Reference [1]:

$$\|\tilde{r}\| \leq \left[1 - \frac{\mu^2(A)}{\|A\|^2}\right]^{1/2} \|r_0\| \tag{2.10}$$

An alternative inequality can be obtained by using the same vector s and the following argument, see, e.g., References [1,13,14]. Write

$$\frac{(x, Ax)^2}{\|x\|^2\|Ax\|^2} = \frac{(Ax, x)}{\|x\|^2} \times \frac{(Ax, x)}{\|Ax\|^2}$$

and note that the first term in the right-hand side can again be bounded from below by $\mu(A)$. For the second term, set $z = Ax$ and write

$$\frac{(Ax, x)}{\|Ax\|^2} = \frac{(z, A^{-1}z)}{\|z\|^2} \geq \mu(A^{-1})$$

Since A is positive definite, then $\mu(A^{-1}) > 0$ and this gives the relation

$$\frac{|(x, Ax)|^2}{\|x\|^2\|Ax\|^2} \geq \mu(A)\ \mu(A^{-1}) \tag{2.11}$$

Therefore,

$$\cos(A) \geq \sqrt{\mu(A)\ \mu(A^{-1})} \tag{2.12}$$

The resulting residual bound similar to (2.10) is given by

$$\|\tilde{r}\| \leq \left[1 - \mu(A)\mu(A^{-1})\right]^{1/2} \|r_0\| \tag{2.13}$$

It is useful to compare the bounds (2.10) and (2.13) in the case when A is symmetric positive definite (SPD). Inequality (2.10) gives a particular case of a result proved in Reference [1]:

$$\|\tilde{r}\| \leq \left[1 - \frac{1}{\kappa_2^2(A)}\right]^{1/2} \|r_0\| \tag{2.14}$$

In the SPD case, $\mu(A)\mu(A^{-1}) = 1/\kappa_2(A)$ and (2.13) becomes

$$\|\tilde{r}\| \leq \left[1 - \frac{1}{\kappa_2(A)}\right]^{1/2} \|r_0\| \tag{2.15}$$

which is much sharper than (2.14) in general.

2.2. Alternative bounds

When A is not indefinite, an improvement to the above well-known inequalities can be obtained by exploiting an inequality similar to a result due to Kantorovitch [13]. The following Lemma is needed.

Lemma 2.1. *Let A be a real square matrix and assume that there exist two non-zero real scalars α and β of the same sign such that*

$$((A - \alpha I)x, (A - \beta I)x) \leq 0, \quad \forall\, x \in \mathbb{R}^n \tag{2.16}$$

Then, A is either positive definite or negative definite, and

$$\frac{(Ax, x)^2}{\|Ax\|^2\, \|x\|^2} \geq \frac{4\alpha\beta}{(\alpha + \beta)^2}, \quad \forall x \neq 0 \tag{2.17}$$

Proof
Consider an arbitrary *unit* vector x and expand (2.16) into

$$\|Ax\|^2 + \alpha\beta \leq (\alpha + \beta)(Ax, x) \tag{2.18}$$

When α and β are both positive then the above inequality shows that A is positive definite. When they are both negative then it shows that A is negative definite. This establishes the first part of the lemma.

Define $\lambda = (Ax, x)$ when A is positive definite and $\lambda = -(Ax, x)$ when A is negative definite. Then use (2.18) to show that

$$\frac{(Ax, x)^2}{\|Ax\|^2} = \frac{|\lambda|^2}{\|Ax\|^2} \geq \frac{|\lambda|^2}{|\alpha + \beta|\, |\lambda| - \alpha\beta}$$

The right-hand side is a function of $|\lambda|$ that takes its minimum for $|\lambda_{\text{opt}}| = 2\alpha\beta/|\alpha + \beta|$. Evaluating the right-hand side of the above equation for this value yields the desired inequality (2.17). ■

A converse result can also be shown: if A is either positive or negative definite, then α and β can be found such that (2.16) is satisfied.

When A is symmetric, then a simple choice for the two parameters α, β is $\alpha = \lambda_1$, $\beta = \lambda_n$, where it is assumed that eigenvalues are labelled from the smallest λ_1 to the largest λ_n. Indeed, the relation $((A - \lambda_1 I)x, (A - \lambda_n I)x) \leq 0$ follows immediately from the fact that the eigenvalues of $(A - \lambda_1 I)(A - \lambda_n I)$ are all non-positive, so $((A - \lambda_1 I)x, (A - \lambda_n I)x) = ((A - \lambda_1 I)(A - \lambda_n I)x, x) \leq 0$ for any x. This results in the following well-known

inequality—see, e.g., Reference [6], which is valid for any SPD matrix A:

$$\frac{(Ax, x)^2}{\|x\|^2 \|Ax\|^2} \geq \frac{4\lambda_n \lambda_1}{(\lambda_n + \lambda_1)^2} \tag{2.19}$$

The following proposition, which follows immediately from the above discussion, summarizes the situation.

Proposition 2.1. *Let A be a matrix that satisfies the assumptions of Lemma 2.1 and S a subspace containing the initial residual vector r_0. Then the residual $\tilde{r}$ obtained from a minimal residual projection method onto S is such that*

$$\|\tilde{r}\| \leq \frac{|\beta - \alpha|}{|\beta + \alpha|} \|r_0\| \tag{2.20}$$

In particular, when A is SPD, then the assumptions of Lemma 2.1 are satisfied with $\beta = \lambda_1, \alpha = \lambda_n$, and we have

$$\|\tilde{r}\| \leq \frac{\kappa_2(A) - 1}{\kappa_2(A) + 1} \|r_0\| \tag{2.21}$$

Proof
Inequality (2.20) follows immediately by using inequalities (2.8), (2.6), and (2.17). Inequality (2.21) is a consequence for the SPD case. ■

In the SPD case, a simpler proof of (2.21) that does not exploit the lemma is based on minimizing $\|I - \alpha A\|$ over α.

Inequality (2.21) resembles a similar result obtained for the steepest descent algorithm and the proofs of these results are very similar. In the SPD case, it can easily be seen that (2.21) is sharper than (2.15). Indeed, this follows from the inequality

$$[1 - t]^{1/2} \geq \frac{1 - t}{1 + t}$$

which is valid for $0 \leq t \leq 1$, when we set $t \equiv \lambda_1/\lambda_n$.

We now provide two examples that show how the previous results can be exploited. It is assumed in both cases that the matrix is positive definite.

Example 1. The condition (2.16) can be rewritten as

$$(A^{\mathrm{T}} Ax, x) \leq (\alpha + \beta) \left(\left(A - \frac{\alpha\beta}{\alpha + \beta} I \right) x, x \right) \quad \forall x \tag{2.22}$$

Assume that A is positive definite and select the shift $\delta \equiv \alpha\beta/(\alpha + \beta)$ such that $A - \delta I$ is positive semi-definite, for example:

$$\frac{\alpha\beta}{\alpha + \beta} = \delta = \frac{1}{2}\mu(A)$$

where μ defined in (2.9) is the smallest eigenvalue of the symmetric part of A. Then (2.22) can be rewritten as

$$\frac{(A^{\mathrm{T}} Ax, x)}{((A - \delta I)\, x, x)} \leq \alpha + \beta, \quad \forall x$$

and is satisfied when $\alpha + \beta$ is the largest eigenvalue of the generalized eigenvalue problem

$$A^{\mathrm{T}} A\, x = \lambda \left(\frac{A + A^{\mathrm{T}}}{2} - \delta I \right) x \tag{2.23}$$

The largest eigenvalue of this problem is positive since the two matrices of the pair are both positive definite. Let $\sigma(\delta)$ be this eigenvalue. Then the condition is that $\alpha + \beta = \sigma(\delta)$. The two conditions

$$\frac{\alpha\beta}{\alpha + \beta} = \delta\,, \quad \alpha + \beta = \sigma(\delta)$$

yield the solution

$$\alpha = \tfrac{1}{2}\left[\sigma(\delta) + \sqrt{\sigma(\delta)^2 - 4\delta\sigma(\delta)}\right], \quad \beta = \tfrac{1}{2}\left[\sigma(\delta) - \sqrt{\sigma(\delta)^2 - 4\delta\sigma(\delta)}\right]$$

For these values the result (2.20) becomes

$$\|\tilde{r}\| \;\leq\; \sqrt{1 - 4\frac{\delta}{\sigma(\delta)}}\, \|r_0\| \tag{2.24}$$

Note that the above result depends on a parameter δ. Later we will provide an inequality in which the best δ is selected.

Example 2. An alternative to the previous approach consists of rewriting inequality (2.16) as

$$\Big((A^{\mathrm{T}} A + \alpha\beta\, I)x, x\Big) \leq (\alpha + \beta)(Ax, x)$$

Similarly to the previous case we impose the condition $\alpha\beta = \delta$ with $\delta > 0$. Then the above inequality is satisfied when

$$\alpha + \beta = \sigma(\delta)$$

where $\sigma(\delta)$ is the largest eigenvalue of the generalized problem

$$(A^{\mathrm{T}} A + \delta I)\, x = \lambda \frac{A + A^{\mathrm{T}}}{2}\, x, \quad \forall x \tag{2.25}$$

The largest eigenvalue of this problem is again positive. The two conditions

$$\alpha\beta = \delta\,, \quad \alpha + \beta = \sigma(\delta)$$

yield the solution,

$$\alpha = \tfrac{1}{2}\left[\sigma(\delta) + \sqrt{\sigma(\delta)^2 - 4\delta}\right], \quad \beta = \tfrac{1}{2}\left[\sigma(\delta) - \sqrt{\sigma(\delta)^2 - 4\delta}\right]$$

And now the result (2.20) becomes

$$\|\tilde{r}\| \;\leq\; \sqrt{1 - 4\frac{\delta}{\sigma(\delta)^2}}\, \|r_0\| \tag{2.26}$$

It is interesting to observe that an eigenvalue $\lambda(\delta)$ of Equation (2.25) is also an eigenvalue

of (2.23) for a different δ, namely for $\delta' = \lambda(\delta) \times \delta$. In addition, the corresponding ratios $\delta/\sigma(\delta)$ and $\delta/\sigma(\delta)^2$ involved in the bounds (2.24) and (2.26), respectively, are identical. Therefore, the best bounds achieved in both cases are also the same, so we need only consider one of the approaches, e.g. the second.

Theorem 2.1. *Let A be a positive definite matrix and for any $\omega > 0$, let $\gamma(\omega)$ be the largest generalized eigenvalue of the pair*

$$\left(\frac{A^{\mathrm{T}}A}{\omega} + \omega I,\ \frac{A + A^{\mathrm{T}}}{2}\right) \tag{2.27}$$

Define $\gamma_{\min}$ to be the minimum of $\gamma(\omega)$ over $\omega > 0$. Then,

$$\|\tilde{r}\| \ \leq\ \sqrt{1 - \frac{4}{\gamma_{\min}^2}}\ \|r_0\| \tag{2.28}$$

Proof
The proof is essentially based on a change of notation. Dividing both sides of Equation (2.25) by $\omega \equiv \sqrt{\delta}$ yields the eigenvalue problem

$$\left(\frac{A^{\mathrm{T}}A}{\omega} + \omega I\right)\ x = \gamma\ \frac{A + A^{\mathrm{T}}}{2}\ x$$

with $\gamma \equiv \lambda/\omega$. With this new notation Equation (2.26) becomes

$$\|\tilde{r}\| \ \leq\ \sqrt{1 - \frac{4}{\gamma(\omega)^2}}\ \|r_0\|$$

The best convergence factor is provided when $\gamma(\omega)$ is minimized. ■

The form of the first matrix in the pair (2.27) suggests that $\gamma(\omega)$ is a function that will decrease from infinity as $\omega = 0$ then reaches a minimum associated with $\gamma_{\min}$ and then increases again to infinity. This has been confirmed experimentally.

We illustrate the above results with two 15×15 matrices defined as

$$A = \begin{pmatrix} 1 & \eta & & & \\ & 1 & \eta & & \\ & & \ddots & \ddots & \\ & & & 1 & \eta \\ & & & & 1 \end{pmatrix}$$

with $\eta = 1$ for the first test and $\eta = 0.9$ for the second. Figure 1 shows the results for the first matrix and Figure 2 shows the results for the second matrix. In the figures, the y-coordinates show the various estimates for the reduction factors $\|\tilde{r}\|/\|r_0\|$. EES1 and EES2 refer to Inequalities (2.10) and (2.13) respectively. The curves labelled EX1 and EX2 refer to the factors obtained from (2.24) and (2.26) respectively, as δ varies. It can be seen that the minima of the two curves are indeed the same.

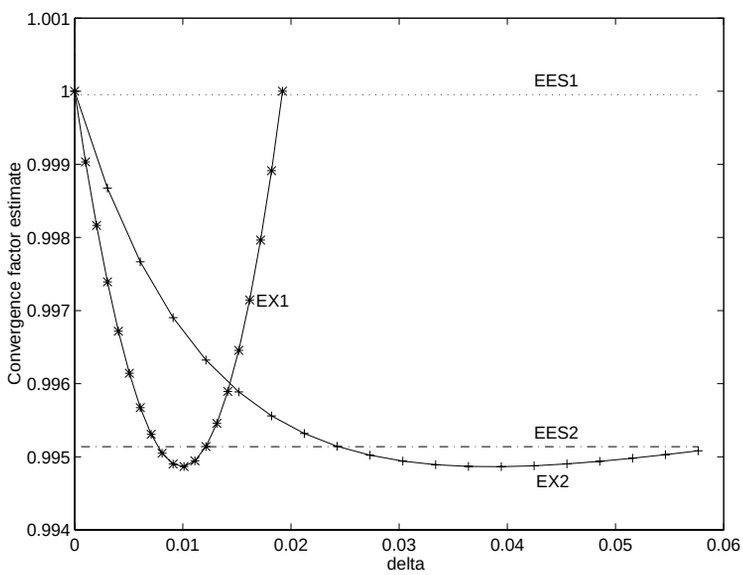

Figure 1. Upper bounds for residual norm reductions obtained from inequalities (2.10) (labelled EES1), (2.13) (labelled EES2), (2.24) (labelled EX1) and (2.26) (labelled EX2), for the first test matrix.

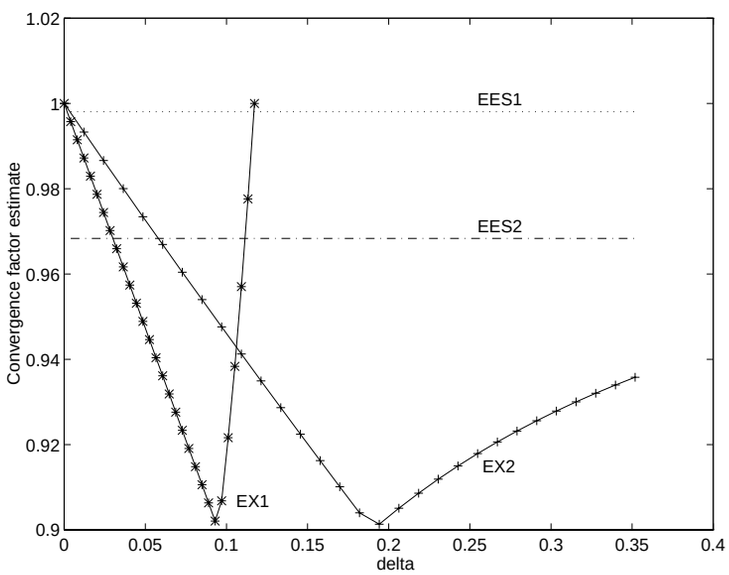

Figure 2. Upper bounds for residual norm reductions obtained from inequalities (2.10) (labelled EES1), (2.13) (labelled EES2), (2.24) (labelled EX1) and (2.26) (labelled EX2), for the second test matrix.

2.3. *Restarted Min-Res algorithms*

A 'restarted' Min-Res iterative process is an algorithm that selects a subspace S at each iteration and minimizes the residual norm in the subspace. Each new iteration starts with the initial guess x_0 set equal to the approximate solution obtained from the previous iteration. It is always assumed that each subspace S contains the initial residual r_0. We begin by observing that by (2.5) and (2.3) the condition

$$\frac{|(Ax, x)|}{(x, x)} \geq c > 0, \quad \forall x$$

guarantees convergence of any restarted iteration since the successive subspaces S are assumed to contain the initial residual r_0. In particular, a well-known convergence result [1] is that *if A is positive definite then a restarted Min-Res iterative process in which each subspace S contains r_0 converges for any intial guess x_0.* The set of all Rayleigh quotients $(Ax, x)/(x, x)$ constitutes the field of values of A. The above condition therefore states that if the field of values (a compact set) excludes the origin, then the method converges. We now show that a weakened form of the converse is also true. Consider the MR method which, at each restart, takes $S \equiv \{r\}$ where r now denotes the current residual at restart. Assume that for this case each step of the projection process reduces the initial residual by a constant $0 \leq \tau < 1$, and this for an arbitrary initial residual. This means that

$$\sin \angle(r, Ar) \leq \tau, \quad \forall\, r$$

or

$$\cos \angle(r, Ar) \geq \sqrt{1 - \tau^2}, \quad \forall\, r$$

which implies that

$$\frac{|(Ar, r)|}{\|Ar\|\|r\|} = \frac{|(Ar, r)|}{\|r\|^2} \times \frac{\|r\|}{\|Ar\|} \geq \sqrt{1 - \tau^2}$$

A result is that the Rayleigh quotients must be bounded from below because

$$\frac{|(Ar, r)|}{\|r\|^2} \geq \frac{\|Ar\|}{\|r\|}\sqrt{1 - \tau^2} \geq \sigma_{\min}\sqrt{1 - \tau^2}$$

where $\sigma_{\min}$ is the smallest singular value of A. This establishes the following result.

Theorem 2.2. *Assume that the field of values of A excludes the origin. Then, each step of a restarted Min-Res projection procedure (in which the subspaces S contain the initial residuals), reduces the residual norm by a factor $\leq \tau < 1$. Conversely, if each step of the MR (i.e. GMRES(1)) algorithm reduces the residual norm by a factor $\leq \tau < 1$, for any initial residual, then the field of values of A excludes the origin.*

Sufficient conditions to guarantee that the field of values excludes the origin are difficult to obtain. One such condition is given by Lemma 2.1.

The main insight provided by Therorem 2.2 is that any convergence result or resisudal bounds for restarted Min-Res algorithms, which prove that the residual is reduced by a certain amount at each step, must assume that A is not indefinite. No convergence result can

be established for an indefinite matrix—unless it proves only sub-geometric convergence. It is not known whether such convergence results can be established that bypass the positive (or negative) definiteness assumption.

3. Residual bounds of the second type

We now return to the general result of (2.3) and consider the case when S is the Krylov subspace K_m of dimension m,

$$K_m = \text{span}\{r_0, Ar_0, \ldots, A^{m-1}r_0\}$$

Thus, a generic vector of K_m is of the form

$$q(A)r_0$$

where q is a polynomial of degree $m-1$. For any such polynomial, we have

$$\cos \angle(r_0, AK_m) \geq \frac{|(r_0, Aq(A)r_0)|}{\|r_0\| \|Aq(A)r_0\|} \tag{3.1}$$

In the ideal situation when $q(A)r_0 = A^{-1}r_0$ then $Aq(A)r_0 = r_0$, which gives a zero angle and a zero residual. However, obtaining residual bounds by using this relationship is rather complex. A few instances are considered.

3.1. Bounds derived from Chebyshev polynomials

The usual error bounds used to analyse the GMRES algorithm in the asymmetric case depend on the condition number of the matrix of eigenvectors. In establishing these results A is assumed to be diagonalizable ($A = XDX^{-1}$)—see however a generalization in References [19,18]. By the optimality of the approximate solution it can be said that $x_m = x_0 + s_m(A)r_0$, where s minimizes the residual norm $\|b - Ax\| = \|b - A(x_0 + s(A)r_0)\|$ over all polynomials s of degree $\leq m-1$. The residual vector for each polynomial s is

$$r = r_0 - As(A)r_0 = [I - As(A)]r_0 \equiv p(A)r_0$$

and by the optimality property, the polynomial p minimizes $\|p(A)r_0\|$ over all polynomials of degree $\leq m$ which are 'consistent' polynomials, i.e. such that $p(0) = 1$. Then, a certain consistent polynomial p is selected to be small on the spectrum of A, and the following argument is used:

$$\begin{aligned} \|r_m\| &\leq \|p(A)r_0\| = \|p(XDX^{-1})r_0\| = \|Xp(D)X^{-1}r_0\| \\ &\leq \kappa_2(X)\, \|p(D)\|\, \|r_0\| \end{aligned}$$

Except when A is normal, the condition number of X can be very large and the above bound can become poor. The alternative discussed next uses information that is computable, and it avoids the condition number of X, leading to a tighter inequality. It is based on a comparison result with a matrix–vector power sequence of the form $B^k v$ where B is a matrix whose spectral radius is known and small. The growth of $\|B^k v\|$ as k tends to infinity is (generally)

of the form $(\rho(B) + \epsilon_k)^k$, where $\rho(B)$ is the spectral radius of B and ϵ_k is a sequence that converges to zero as $k \rightarrow \infty$.

Assume that we can find two scalars α and $\beta > 1$ such that the eigenvalues of the shifted and scaled matrix

$$\hat{A} = \beta I - \alpha A \tag{3.2}$$

are contained in the ellipse centred at the origin and with focal distance one. Then a good polynomial to use is

$$t_m(A) = T_m(\beta I - \alpha A)/T_m(\beta) \equiv T_m(\hat{A})/T_m(\beta)|$$

and the usual bound derived from using this polynomial is

$$\|r_m\| \leq \|t_m(A)r_0\|$$

Note that the polynomials t_m satisfy the consistency condition $t_m(0) = 1$. A consequence of the three-term recurrence relation of Chebyshev polynomials

$$T_{m+1}(\hat{A}) = 2T_m(\hat{A}) - T_{m-1}(\hat{A}), m \geq 1, \quad T_1(\hat{A}) = \hat{A},\ T_0(\hat{A}) = I$$

is that

$$\begin{pmatrix} T_{m+1}(\hat{A})r_0 \\ T_m(\hat{A})r_0 \end{pmatrix} = \begin{pmatrix} 2\hat{A} & -I \\ I & 0 \end{pmatrix} \begin{pmatrix} T_m(\hat{A})r_0 \\ T_{m-1}(\hat{A})r_0 \end{pmatrix}$$

Denote $T_m(\beta)$ by σ_m and define

$$\mathcal{B} = \begin{pmatrix} 2\hat{A} & -I \\ I & 0 \end{pmatrix} \tag{3.3}$$

Then, we have the relations,

$$\begin{aligned} \begin{pmatrix} \sigma_{m+1}t_{m+1}(A)r_0 \\ \sigma_m t_m(A)r_0 \end{pmatrix} &= \begin{pmatrix} T_{m+1}(\hat{A})r_0 \\ T_m(\hat{A})r_0 \end{pmatrix} = \mathcal{B} \begin{pmatrix} T_m(\hat{A})r_0 \\ T_{m-1}(\hat{A})r_0 \end{pmatrix} = \cdots \\ &= \mathcal{B}^{m+1} \begin{pmatrix} r_0 \\ \hat{A}r_0 \end{pmatrix} \end{aligned}$$

As a result, letting

$$w_0 \equiv \begin{pmatrix} r_0 \\ \hat{A}r_0 \end{pmatrix}$$

we obtain

$$\sqrt{\sigma_{m+1}^2 \|r_{m+1}\|^2 + \sigma_m^2 \|r_m\|^2} \leq \|\mathcal{B}^{m+1} w_0\| \tag{3.4}$$

We can now state the following result.

Theorem 3.1. *Assume that an ellipse $E(c, d, a)$ with centre c, focal distance d and major semi-axis a can be found that includes the spectrum of A with at least one eigenvalue on the boundary. Let $\hat{A} = \beta I - \alpha A$ with $\alpha = 1/d$ and $\beta = c/d$, and $\mathcal{B}$ defined by (3.3). Then the residual vector r_{m+1} obtained from a Minimal Residual method using a Krylov subspace*

of dimension $m+1$ satisfies the inequality

$$\|r_{m+1}\| \leq \frac{\|\mathcal{B}^{m+1} w_0\|}{\sqrt{T_{m+1}^2(c/d) + T_m^2(c/d)}} \tag{3.5}$$

where w_0 is the $2n$ vector consisting of the subvectors r_0 and $\hat{A}r_0$. The spectral radius of the matrix $\mathcal{B}$ is given by

$$\rho(\mathcal{B}) = \frac{a}{d} + \sqrt{\left(\frac{a}{d}\right)^2 - 1} \tag{3.6}$$

and in particular when $a = d$ (real spectrum) then $\rho(\mathcal{B}) = 1$.

Proof
Inequality (3.5) follows immediately from (3.4) and the inequality $\|r_{m+1}\| \leq \|r_m\|$. The scalars α, β as given above transform the ellipse $E(c, d, a)$ into an ellipse centred at the origin and with focal distance unity. The major semi-axis of the transformed ellipse is then given by

$$\hat{a} = \frac{a}{d}$$

The eigenvalues of $\mathcal{B}$ are

$$\theta_{i,\pm} = \frac{1}{2}\left(\hat{\lambda}_i \pm \sqrt{\hat{\lambda}_i^2 - 1}\right)$$

where each $\hat{\lambda}_i = \beta - \alpha\lambda_i$ is an eigenvalue of $\hat{A}$. It is well-known [13] that the $\hat{\lambda}$'s are transformed from the θ's by means of the Joukowsky transform

$$\hat{\lambda} = \tfrac{1}{2}(\theta + \theta^{-1})$$

which maps a circle of radius r centred at the origin into an ellipse centred at the origin, with focal distance 1, and major semi-axis $\hat{a} = (r + r^{-1})/2$. Each point on the ellipse is mapped from a point of the circle, so an eigenvalue on the ellipse is transformed into an eigenvalue of equal modulus r. All eigenvalues inside the ellipse will be transformed into points inside the circle. Therefore the spectral radius of $\mathcal{B}$ is the radius of the circle associated with the eigenvalue (s) on the ellipse. This radius can be obtained from transforming the major semi-axis a with the inverse function $z + \sqrt{z^2 - 1}$ and this gives

$$\rho(\mathcal{B}) = \hat{a} + \sqrt{\hat{a}^2 - 1} = \frac{a}{d} + \sqrt{\left(\frac{a}{d}\right)^2 - 1}$$

This completes the proof. ■

The scalars d and c are typically selected in a certain optimal way, to make the ratio c/d as large as possible. A weakened version of this inequality is obtained by exploiting the inequality $\|\mathcal{B}^m w_0\| \leq \|\mathcal{B}^m\| \|w_0\|$ and the well-known result [15,13,16]:

$$\lim_{k \to \infty} \|B^k\|^{1/k} = \rho(B)$$

from which it follows immediately that

$$\|B^k\| = (\rho(B) + \epsilon_k)^k$$

where $\epsilon_k = \|B^k\|^{1/k} - \rho(B)$, converges to zero as k converges to infinity. This gives,

$$\|r_{m+1}\| \leq \frac{\|\mathcal{B}^{m+1}\| \, \|w_0\|}{\sqrt{T_{m+1}^2(\beta) + T_m^2(\beta)}} = \frac{\left[\rho(\mathcal{B}) + \epsilon_{m+1}\right]^{m+1}}{\sqrt{T_{m+1}^2(\beta) + T_m^2(\beta)}} \, \|w_0\| \tag{3.7}$$

However, there are disadvantages in using the upper bound given above since the norm of $\mathcal{B}^k w_0$ can be very poorly estimated by $\|\mathcal{B}^k\|\|w\|$ when $\mathcal{B}$ is highly non-normal.

The main difference between the inequality of the above theorem and the classical ones, such as Inequality (1.1), is that it does not involve the condition number of the matrix of eigenvectors. In contrast, it provides only a comparison result with a sequence that captures the effects of nonnormality without trying to model them. As is shown by the experiments, attempts to model these effects, e.g. as in Inequalities (1.1) or (3.7) may lead to bounds that are too loose to be of any interest. The rationale of the above theorem is that the usual tools provided by norms and spectral analysis are insufficient for analysing the behaviour of certain iterative processes. The condition number of X in Inequality (1.1) is not only unavailable in practice (for large matrices) but it may also lead to results that are too pessimistic. The bound (3.5) involves a sequence that is also unavailable *a priori*, but that is more easily computable than the constant $\kappa_2(X)$. All that is required is to get some understanding of the predicted behaviour of this sequence in particular situations. Though the spectral norm of $\mathcal{B}$ given by (3.6) gives an idea of the asymptotic behaviour of this sequence, the transient behaviour is not easily modelled in the case of highly non-normal matrices and it is better not to attempt to capture it with an estimate based on norms.

In the case when all eigenvalues of A are real then $a = d$ and the term in brackets in the numerator of Inequality (3.7) becomes $1 + \epsilon_{m+1}$. In addition, $\epsilon_i = 0$ when $\mathcal{B}$ is normal. As it turns out, $\mathcal{B}$ is normal if and only if the matrix $\hat{A}$ is skew-Hermitian:

$$\begin{aligned}
\mathcal{B}^{\mathrm{H}}\mathcal{B} - \mathcal{B}\mathcal{B}^{\mathrm{H}} &= \begin{pmatrix} 2\hat{A}^{\mathrm{H}} & I \\ -I & 0 \end{pmatrix}\begin{pmatrix} 2\hat{A} & -I \\ I & 0 \end{pmatrix} - \begin{pmatrix} 2\hat{A} & -I \\ I & 0 \end{pmatrix}\begin{pmatrix} 2\hat{A}^{\mathrm{H}} & I \\ -I & 0 \end{pmatrix} \\
&= \begin{pmatrix} 4\hat{A}^{\mathrm{H}}\hat{A} + I & -2\hat{A}^{\mathrm{H}} \\ -2\hat{A} & I \end{pmatrix} - \begin{pmatrix} 4\hat{A}\hat{A}^{\mathrm{H}} + I & 2\hat{A} \\ 2\hat{A}^{\mathrm{H}} & I \end{pmatrix} \\
&= 4\begin{pmatrix} \hat{A}^{\mathrm{H}}\hat{A} - \hat{A}\hat{A}^{\mathrm{H}} & -\frac{1}{2}(\hat{A} + \hat{A}^{\mathrm{H}}) \\ -\frac{1}{2}(\hat{A} + \hat{A}^{\mathrm{H}}) & 0 \end{pmatrix}
\end{aligned}$$

Unfortunately, when $\hat{A}$ (or A) is normal but not skew-Hermitian, $\mathcal{B}$ is not normal in general. Because the non-Hermitian matrix $\mathcal{B}$ is used to derive the bounds (3.5) and (3.7), a natural question is whether or not the resulting estimates will be weaker than those provided by the classical inequality (1.1). The answer is that the two inequalities are very close to one another in this case, as one should expect. Recall that

$$\mathcal{B}^{m+1} w_0 = \begin{pmatrix} T_{m+1}(\hat{A}) r_0 \\ T_m(\hat{A}) r_0 \end{pmatrix}$$

We have

$$\|T_m(\hat{A}) r_0\| \leq \|T_m(\hat{A})\| \, \|r_0\|$$

and when $\hat{A}$ is normal then

$$\|T_m(\hat{A})\| = \max_{i=1,\ldots,n} |T_m(\hat{\lambda}_i)| = \max_{i=1,\ldots,n} \frac{\theta_i^m + \theta_i^{-m}}{2} = T_m(a/d)$$

Hence

$$\|\mathcal{B}^{m+1} w_0\| \leq \sqrt{T_{m+1}^2(a/d) + T_m^2(a/d)}\|r_0\|$$

leading to the inequality

$$\|r_{m+1}\| \leq \frac{\sqrt{T_{m+1}^2(a/d) + T_m^2(a/d)}}{\sqrt{T_{m+1}^2(c/d) + T_m^2(c/d)}}\|r_0\|$$

which, for large m, is indeed very close to Inequality (1.1) with $\kappa_2(X) = 1$.

3.2. Bounds from the Arnoldi matrix $\bar{H}_m$

It is possible also to analyse the behaviour of Min-Res methods from some *a posteriori* information extracted from the process, such as the eigenvalue or singular value estimates obtained from the projection matrix. We distinguish again between basic bounds and bounds that attempt to mimic the optimality of the residual polynomials.

3.2.1. Basic results. The basic relation that arises from the Arnoldi algorithm is the following:

$$AV_m = V_{m+1}\bar{H}_m \tag{3.8}$$

Here, the column-vectors of V_m are the Arnoldi vectors obtained from a Gram–Schmidt orthogonalization starting with $v_1 = r_0/\beta$ in which each new vector is the product of the current basis vector v_j by A. Therefore,

$$\begin{aligned}
\cos\angle(r_0, AK_m) &= \max_y \frac{|(AV_m y, v_1)|}{\|AV_m y\|} \\
&= \max_y \frac{|(V_{m+1}\bar{H}_m y, v_1)|}{\|V_{m+1}\bar{H}_m y\|} \\
&= \max_y \frac{|(\bar{H}_m y, V_{m+1}^{\mathrm{H}} v_1)|}{\|V_{m+1}\bar{H}_m y\|} \\
&= \max_y \frac{|(\bar{H}_m y, e_1)|}{\|\bar{H}_m y\|}
\end{aligned} \tag{3.9}$$

Thus, the cosine of the angle between r_0 and AK_m is equal to the maximum cosine of angles spanned between an arbitrary linear combination of columns of $\bar{H}_m$ and the vector e_1. By selecting various test vectors y we can get bounds on the cosine. For example, taking $y = e_1$ gives the simple bound,

$$\cos\angle(r_0, AK_m) \geq \frac{|h_{11}|}{\sqrt{h_{21}^2 + h_{11}^2}}$$

We have already seen this lower bound in a different form. Indeed,

$$\frac{|h_{11}|}{\sqrt{h_{21}^2 + h_{11}^2}} = \frac{|(Av_1, v_1)|}{\|Av_1\|}$$

and, since $v_1 = r_0/\|r_0\|$, this gives

$$\cos \angle(r_0, AK_m) \geq \frac{|(r_0, Ar_0)|}{\|r_0\|\|Ar_0\|}$$

which is identical with (2.5).

This can be extended by taking, similarly, $y = e_j$ to obtain

$$\cos \angle(r_0, AK_m) \geq \frac{|h_{1j}|}{\sqrt{h_{1j}^2 + \cdots + h_{j+1,j}^2}}$$

Note that the right-hand side is also equal to $|(Av_j, v_1)|/\|Av_j\|$, which represents the cosine of the angle between Av_j and r_0. A consequence of the above inequality is that

$$\cos \angle(r_0, AK_m) \geq \max_{j=1,\ldots,m} \frac{|h_{1j}|}{\sqrt{h_{1j}^2 + \cdots + h_{j+1,j}^2}} \tag{3.10}$$

which is a readily computable quantity.

Another natural way to select a test vector y is to rewrite (3.9) as

$$\cos \angle(r_0, AK_m) = \max \frac{|(y, \bar{H}_m^{\mathrm{H}} e_1)|}{\|\bar{H}_m y\|}$$

and attempt to make the numerator as large as possible modulo a scaling of y. This gives the choice $y = \bar{H}_m^{\mathrm{H}} e_1$ which results in

$$\cos \angle(r_0, AK_m) \geq \frac{|(\bar{H}_m^{\mathrm{H}} e_1, \bar{H}_m^{\mathrm{H}} e_1)|}{\|\bar{H}_m \bar{H}_m^{\mathrm{H}} e_1\|} = \frac{\|\bar{H}_m^{\mathrm{H}} e_1\|^2}{\|\bar{H}_m \bar{H}_m^{\mathrm{H}} e_1\|} \tag{3.11}$$

If we denote the elements of the matrix $\bar{H}_m \bar{H}_m^{\mathrm{H}}$ by s_{ij}, then the right-hand side is equal to

$$\frac{s_{11}}{\sqrt{s_{11}^2 + \cdots + s_{m+1,1}^2}}$$

The above inequality can be extended using other columns $y = \bar{H}_m^{\mathrm{H}} e_j$ and then sharpened by taking the maximum over j to yield a bound similar to (3.10) but based on the s_{ij}'s.

Let now $\sigma_{\min}$ and $\sigma_{\max}$ be the smallest and largest singular values of $\bar{H}_m$. Then, relation (3.11) yields

$$\cos \angle(r_0, AK_m) \geq \frac{\sigma_{\min}^2(\bar{H}_m)}{\sigma_{\max}^2(\bar{H}_m)} = \frac{1}{\kappa_2^2(\bar{H}_m)} \tag{3.12}$$

A rather interesting consequence of this is the following corollary.

Proposition 3.1. *Assume that GMRES(m), the restarted GMRES algorithm using Krylov subspaces of dimension m, is used to solve a non-singular system and that at each restart the condition number $\kappa_2(\bar{H}_m)$ is bounded from above by a constant τ. Then the algorithm will converge.*

Note that the proposition utilizes information on the projected problem that is not available beforehand. However, the insight it provides is that convergence goes hand in hand with a good conditioning of the matrices $\bar{H}_m$.

We now show how to adapt the results of Section 2.1 to Krylov subspaces. The result of Proposition 2.1 would still be valid if we replace the condition (2.16) on α, β of Lemma 2.1 by a similar condition that must be satisfied for all x in K_m instead of all x in $\mathbb{R}^n$. The reason is that r_0 belongs to K_m. Then, writing an arbitrary vector x in K_m in the form $V_m y$, where the columns of V_m form the Arnoldi orthonormal basis of K_m, we get the condition that

$$((A - \alpha I)V_m y, (A - \beta I)V_m y) \leq 0, \quad \forall\, y \in \mathbb{R}^m$$

Denote by $\bar{I}_m$ the $(m+1) \times m$ identity matrix whose entries are equal to δ_{ij} and notice that by (3.8) we have

$$\begin{aligned}((A - \alpha I)V_m y, (A - \beta I)V_m y) &\\ &= (V_{m+1}\bar{H}_m y - \alpha V_m y, V_{m+1}\bar{H}_m y - \beta V_m y)\\ &= (V_{m+1}(\bar{H}_m y - \alpha \bar{I}_m y), V_{m+1}(\bar{H}_m y - \beta \bar{I}_m y))\\ &= (\bar{H}_m y - \alpha \bar{I}_m y, \bar{H}_m y - \beta \bar{I}_m y)\end{aligned}$$

Then, the requirement that (2.16) be valid in K_m translates into the condition

$$(\bar{H}_m y - \alpha \bar{I}_m y, \bar{H}_m y - \beta \bar{I}_m y) \leq 0, \quad \forall\, y \in \mathbb{R}^m \tag{3.13}$$

Denoting by H_m the $m \times m$ matrix obtained from $\bar{H}_m$ by deleting its last row, an expansion of (3.13) yields,

$$(\bar{H}_m^{\mathrm{T}}\bar{H}_m y, y) - (\alpha + \beta)\left(\frac{H_m + H_m^{\mathrm{T}}}{2} y, y\right) + \alpha\beta(y, y) \leq 0, \quad \forall\, y$$

This means that a result similar to that of Theorem 2.1 can be shown. The requirement is now that H_m be positive definite, instead of A. We state this result without proof.

Corollary 3.1. *Let $\bar{H}_m$ and H_m be the $(m+1)\times m$ and $m \times m$ Hessenberg matrices obtained from m Arnoldi steps applied to a matrix A and assume that H_m is positive definite. For any $\omega > 0$, let $\tilde{\gamma}(\omega)$ be the largest generalized eigenvalue of the pair*

$$\left(\frac{\bar{H}_m^{\mathrm{T}}\bar{H}_m}{\omega} + \omega I, \frac{H_m + H_m^{\mathrm{T}}}{2}\right) \tag{3.14}$$

Define $\tilde{\gamma}_{min}$ to be the minimum of $\tilde{\gamma}(\omega)$ over $\omega > 0$. Then

$$\|\tilde{r}\| \leq \sqrt{1 - \frac{4}{\tilde{\gamma}_{\min}^2}}\, \|r_0\| \tag{3.15}$$

3.2.2. Chebyshev polynomials of the Hessenberg matrix. The relation (3.8) can be rewritten as

$$AV_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^{\mathrm{T}}$$

where H_m is the leading $m \times m$ upper block of $\bar{H}_m$. In Arnoldi's method some of the eigenvalues of A are approximated by eigenvalues of H_m. The question addressed in this section is whether or not these approximations can be used instead of the exact ones in an inequality such as (3.7). It is clear that we are free to select anything we want for α and β when defining $\hat{A}$. In particular we can use those scalars that are associated with an ellipse that encloses the spectrum of H_m. We will then get an inequality similar to (3.7), except that all quantities defined will be related to the parameters of the optimal ellipse for H_m instead of A. Specifically, we can state the following result, which is a straightforward consequence of Theorem 3.1.

Corollary 3.2. *Assume that an ellipse $E(c_m, d_m, a_m)$ with centre c_m, focal distance d_m and major semi-axis a_m can be found that includes the spectrum of H_m. Let $\hat{A}_m = \beta_m I - \alpha_m A$ with $\alpha_m = 1/d_m$ and $\beta_m = c_m/d_m$, and $\mathcal{B}_m$ defined by (3.3), in which $\hat{A}$ is replaced by $\hat{A}_m$. Then the residual vector r_m obtained from a Minimal Residual method using a Krylov subspace of dimension m satisfies the inequality*

$$\|r_m\| \leq \frac{\|\mathcal{B}_m^m w_{0,m}\|}{\sqrt{T_m^2(c_m/d_m) + T_{m-1}^2(c_m/d_m)}} \tag{3.16}$$

where $w_{0,m}$ is the $2n$ vector consisting of the subvectors r_0 and $\hat{A}_m r_0$.

One difficulty with the above inequality is that we do not know if the vector $\mathcal{B}_m^m w_{0,m}$ can be very large since the ellipse enclosing the eigenvalues of H_m may potentially miss eigenvalues of A that could cause the Chebyshev polynomials $T_m(A_m)r_0$ to be very large. The spectral radius of $\mathcal{B}$ is provided in Theorem 3.1 to show that the asymptotic growth of the numerator in (3.5) is slower than that of the denominator. This spectral radius is equal to one when the eigenvalues are real and should be close to one in other situations. In fact, as is shown next, the term in the numerator of (3.16) also grows more slowly than the denominator. Its growth *is actually governed by the eigenvalues of H_m, not those of A*, i.e. they act as a power sequence associated with a matrix with spectral radius $a_m/d_m + \sqrt{(a_m/d_m)^2 - 1}$. We start by recalling the following result.

Lemma 3.1. *Let A be any matrix and V_m, H_m the results of m steps of the Arnoldi or Lanczos method applied to A. Then for any polynomial p_j of degree $j \leq m-1$ the following equality holds:*

$$p_j(A)v_1 = V_m p_j(H_m)e_1 \tag{3.17}$$

For a proof see, e.g., References [17,13]. Since the residual polynomial is of degree m, it is convenient to extend this result to polynomials of degree $\leq m$. For this we need to define for an arbitrary scalar, τ, the square matrix obtained by appending a column of zeros except in the diagonal position $(m+1, m+1)$, where the value τ is inserted. In other words,

$$H_m^{\square} = (\bar{H}_m,\ \tau e_{m+1}) \tag{3.18}$$

Then the desired extension of the above lemma can be stated.

Lemma 3.2. *For any polynomial p_k of degree $k \geq 1$, we have*

$$p_k(H_m^{\square}) = \begin{pmatrix} p_k(H_m) & 0 \\ e_m^{\mathrm{T}} q_{k-1}(H_m) & p_k(\tau) \end{pmatrix} \tag{3.19}$$

in which q_{k-1} is a certain polynomial of degree $k-1$. As a result, for $1 \leq k \leq m-1$,

$$p_k(H_m^{\square}) e_1 = \begin{pmatrix} p_k(H_m) e_1 \\ 0 \end{pmatrix} \tag{3.20}$$

and for $1 \leq k \leq m$,

$$p_k(A) v_1 = V_{m+1} p_k(H_m^{\square}) e_1 \tag{3.21}$$

Proof
To prove the first part, it is sufficient to establish the result for the particular polynomials $p_k(t) \equiv t^k$. The proof is by induction and is straightforward. The relation (3.20) follows from (3.19) and the fact that the vector $H_m^k e_1$ has non-zero components only in locations $1, 2, \ldots, k+1$. Finally, to prove (3.21) write the polynomial p_k of degree $\leq m$ in the form

$$p_k(t) = \eta + t s_{k-1}(t)$$

where s_{k-1} is of degree $\leq m-1$. Then,

$$\begin{aligned} p_k(A) v_1 &= \eta v_1 + A s_{k-1}(A) v_1 \\ &= \eta v_1 + A V_m s_{k-1}(H_m) e_1 \\ &= V_{m+1} \left[\eta e_1 + \bar{H}_m s_{k-1}(H_m) e_1 \right] \\ &= V_{m+1} \left[\eta e_1 + [\bar{H}_m, \tau e_{m+1}] \begin{pmatrix} s_{k-1}(H_m) e_1 \\ 0 \end{pmatrix} \right] \\ &= V_{m+1} \left[\eta I + H_m^{\square} s_{k-1}(H_m^{\square}) \right] e_1 \\ &= V_{m+1} p_k(H_m^{\square}) e_1 \end{aligned}$$

■

The result we sought now follows immediately. Its goal is essentially to relate the vector sequence $\mathcal{B}^m w_{0,m}$ which appears in (3.16), with a similar vector sequence obtained from $H_m^{\square}$. This latter sequence is then easier to analyse.

Lemma 3.3. *Assume that an ellipse $E(c_m, d_m, a_m)$ with centre c_m, focal distance d_m and major semi-axis a_m can be found that includes the spectrum of H_m with at least one eigenvalue on the boundary. For any $\tau \in E(c_m, d_m, a_m)$, define the matrices*

$$\hat{H}_m = \beta_m I - \alpha_m H_m^{\square} \quad \text{and} \quad \mathcal{H}_m = \begin{pmatrix} 2\hat{H}_m & -I \\ I & 0 \end{pmatrix} \tag{3.22}$$

with $\alpha_m = 1/d_m$ and $\beta_m = c_m/d_m$. Then the following equality holds:

$$\| \mathcal{B}_m^m w_{0,m} \| = \| \mathcal{H}_m^m z_m \| \tag{3.23}$$

where z_m is the $2(m+1)$-dimensional vector consisting of the subvectors $\|r_0\|e_1$ and $\hat{H}_m(\|r_0\|e_1)$. The spectral radius of the matrix $\mathcal{H}_m$ is given by

$$\rho(\mathcal{H}_m) = \frac{a_m}{d_m} + \sqrt{\left(\frac{a_m}{d_m}\right)^2 - 1} \tag{3.24}$$

and in particular when $a_m = d_m$ (spectrum of $\mathcal{H}_m$ is real) then $\rho(\mathcal{H}_m) = 1$.

Proof
By definition,

$$\mathcal{B}_m^m w_{0,m} = \begin{pmatrix} T_m(\hat{A}_m)r_0 \\ T_{m-1}(\hat{A}_m)r_0 \end{pmatrix}, \quad \mathcal{H}_m^m z_m = \begin{pmatrix} T_m(\hat{H}_m)e_1 \\ T_{m-1}(\hat{H}_m)e_1 \end{pmatrix}$$

For the previous lemma

$$T_m(\hat{A}_m)r_0 = V_{m+1}T_m(\hat{H}_m)\|r_0\|e_1$$

and similarly for $T_{m-1}(\hat{A}_m)r_0$. Thus the equality (3.23) follows immediately. It remains to determine the spectral radius of $\mathcal{H}_m$. The result is similar to that of Theorem 3.1. However, the matrix $H_m^\square$ has the extra eigenvalue τ in addition to the eigenvalues of H_m. Since by assumption τ belong to the ellipse enclosing the spectrum of H_m, this ellipse also contains all the eigenvalues of $H_m^\square$, with at least one on the boundary. ■

Incidentally, it is interesting to note that the residual polynomial is of degree m and as a result the GMRES polynomial minimizes the norm of $p(H_m^\square)e_1$ over all polynomials of degree $\leq m$ such that $p(0) = 1$. The result of Corollary 3.2 replaces the GMRES polynomial in this minimization by a Chebyshev polynomial to provide an upper bound.

4. Numerical examples

The behaviour of the various Chebyshev bounds is now illustrated on two simple examples. We consider an upper triangular matrix of size $n = 50$, with diagonal entries

$$a_{jj} = \sqrt{1/j}\ j = 1, \ldots, n$$

and non-diagonal elements are equal to a constant $-\gamma$. This matrix can become highly non-normal (having an ill-conditioned set of eigenvectors) even for moderate values of γ. For $\gamma = 0.1$ the condition number of the matrix of eigenvectors exceeds 10^{40}. This can easily seen by computing the eigenvectors explicitly. We tested three cases and performed all experiments in Matlab. First, we took $\gamma = 0.001$ and $n = 50$, which produces a moderate condition number of $\kappa_2(X) \approx 8.8$ for the matrix of eigenvectors. The other two examples used a matrix of the same size $n = 50$ but $\gamma = 0.005$ leading to $\kappa_2(X) \approx 3.6 \times 10^4$ and then $\gamma = 0.01$ leading to $\kappa_2(X) \approx 6.8 \times 10^7$. The initial residual is selected to be a random vector. A comparison of the residual norms produced by the (full) GMRES algorithm and three upper bounds is shown in Figures 3, 4 and 5 for these three tests.

Our second set of test matrices used arises from the centred difference discretization of

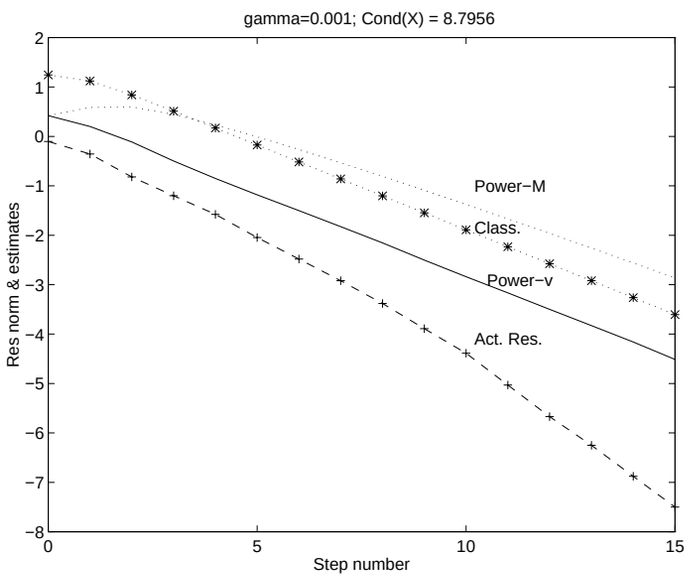

Figure 3. Actual GMRES residuals (+) and a comparison with the classical bound given by Inequality (1.1) (∗), the vector power bound (3.5) (solid line), and the matrix power bound (3.7) (dotted line).

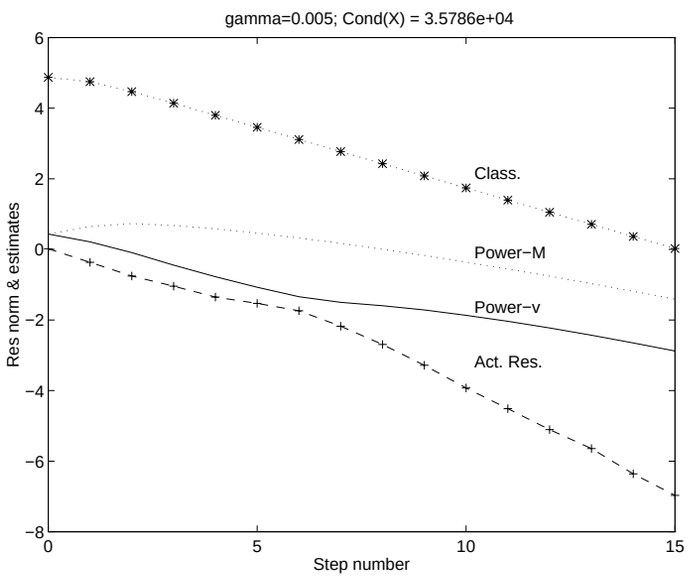

Figure 4. Actual GMRES residuals (+) and a comparison with the classical bound given by Inequality (1.1) (∗), the vector power bound (3.5) (solid line), and the matrix power bound (3.7) (dotted line).

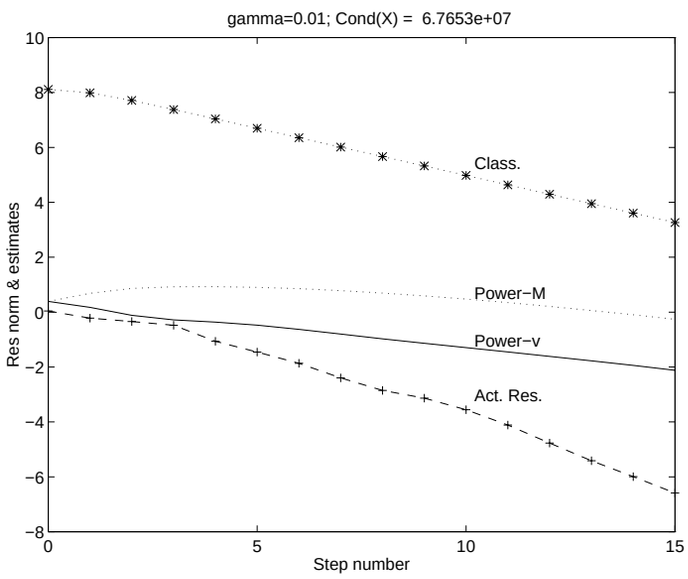

Figure 5. Actual GMRES residuals (+) and a comparison with the classical bound given by Inequality (1.1) (∗), the vector power bound (3.5) (solid line), and the matrix power bound (3.7) (dotted line).

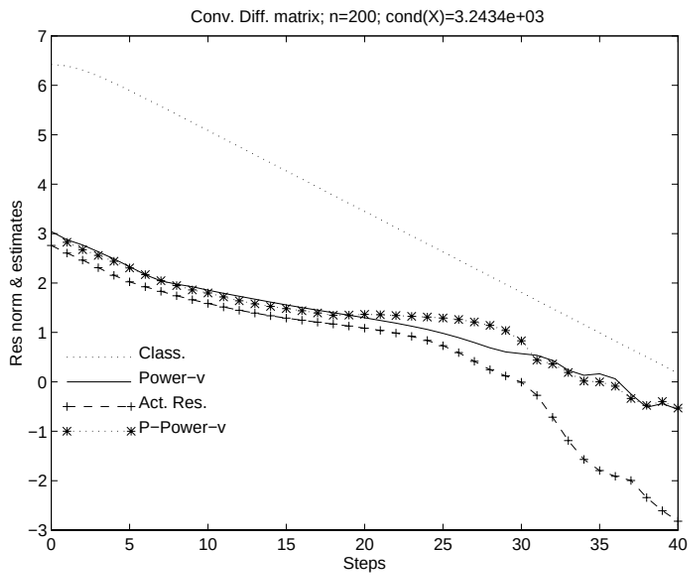

Figure 6. Actual GMRES residuals ('Act. Res.') and a comparison with the classical bound ('Class.') given by Inequality (1.1), the vector power bound ('Power-v') given by Inequality (3.5), and the projected version ('P-Power-v') given by Inequality (3.16).

convection–diffusion operators. Specifically, we selected A in the form

$$A = \begin{pmatrix} B & -I & & & \\ -I & B & -I & & \\ & \ddots & \ddots & \ddots & \\ & & & & -I \\ & & & -I & B \end{pmatrix}, \quad B = \begin{pmatrix} 4 & \delta_- & & & \\ \delta_+ & 4 & \delta_- & & \\ & \ddots & \ddots & \ddots & \\ & & & & \delta_- \\ & & & \delta_+ & 4 \end{pmatrix}$$

with

$$\delta_+ = -1 + \gamma, \quad \delta_- = -1 - \gamma$$

For $0 < \gamma < 1$, the eigenvalues of A are real. We generated a matrix of size 200 by taking a 20×10 grid, leading to a matrix B of size 20 and a block size of 10. The parameter γ was taken equal to 0.4 and this yields a condition number of about 3243 for the matrix of eigenvectors. The plot shown in Figure 6 illustrates the behaviour of the inequalities shown earlier. The matrix power bound was omitted from this experiment but results with the projected version (3.16) of the vector power bound are shown.

A general observation is that the vector power bound (3.5), and its projected version (3.16) are fairly accurate—especially at the beginning of the process. This seems to be true for highly non-normal matrices as well. Because of the similarity of these two bounds, one should expect them to behave similarly. In fact in many of our examples the projected

and non-projected bounds were so close as to be hard to distinguish. The advantage of the projected bound is that all quantities in it are computable, though not *a priori*, and that it does not require eigenvalue estimates from the original matrix.

What distinguishes the bounds provided in this paper from existing ones is precisely this mix of *a priori* and *a posteriori* features. This middle-ground solution cannot be avoided if accurate estimates are desired. It is of course possible to estimate $\|\mathcal{B}^{m+1}w_0\|_2$, and porting this estimate in Inequality (3.5) would give us again a fully *a posteriori* bound. However, the point made here is that this is not advisable in general because of the loss incurred in the quality of the bound. A bound that captures the main behaviour is likely to be better than one that is useless because it is too loose.

5. Conclusion

The error bounds used to analyse the behaviour of Krylov subspace methods in the non-Hermitian case are often too pessimistic and utilize information that is not readily available. We have shown a number of such bounds and a few variants. In general, any attempt to model the behaviour of iterates in the highly non-normal case will lead to poor estimates. One of the main reasons for this is that the tools available for modelling simple matrix-polynomial behaviour are not accurate in the highly non-normal case, i.e. lower bounds obtained by using these tools are typically too loose. One possible solution advocated in some of the results in this paper is to exploit comparisons with sequences of the form $B^k v$ where B is a matrix whose spectral radius is known. In this way, the asymptotic behaviour is understood, as in the classical bounds, but the intermediate upper bounds are not too pessimistic. The use of such bounds has been demonstrated in a few examples, indicating that they are fairly accurate at the initial stages of the process before the super-linear behaviour of GMRES sets in. It is anticipated that these residual bounds that combine *a priori* analysis and *a posteriori* information will in general be sharper than those based on purely analytical results alone. Since the purely *a priori* results are likely to be very poor in the highly non-normal case, there is no alternative but to extract the information that is hard to analyse and treat it as *a posteriori* information. The rest of the information is then analysed in an *a priori* fashion. Of course, one could simply compute the actual residual norm, the ultimate *a posteriori* information, but this provides no insight on the behaviour of the iterative method.

Acknowledgements

This work was supported by NSF Grant CCR-9618827 and by the Minnesota Supercomputer Institute.

REFERENCES

1. Eisenstat SC, Elman HC, Schultz MH. Variational iterative methods for nonsymmetric systems of linear equations. *SIAM Journal on Numerical Analysis*, 1983;**2**:345–357.
2. Chapman A, Saad Y. Deflated and augmented Krylov subspace techniques. *Numerical Linear Algebra with Applications*, 1997;**4**:43–66.
3. Morgan RB. A restarted GMRES method augmented with eigenvectors. *SIAM Journal on Matrix Analysis and Applications*, 1995;**16**:1154–1171.

4. Sorensen DC. Implicit application of polynomial filters in a k-step Arnoldi method. *SIAM Journal on Matrix Analysis and Applications*, 1992;**13**:357–385.
5. Stathopoulos A, Saad Y, Wu K. Dynamic thick restarting of the davidson, and the implicitly restarted arnoldi methods. *SIAM Journal on Scientific Computing*, 1998;**19**:227–245.
6. Gustafson K. Operator trigonometry of iterative methods. *Numerical Linear Algebra with Applications*, 1997;**4**:333–347.
7. Nachtigal NM, Reddy SC, Trefethen LN. How fast are nonsymmetric matrix iterations. *SIAM Journal on Matrix Analysis and Applications*, 1992;**13**:778–795.
8. Greenbaum A, Trefethen LN. GMRES/CR and Arnoldi/Lanczos as matrix approximation problems. *SIAM Journal on Scientific Computing*, 1994;**15**:359–368.
9. Greenbaum A, Strakos Z. Matrices that generate the same Krylov residual spaces. In *Recent Advances in Iterative Methods*, IMA Volumes in Mathematics and Its Applications, Golub G, Luskin M, Greenbaum A (eds), vol. 60. Springer Verlag: New York, 1994; 95–119.
10. Nevanlinna O. How fast can iterative methods be? In *Recent Advances in Iterative Methods*, IMA Volumes in Mathematics and Its Applications, Golub G, Luskin M, Greenbaum A (eds), vol. 60. Springer Verlag: New York, 1994.
11. Godunov SK, Antonov AG, Kirilyuk OP, Kostin VI. *Guaranteed accuracy in numerical linear algebra*. Kluwer, Dordrecht, 1994.
12. Trefethen LN. Pseudospectra of matrices. In *Numerical Analysis, 1991*, Griffiths DF, Watson GA (eds), Longman, 1992; 234–246.
13. Saad Y. *Iterative Methods for Sparse Linear Systems*. PWS publishing: New York, 1996.
14. Elman HC. *Iterative Methods for Large Sparse Nonsymmetric Systems of Linear Equations*. PhD thesis, Yale University, Computer Science Dept., New Haven, CT., 1982.
15. Kato T. *Perturbation Theory for Linear Operators*. Springer Verlag: New York, 1965.
16. Golub GH, Van Loan C. *Matrix Computations*. The John Hopkins University Press, Baltimore, 1989.
17. Saad Y. Analysis of some Krylov subspace approximations to the matrix exponential operator. *SIAM Journal on Numerical Analysis*, 1992;**29**:209–228.
18. Axelsson O. *Iterative Solution Methods*. Cambridge University Press: New York, 1994.
19. Axelsson O, Makarov M. On a generalized conjugate gradient orthogonal residual method. *Numerical Linear Algebra with Applications*, 1995;**2**:467–479.

