

Sharpness in Rates of Convergence
For CG and Symmetric Lanczos Methods¹

Ren-Cang Li²

January 2005

ABSTRACT

Conjugate Gradient (CG) method is often used to solve a positive definite linear system $Ax = b$. Existing bounds suggest that the residual of the k th approximate solution by CG goes to zero like $[(\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)]^k$, where $\kappa \equiv \kappa(A) = \|A\|_2 \|A^{-1}\|_2$ is A 's spectral condition number. It is well-known that for a given positive definite linear system, CG may converge (much) faster, known as superlinear convergence. The question is “*do the existing bounds tell the correct convergence rate in general?*”. An affirmative answer is given here by examples whose CG solutions have errors comparable to the error bounds for all k .

A similar question for the convergence rate of Lanczos algorithm for symmetric eigenvalue problems is addressed and answered firmly, too. Conceivably examples devised here may be good testing problems for linear system and eigensystem solvers.

¹This report is available on the web at <http://www.ms.uky.edu/~math/MAREport/>.

²Department of Mathematics, University of Kentucky, Lexington, KY 40506 (rccli@ms.uky.edu.) This work was supported in part by the National Science Foundation CAREER award under Grant No. CCR-9875201.

Sharpness in Rates of Convergence For CG and Symmetric Lanczos Methods

Ren-Cang Li *

January 2005

Abstract

Conjugate Gradient (CG) method is often used to solve a positive definite linear system $Ax = b$. Existing bounds suggest that the residual of the k th approximate solution by CG goes to zero like $[(\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)]^k$, where $\kappa \equiv \kappa(A) = \|A\|_2 \|A^{-1}\|_2$ is A 's spectral condition number. It is well-known that for a given positive definite linear system, CG may converge (much) faster, known as superlinear convergence. The question is “*do the existing bounds tell the correct convergence rate in general?*”. An affirmative answer is given here by examples whose CG solutions have errors comparable to the error bounds for all k .

A similar question for the convergence rate of Lanczos algorithm for symmetric eigenvalue problems is addressed and answered firmly, too. Conceivably examples devised here may be good testing problems for linear system and eigensystem solvers.

1 Introduction

Krylov subspace projection methods are widely used for large scale matrix computations because they typically require just matrix-vector products to extract enough information to compute desired solutions. Two particular popular and efficient ones for Hermitian matrices are Conjugate Gradient (CG) method for positive definite linear system $Ax = b$ and Lanczos algorithm for symmetric eigenvalue problem $Ax = \lambda x$. Both have well-established convergence theories to go with them in terms of error bounds indicating how fast the computed solutions converge to the desired ones. These bounds usually underestimate the speed of convergence, however. In practice, often the observed convergence is (much) faster than these error bounds suggest. This paper investigates the attainability of these bounds in general.

Consider positive definite linear system $Ax = b$, where A is n -by- n and Hermitian positive definite and b is a vector of dimension n . In exact mathematics, the k th approximate solution x_k by CG is the optimal one in the sense that [3]

$$\|r_k\|_{A^{-1}} = \min_{x \in \mathcal{K}_k} \|b - Ax\|_{A^{-1}}, \quad (1.1)$$

where $r_k = b - Ax_k$, $\mathcal{K}_k \equiv \mathcal{K}_k(A, b)$ is the k th Krylov subspace of A on b defined as

$$\mathcal{K}_k \equiv \mathcal{K}_k(A, b) \stackrel{\text{def}}{=} \text{span}\{b, Ab, \dots, A^{k-1}b\}, \quad (1.2)$$

*Department of Mathematics, University of Kentucky, Lexington, KY 40506 (rccli@ms.uky.edu.) Supported in part by the National Science Foundation CAREER award under Grant No. CCR-9875201.

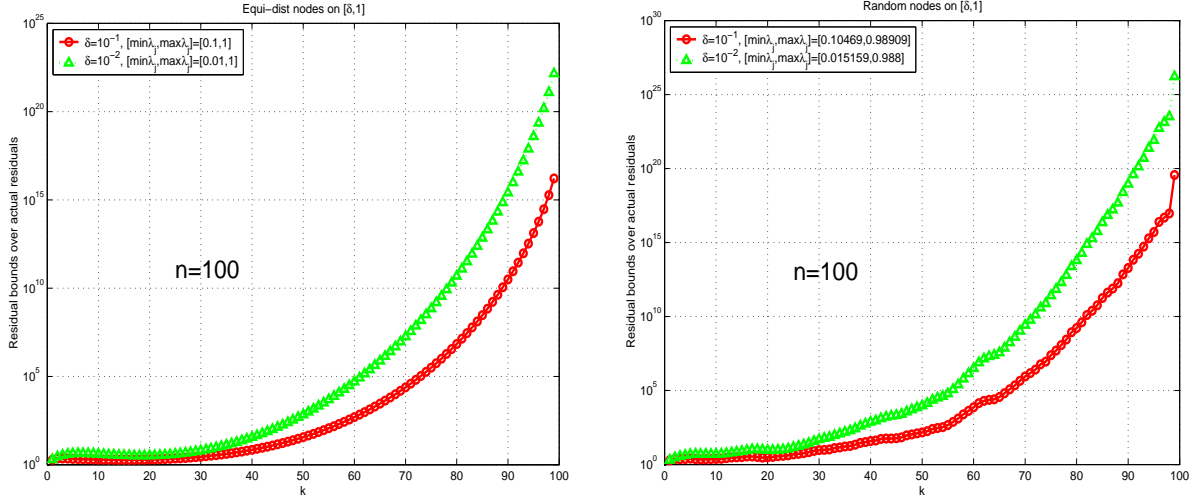


Figure 1.1: Conjugate Gradient Method for $Ax = b$ with $A = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ and b the vector of all ones – Ratios of (1.3) over the actual residuals for equidistant or random distributed λ_j 's

and A^{-1} -vector norm $\|z\|_{A^{-1}} \stackrel{\text{def}}{=} z^* A^{-1} z$. Here the superscript “ $*$ ” takes conjugate transpose. In practice, x_k is computed recursively from x_{k-1} via short term recurrences [3, 5, 6, 18].

CG always converges for positive definite A . In fact with $x_0 = 0$ (and thus $r_0 = b$), we have the following error bound due to Meinardus [14] (see also [3, 6, 18]):

$$\frac{\|r_k\|_{A^{-1}}}{\|r_0\|_{A^{-1}}} = \min_{x \in \mathcal{K}_k} \frac{\|b - Ax\|_{A^{-1}}}{\|r_0\|_{A^{-1}}} \leq 2 \left[\Delta_\kappa^k + \Delta_\kappa^{-k} \right]^{-1}, \quad (1.3)$$

where $\kappa \equiv \kappa(A) = \|A\|_2 \|A^{-1}\|_2$ is the spectral condition number, generic notation $\|\cdot\|_2$ is for either the spectral norm (the largest singular value) of a matrix or the euclidian length of a vector, and

$$\Delta_t \stackrel{\text{def}}{=} \frac{\sqrt{t} + 1}{|\sqrt{t} - 1|} \quad \text{for } t > 0 \quad (1.4)$$

that will be used frequently later for different t . The right-hand side of (1.3) is in fact $T_k(1 + 2\kappa)$, where T_k is the k th Chebyshev polynomial of the first kind; more in Section 2. But CG may converge much faster than this bound indicates for a given linear system, for example, when A has only two different eigenvalues, $r_k \equiv 0$ for $k \geq 2$. Then naturally one may ask: how good is this bound in general? or, does (1.3) always overestimate the speed of convergence substantially? or,

$$\text{is } \sup_{\kappa \leq \delta^{-1}} \max_{1 \leq k \leq n-1} 2 \left[\Delta_\delta^k + \Delta_\delta^{-k} \right]^{-1} \left/ \frac{\|r_k\|_{A^{-1}}}{\|r_0\|_{A^{-1}}} \right. \text{ modestly bounded?} \quad (1.5)$$

Consider $A = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ and b the vector of all ones, where $\lambda_j \in [\delta, 1]$ either randomly or equidistantly distributed on the interval. Figure 1.1 plot the ratios of residual bounds by (1.3) over the actual residuals. What it shows is that initially for small k , bounds by (1.3) are good indications of actual residuals, but as k becomes larger and larger, this bound overestimates the actual ones too much to be of any use. Is the phenomena in Figure 1.1 representative? Often this is what people observed [20], known as superlinear convergence. In [14], Meinardus gave an incomplete answer to (1.3) by devising an $n \times n$ positive

definite linear system $Ax = b$ for which

$$\frac{\|r_{n-1}\|_{A^{-1}}}{\|r_0\|_{A^{-1}}} = 2 \left[\Delta_{\kappa}^{n-1} + \Delta_{\kappa}^{-(n-1)} \right]^{-1},$$

but without saying anything about all other $1 \leq k < n - 1$. Kaniel [9] later offered an existence proof of this result. One of our main contributions of this paper is to show

There is a positive linear system $Ax = b$ with $\kappa(A) \leq \delta^{-1}$ for which the residual of the k th CG approximations is comparable to $2 \left[\Delta_{\delta}^k + \Delta_{\delta}^{-k} \right]^{-1}$ for all $1 \leq k < n$.

(1.6)

Thus the existing bound (1.3) tells the correct rate of convergence for CG in general. Without knowing any additional property for a given positive definite linear system it cannot be substantially improved.

A theory of Kaniel [9] and Saad [17] established similar error bounds on the computed eigenvectors and eigenvalues by Lanczos algorithm on Hermitian matrices, too. Their bounds often suggest slower speed of convergence than the observed rate, also known as superlinear convergence [7, 22]. A similar question to what we ask to CG arises, too, and will be answered firmly by a conclusion like (1.6) to CG.

The rest of this paper is organized as follows. Section 2 derives a decomposition (essentially a QR decomposition) of a Vandermonde matrix with translated Chebyshev zero nodes. A minimization problem whose solution is the key to our main results here is proposed and solved in Section 3. Section 4 presents a result that show the existing error bounds for CG is sharp in general, modulo a modest factor. For similarity reason, the result is stated with applicability to MINRES for positive definite systems, too. In Section 5, we comment on why MINRES could be very slow. Section 6 investigates the sharpness of the existing error bounds for symmetric Lanczos algorithm. Finally concluding remarks are given in Section 7.

Notation. Throughout this paper, $\mathbb{C}^{n \times m}$ is the set of all $n \times m$ complex matrices, $\mathbb{C}^n = \mathbb{C}^{n \times 1}$, and $\mathbb{C} = \mathbb{C}^1$. Similarly define $\mathbb{R}^{n \times m}$, \mathbb{R}^n , and \mathbb{R} except replacing the word *complex* by *real*. I_n (or simply I if its dimension is clear from the context) is the $n \times n$ identity matrix, and e_j is its j th column. The superscript “ \cdot^* ” takes conjugate transpose while “ \cdot^T ” takes transpose only. We shall also adopt MATLAB-like convention to access the entries of vectors and matrices. $i : j$ is the set of integers from i to j inclusive and $i : i = \{i\}$. For vector u and matrix X , $u_{(j)}$ is u 's j th entry, $X_{(i,j)}$ is X 's (i, j) th entry, $\text{diag}(u)$ is the diagonal matrix with $(\text{diag}(u))_{(j,j)} = u_{(j)}$; X 's submatrices $X_{(k:\ell,i:j)}$, $X_{(k:\ell,:)}$, and $X_{(:,i:j)}$ consists of intersections of row k to row ℓ and column i to column j , row k to row ℓ , and column i to column j , respectively. For $t > 0$ and integer $k \geq 1$,

$$\Xi_{t,k} \stackrel{\text{def}}{=} 2 + \sum_{j=1}^{k-1} \left(\Delta_t^j + \Delta_t^{-j} \right)^2 = 2(k-1) + \frac{\Delta_t^{2k} - 1}{\Delta_t^2 - 1} + \frac{1 - \Delta_t^{-2k}}{1 - \Delta_t^{-2}}. \quad (1.7)$$

2 Vandermonde matrix with translated Chebyshev zero nodes

Given n numbers $\alpha_1, \alpha_2, \dots, \alpha_n$ called *nodes*, the associated *Vandermonde Matrix* is defined as

$$V_n \stackrel{\text{def}}{=} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^{n-1} & \alpha_2^{n-1} & \cdots & \alpha_n^{n-1} \end{pmatrix}, \quad (2.1)$$

and we also set $V_{k,n}$ to be its submatrices consisting of its first k rows:

$$V_{k,n} \equiv (V_n)_{(1:k,:)}. \quad (2.2)$$

At their later occurrences, nodes α_j will be explicitly specified. It is perhaps one of the best known structural matrices, and notoriously ill-conditioned for real nodes α_j . Recently various asymptotically optimal lower bounds on its condition number were obtained [1, 13]. As by-products, it is shown that Vandermonde matrices whose nodes are zeros of (translated) Chebyshev polynomials are nearly optimally conditioned among real Vandermonde matrices with nodes on symmetric intervals or nonnegative intervals.

In most part of this paper, V_n 's nodes are the zeros of (translated) Chebyshev polynomials, too. Let

$$T_n(t) = \cos(n \arccos t) \quad \text{for } |t| \leq 1, \quad (2.3)$$

$$= \frac{1}{2} \left(t + \sqrt{t^2 - 1} \right)^n + \frac{1}{2} \left(t - \sqrt{t^2 - 1} \right)^n \quad \text{for } |t| \geq 1 \quad (2.4)$$

which is the n th Chebyshev polynomial of the 1st kind. It frequently shows up in numerical analysis and computations because of its numerous nice properties, for example $|T_n(t)| \leq 1$ for $|t| \leq 1$ and $|T_n(t)|$ grows extremely fast for $|t| > 1$. Later we will need

$$\left| T_n \left(\frac{1+t}{1-t} \right) \right| \equiv \left| T_n \left(\frac{t+1}{t-1} \right) \right| = \frac{1}{2} [\Delta_t^n + \Delta_t^{-n}] \quad \text{for } 1 \neq t > 0. \quad (2.5)$$

The first equality holds because that $T_n(-t) = (-1)^n T_n(t)$. We shall prove the 2nd equality for $0 < t < 1$ only and a proof for $t > 1$ is similar. For $0 < t < 1$, we have

$$\frac{1+t}{1-t} + \sqrt{\left(\frac{1+t}{1-t} \right)^2 - 1} = \frac{1+t+2\sqrt{t}}{1-t} = \frac{1+\sqrt{t}}{1-\sqrt{t}} = \Delta_t,$$

which proves (2.5) for $0 < t < 1$.

Given two real numbers $\alpha < \beta$, and therefore an interval $[\alpha, \beta]$. Define

$$\omega = \frac{\beta - \alpha}{2} > 0, \quad \tau = -\frac{\alpha + \beta}{\beta - \alpha}. \quad (2.6)$$

Then the linear transformation

$$t(x) = \frac{x}{\omega} + \tau = \frac{2}{\beta - \alpha} \left(x - \frac{\alpha + \beta}{2} \right) \quad (2.7)$$

maps $x \in [\alpha, \beta]$ one-to-one and onto $t \in [-1, 1]$. The inverse transformation is $x(t) = \omega(t - \tau)$. The n th *Translated Chebyshev Polynomial* in x of degree n is defined by

$$T_n(x; \omega, \tau) \stackrel{\text{def}}{=} T_n(x/\omega + \tau), \quad (2.8)$$

$$= a_{nn}x^n + a_{n-1n}x^{n-1} + \cdots + a_{1n}x + a_{0n}, \quad (2.9)$$

where $a_{jn} \equiv a_{jn}(\omega, \tau)$ are functions of ω and τ in (2.6). Their explicit dependence on ω and τ is often suppressed for convenience. Define

$$\text{Chebyshev zero nodes: } t_{jn} = \cos \theta_{jn}, \theta_{jn} = \frac{2j-1}{2n}\pi, 1 \leq j \leq n, \quad (2.10)$$

$$\text{translated Chebyshev zero nodes: } t_{jn}^{\text{tr}} = \omega(t_{jn} - \tau), 1 \leq j \leq n. \quad (2.11)$$

It can be seen that t_{jn} ($1 \leq j \leq n$) are the zeros of $T_n(t)$, while t_{jn}^{tr} ($1 \leq j \leq n$) are the zeros of $T_n(x; \omega, \tau)$. Define upper triangular $R_n \in \mathbb{R}^{n \times n}$, a matrix-valued function in ω and τ ,

$$R_n \equiv R_n(\omega, \tau) \stackrel{\text{def}}{=} \begin{pmatrix} a_{00} & a_{01} & a_{02} & \cdots & a_{0n-1} \\ & a_{11} & a_{12} & \cdots & a_{1n-1} \\ & & a_{22} & \cdots & a_{2n-1} \\ & & & \ddots & \vdots \\ & & & & a_{n-1n-1} \end{pmatrix}, \quad (2.12)$$

i.e., the j th column consists of the coefficients of $T_{j-1}(x; \omega, \tau)$.

In the rest of this section, V_n will have the translated Chebyshev zero nodes $\alpha_j = t_{jn}^{\text{tr}}$ ($1 \leq j \leq n$). It can be seen that

$$V_n^T R_n = \mathbf{T}_n \stackrel{\text{def}}{=} \begin{pmatrix} T_0(t_{1n}) & T_1(t_{1n}) & T_2(t_{1n}) & \cdots & T_{n-1}(t_{1n}) \\ T_0(t_{2n}) & T_1(t_{2n}) & T_2(t_{2n}) & \cdots & T_{n-1}(t_{2n}) \\ \vdots & \vdots & \vdots & & \vdots \\ T_0(t_{nn}) & T_1(t_{nn}) & T_2(t_{nn}) & \cdots & T_{n-1}(t_{nn}) \end{pmatrix}. \quad (2.13)$$

We claim that

$$\mathbf{T}_n^T \mathbf{T}_n = (n/2) \text{diag}(2, 1, 1, \dots, 1). \quad (2.14)$$

The orthogonality among the columns of \mathbf{T}_n is a well-know fact, as the result of Gaussian quadrature formula. But we need the diagonal entries of $\mathbf{T}_n^T \mathbf{T}_n$ as well. To this end, we notice $(\mathbf{T}_n)_{(i,j+1)} = T_j(t_{in}) = \cos j\theta_i = \cos \frac{j(2i-1)}{2n}\pi$, and therefore for $0 \leq i, j \leq n-1$

$$\begin{aligned} (\mathbf{T}_n^T \mathbf{T}_n)_{(i+1,j+1)} &= \sum_{k=1}^n (\mathbf{T}_n^T)_{(i+1,k)} (\mathbf{T}_n)_{(k,j+1)} \\ &= \sum_{k=1}^n T_i(t_{kn}) T_j(t_{kn}) \\ &= \sum_{k=1}^n \cos i\theta_{kn} \cos j\theta_{kn} \\ &= \frac{1}{2} \sum_{k=1}^n \cos(i+j)\theta_{kn} + \frac{1}{2} \sum_{k=1}^n \cos(i-j)\theta_{kn}. \end{aligned} \quad (2.15)$$

We now compute $\sum_{k=1}^n \cos \ell \theta_{kn}$, where ℓ is an integer. We claim that

$$\sum_{k=1}^n \cos \ell \theta_{kn} = \begin{cases} n, & \text{for } \ell = 0 \\ 0, & \text{otherwise.} \end{cases} \quad (2.16)$$

The case $\ell = 0$ is clear. Assume that $\ell \neq 0$ for now. Denote $\iota = \sqrt{-1}$ and $\phi = \ell\pi/(2n)$. We have

$$\begin{aligned} 2 \sum_{k=1}^n \cos \ell \theta_{kn} &= 2 \sum_{k=1}^n \cos(2k-1)\phi = \sum_{k=1}^n \left[e^{\iota(2k-1)\phi} + e^{-\iota(2k-1)\phi} \right] \\ &= e^{\iota\phi} \sum_{k=1}^n \left[e^{\iota 2\phi} \right]^{k-1} + e^{-\iota\phi} \sum_{k=1}^n \left[e^{-\iota 2\phi} \right]^{k-1} \\ &= e^{\iota\phi} \frac{1 - [e^{\iota 2\phi}]^n}{1 - e^{\iota 2\phi}} + e^{-\iota\phi} \frac{1 - [e^{-\iota 2\phi}]^n}{1 - e^{-\iota 2\phi}} \\ &= e^{\iota\phi} \frac{1 - e^{\iota 2n\phi}}{1 - e^{\iota 2\phi}} + e^{-\iota\phi} \frac{1 - e^{-\iota 2n\phi}}{1 - e^{-\iota 2\phi}}. \end{aligned}$$

Now if ℓ is odd, then $e^{\pm \iota 2n\phi} = e^{\pm \iota \ell\pi} = -1$ and thus

$$2 \sum_{k=1}^n \cos \ell \theta_{kn} = e^{\iota\phi} \frac{1}{1 - e^{\iota 2\phi}} + e^{-\iota\phi} \frac{1}{1 - e^{-\iota 2\phi}} = 0;$$

if $\ell \neq 0$ is even, then $e^{\pm \iota 2n\phi} = e^{\pm \iota \ell\pi} = 1$ and thus $\sum_{k=1}^n \cos \ell \theta_{kn} = 0$, too. (2.14) is a consequence of (2.15) and (2.16).

Equation (2.13) yields $V_n^T = \mathbf{T}_n R_n^{-1}$ which essentially gives a QR decomposition for V_n^T after normalizing \mathbf{T}_n 's columns to have unit norm. Extracting the first k columns from the both sides of $V_n^T = \mathbf{T}_n R_n^{-1}$ yields the following theorem.

Theorem 2.1 *Let V_n have the translated Chebyshev zero nodes $\alpha_j = t_{jn}^{\text{tr}}$ ($1 \leq j \leq n$) on $[\alpha, \beta]$, and let upper triangular R_k be defined as in (2.12) with n replaced by k and \mathbf{T}_n as in (2.13). Then $V_{k,n}^T = (\mathbf{T}_n)_{(:,1:k)} R_k^{-1}$.*

3 A minimization problem

Any normal matrix A admits the following eigen-decomposition:

$$A = U \Lambda U^*, \quad U^* U = I_n, \quad \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n). \quad (3.1)$$

For CG, A is Hermitian positive definite and thus all $\lambda_j > 0$, and

$$\begin{aligned} \min_{x \in \mathcal{K}_k} \|b - Ax\|_{A^{-1}} &= \min_{\phi_k(0)=1} \|\phi_k(A)b\|_{A^{-1}} \\ &= \min_{\phi_k(0)=1} \|\phi_k(\Lambda)\Lambda^{-1/2}U^*b\|_2 \\ &= \min_{|u_{(1)}|=1} \|\text{diag}(g) V_{k+1,n}^T u\|_2, \end{aligned} \quad (3.2)$$

where $\phi_k(t)$ is a polynomial of degree k , $g = \Lambda^{-1/2}U^*b$, $u \in \mathbb{C}^{k+1}$, and $\alpha_j = \lambda_j$ ($1 \leq j \leq n$) for $V_{k+1,n}$. In general for normal A , including MINRES proposed in [15] for possibly indefinite Hermitian A ,

$$\begin{aligned} \min_{x \in \mathcal{K}_k} \|b - Ax\|_2 &= \min_{\phi_k(0)=1} \|\phi_k(A)b\|_2 \\ &= \min_{\phi_k(0)=1} \|\phi_k(\Lambda)U^*b\|_2 \\ &= \min_{|u_{(1)}|=1} \|\text{diag}(g) V_{k+1,n}^T u\|_2, \end{aligned} \quad (3.3)$$

where $g = U^*b$. Therefore the convergence analysis for both CG and MINRES ends up essentially with bounding $\min_{|u_{(1)}|=1} \|\text{diag}(g) V_{k+1,n}^T u\|_2$, where $g \in \mathbb{C}^n$, $u \in \mathbb{C}^{k+1}$.

In the rest of this section, we shall study minimization problem

$$\min_{|u_{(1)}|=1} \|\text{diag}(g) V_{k,n}^T u\|_2. \quad (3.4)$$

Here for convenience, $k+1$ is relabelled as k , comparing to (3.2) and (3.3). Recall that nodes of $V_{k,n}$ are α_j 's and they are A 's eigenvalues for CG and MINRES. We also need the concept of pseudo-inverse. Let $X \in \mathbb{C}^{n \times m}$ have singular value decomposition (SVD) $X = W\Sigma V^*$, where $W \in \mathbb{C}^{n \times n}$ and $V \in \mathbb{C}^{m \times m}$ are unitary, and $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots)$. The *Moore-Penrose inverse* $X^\dagger \in \mathbb{C}^{m \times n}$ is defined as $X^\dagger \stackrel{\text{def}}{=} V\Sigma^\dagger W^*$, where $\Sigma = \text{diag}(\sigma_1^\dagger, \sigma_2^\dagger, \dots)$ and σ_j^\dagger is σ_j^{-1} if $\sigma_j \neq 0$ and 0 otherwise. It is known that $P_X \stackrel{\text{def}}{=} XX^\dagger$ is the orthogonal projector onto $\text{span}(X)$ [21].

Theorem 3.1 For $Z \in \mathbb{C}^{n \times m}$,

$$\min_{|u_{(1)}|=1} \|Zu\|_2 = \left\| (I - P_{Z(:,2:m)})Z_{(:,1)} \right\|_2 \quad \text{always}, \quad (3.5)$$

$$= \|e_1^T Z^\dagger\|_2^{-1} = [e_1^T (Z^*Z)^{-1} e_1]^{-1/2} \quad \text{if rank}(Z) = m. \quad (3.6)$$

The min is achieved at $u_{\text{opt}} = (1 \quad -Z_{(:,2:m)}^\dagger Z_{(:,1)}^\dagger)^T$ and is $(Z^*Z)^{-1} e_1 / e_1^T (Z^*Z)^{-1} e_1$ if $\text{rank}(Z) = m$.

Proof: (3.5) is a result of re-interpreting $\min_{|u_{(1)}|=1} \|Zu\|_2$ as a linear least squares problem. We now prove (3.6). Set $v = Zu$. Since Z has full column rank, $Z^\dagger = (Z^*Z)^{-1}Z^*$ and thus $u = Z^\dagger v$. This gives a one-one and onto mapping between $u \in \mathbb{C}^m$ and the column space $v \in \text{span}(Z)$. Now

$$\min_{|u_{(1)}|=1} \|Zu\|_2 = \min_u \frac{\|Zu\|_2}{|u_{(1)}|} = \min_{v \in \text{span}(Z)} \frac{\|v\|_2}{|e_1^T Z^\dagger v|} \geq \min_v \frac{\|v\|_2}{|e_1^T Z^\dagger v|} = \|e_1^T Z^\dagger\|_2^{-1}, \quad (3.7)$$

where the last min is achieved at

$$v_{\text{opt}} = \left(e_1^T Z^\dagger \right)^* = Z(Z^*Z)^{-1} e_1 \in \text{span}(Z)$$

which implies the “ \geq ” in (3.7) is actually an equality, and $u_{\text{opt}} = Z^\dagger v_{\text{opt}} / e_1^T Z^\dagger v_{\text{opt}}$. Finally if $\text{rank}(Z) = m$, $Z^\dagger = (Z^*Z)^{-1}Z^*$ and thus

$$\|e_1^T Z^\dagger\|_2 = \sqrt{e_1^T Z^\dagger (Z^\dagger)^* e_1} = \sqrt{e_1^T (Z^*Z)^{-1} e_1}.$$

This completes the proof. ■

REMARK 3.1 It can be seen that for any nonsingular $D \in \mathbb{C}^{m \times m}$ whose 1st row is e_1^T ,

$$\min_{|u_{(1)}|=1} \|ZDu\|_2 = \min_{|u_{(1)}|=1} \|Zu\|_2. \quad (3.8)$$

In particular (3.8) holds for any diagonal D with $D_{(1,1)} = 1$. Theorem 3.1 can be easily modified for minimization subject to $|u_{(j)}| = 1$ for any given $1 \leq j \leq m$. Detail is omitted.

Theorem 3.2 *In this theorem α_j 's are not necessarily real. We have*

$$\min_{|u_{(1)}|=1} \|\text{diag}(g) V_{k,n}^T u\|_2 = \left\| \left(I - P_{\text{diag}(g) (V_{k,n}^T)_{(:,2:k)}} \right) \text{diag}(g) (V_{k,n}^T)_{(:,1)} \right\|_2. \quad (3.9)$$

If, addition, g has no zero entries and all $\alpha_j \neq 0$, then

$$\min_{|u_{(1)}|=1} \|\text{diag}(g) V_{k,n}^T u\|_2 = \begin{cases} 0, & \text{if less than } k \text{ distinct values among all } \alpha_j, \\ \|e_1^T Z^\dagger\|_2^{-1}, & \text{otherwise,} \end{cases} \quad (3.10)$$

where $Z = \text{diag}(g) V_{k,n}^T$.

Proof: (3.9) is a direct consequence of (3.5). We now prove (3.10). Suppose there are ℓ distinct values $\alpha_1, \alpha_2, \dots, \alpha_\ell$ among all α_j . When $\ell < k$, let $v \in \mathbb{C}^k$ whose entries $v_{(j)}$ are the coefficients of t^{j-1} in the polynomial $\prod_{j=1}^\ell (t - \alpha_j)$. Then $v_{(1)} = (-1)^\ell \prod_{j=1}^\ell \alpha_j \neq 0$ and $\text{diag}(g) V_{k,n}^T v = 0$. Hence

$$\min_{|u_{(1)}|=1} \|\text{diag}(g) V_{k,n}^T u\|_2 \leq \|\text{diag}(g) V_{k,n}^T v\|_2 / |v_{(1)}| = 0.$$

Consider now $\ell \geq k$. Then $V_{k,n}$ has a submatrix that is a $k \times k$ Vandermonde matrix with distinct nodes and thus nonsingular; therefore $V_{k,n}$ has full column rank, so does $Z = \text{diag}(g) V_{k,n}^T$ since g has no zero entries. Apply Theorem 3.1 to complete the proof. \blacksquare

Theorem 3.2, combined with (3.2) and (3.3), yield exact formulas of the residuals for CG and MINRES approximations. But such formulas are not very practical and perhaps only useful for theoretical understanding because in practice it is unlikely that the eigen-decomposition (3.1) is known. Nevertheless often estimates to $\min_j \lambda_j$ and $\max_j \lambda_j$, and consequently to $\kappa = \max_j \lambda_j / \min_j \lambda_j$ are available for positive definite A . Therefore bounds on $\min_{|u_{(1)}|=1} \|\text{diag}(g) V_{k,n}^T u\|_2$ in terms of κ can be potentially useful. We shall do so now.

Theorem 3.3 *Assume all $\alpha_j > 0$, and let $\kappa = \max_j \alpha_j / \min_j \alpha_j$. Then*

$$\min_{|u_{(1)}|=1} \|\text{diag}(g) V_{k,n}^T u\|_2 \leq 2\|g\|_2 \left[\Delta_\kappa^{k-1} + \Delta_\kappa^{-(k-1)} \right]^{-1}.$$

Proof: In view of Remark 3.1, we assume $\max_j \alpha_j = 1$ and $\min_j \alpha_j = \delta > 0$ (and thus $\kappa = 1/\delta$). For ω and τ in (2.6) with $[\alpha, \beta] = [\delta, 1]$, $|T_{k-1}(\alpha_j/\omega + \tau)| \leq 1$. Let $v \in \mathbb{R}^k$ with $v_{(j)} = a_{j-1k-1}(\omega, \tau)$ as defined in (2.9). Then

$$\begin{aligned} \min_{|u_{(1)}|=1} \|\text{diag}(g) V_{k,n}^T u\|_2 &\leq \|\text{diag}(g) V_{k,n}^T v\|_2 / |v_{(1)}| \\ &= \sqrt{\sum_{j=1}^n |g_{(j)} T_{k-1}(\alpha_j/\omega + \tau)|^2} / |T_{k-1}(\tau)| \\ &\leq \|g\|_2 / |T_{k-1}(\tau)|, \end{aligned} \quad (3.11)$$

as expected, upon noticing (2.5) and $\tau = (\delta + 1)/(\delta - 1)$. \blacksquare

Theorem 3.3 is not new. In fact it is the well-known error bound (1.3) in disguise. We re-prove it here for completeness and because the proof is short.

Theorem 3.4 *Let $0 < \alpha < \beta$, and let α_j 's be the translated Chebyshev zero nodes t_{jn}^{tr} on $[\alpha, \beta]$, as defined in (2.11). Then*

$$\min_{|u_{(1)}|=1} \frac{\|V_{k,n}^T u\|_2}{\sqrt{n}} = \sqrt{\frac{2}{\Xi_{\delta,k}}}, \quad (3.12)$$

where $\delta = \alpha/\beta$, $\Xi_{\delta,k}$ is defined as in (1.4) and (1.7). When the min is achieved,

$$u_{\text{opt}} = \frac{4}{\Xi_{\delta,k}} R_k \begin{pmatrix} \frac{1}{2}T_0(\tau) \\ T_1(\tau) \\ \vdots \\ T_k(\tau) \end{pmatrix}, \quad V_{k,n}^T u_{\text{opt}} = \frac{4}{\Xi_{\delta,k}} (\mathbf{T}_n)_{(:,1:k)} \begin{pmatrix} \frac{1}{2}T_0(\tau) \\ T_1(\tau) \\ \vdots \\ T_k(\tau) \end{pmatrix}.$$

For $g \in \mathbb{C}^n$, we have

$$\frac{\sqrt{n} \min_j |g_{(j)}|}{\|g\|_2} \sqrt{\frac{2}{\Xi_{\delta,k}}} \leq \min_{|u_{(1)}|=1} \frac{\|\text{diag}(g) V_{k,n}^T u\|_2}{\|g\|_2} \leq \frac{\sqrt{n} \max_j |g_{(j)}|}{\|g\|_2} \sqrt{\frac{2}{\Xi_{\delta,k}}}. \quad (3.13)$$

Proof: Since $V_{k,n}^T$ has full column rank, by Theorem 3.1,

$$\min_{|u_{(1)}|=1} \|V_{k,n}^T u\|_2 = [e_1^T (V_{k,n} V_{k,n}^T)^{-1} e_1]^{-1/2}. \quad (3.14)$$

By Theorem 2.1, we have

$$\begin{aligned} V_{k,n} V_{k,n}^T &= R_k^{-T} [(\mathbf{T}_n)_{(:,1:k)}]^T (\mathbf{T}_n)_{(:,1:k)} R_k^{-1} \\ &= R_k^{-T} (\mathbf{T}_n^T \mathbf{T}_n)_{(1:k,1:k)} R_k^{-1} \\ &= (n/2) R_k^{-T} \text{diag}(2, 1, 1, \dots, 1) R_k^{-1}, \\ (V_{k,n} V_{k,n}^T)^{-1} &= (2/n) R_k \text{diag}(2^{-1}, 1, 1, \dots, 1) R_k^T, \\ e_1^T (V_{k,n} V_{k,n}^T)^{-1} e_1 &= \frac{1}{n} a_{00}^2 + \frac{2}{n} \sum_{j=1}^{k-1} a_{0j}^2, \end{aligned}$$

where $a_{0j} \equiv a_{0j}(\omega, \tau)$ as in (2.9) is the constant term of $T_j(x/\omega + \tau)$ and therefore, by (2.5), $|a_{0j}| = |T_j(\tau)| = \frac{1}{2} [\Delta_\delta^j + \Delta_\delta^{-j}]$. Hence

$$e_1^T (V_{k,n} V_{k,n}^T)^{-1} e_1 = \frac{1}{2n} \left[2 + \sum_{j=1}^{k-1} (\Delta_\delta^j + \Delta_\delta^{-j})^2 \right] = \frac{\Xi_{\delta,k}}{2n}.$$

which, together with (3.14), complete the proof of (3.12). (3.13) follows from

$$\min_j |g_{(j)}| \min_{|u_{(1)}|=1} \|V_{k,n}^T u\|_2 \leq \min_{|u_{(1)}|=1} \|\text{diag}(g) V_{k,n}^T u\|_2 \leq \max_j |g_{(j)}| \min_{|u_{(1)}|=1} \|V_{k,n}^T u\|_2. \quad (3.15)$$

and (3.12). ■

An immediate consequence of Theorem 3.3 for the $V_{k,n}$ in Theorem 3.4 is

$$\sqrt{\frac{2}{\Xi_{\delta,k}}} = \min_{|u_{(1)}|=1} \frac{\|V_{k,n}^T u\|_2}{\sqrt{n}} \leq 2 \left[\Delta_{\delta}^{k-1} + \Delta_{\delta}^{-(k-1)} \right]^{-1}. \quad (3.16)$$

Let us see how tight this upper bound is. We have

$$\begin{aligned} \frac{\text{RHS of (3.16)}}{\text{LHS of (3.16)}} &\leq \sqrt{2} \frac{\left[3(k-1) + 1 + \frac{\Delta_{\delta}^{2k-1}}{\Delta_{\delta}^2 - 1} \right]^{1/2}}{\Delta_{\delta}^{k-1}} \\ &\leq \sqrt{2} \left[\frac{3(k-1) + 1}{\Delta_{\delta}^{2(k-1)}} + \frac{\Delta_{\delta}^2}{\Delta_{\delta}^2 - 1} \right]^{1/2}, \\ \frac{\text{RHS of (3.16)}}{\text{LHS of (3.16)}} &\rightarrow \frac{\sqrt{2} \Delta_{\delta}}{\sqrt{\Delta_{\delta}^2 - 1}} = \frac{1 + \sqrt{\delta}}{\sqrt{2} \sqrt[4]{\delta}} \quad \text{as } k \rightarrow \infty. \end{aligned}$$

At the left of Figure 3.1, this ratio is plotted for $1 \leq k-1 \leq 49$ for $\delta = 10^{-1}$ and 10^{-2} , respectively. Notice the ratio quickly converges to

$$\frac{1 + \sqrt{\delta}}{\sqrt{2} \sqrt[4]{\delta}} = \begin{cases} 1.655068794066460\dots, & \text{for } \delta = 10^{-1}, \\ 2.459674775249768\dots, & \text{for } \delta = 10^{-2}. \end{cases} \quad (3.17)$$

Also plotted at the right of Figure 3.1 are the ratios of upper bounds by $2 \left[\Delta_{\delta}^{k-1} + \Delta_{\delta}^{-(k-1)} \right]^{-1}$ over the actual quantities $\min_{|u_{(1)}|=1} \|\text{diag}(g) V_{k,n}^T u\|_2 / \|g\|_2$, where g is randomly chosen and the nodes for $V_{k,n}$ are the translated Chebyshev zero nodes on the indicated interval. It shows that the existing upper bound is still pretty good for the case, too, as guaranteed by (3.13).

Corollary 3.1 *Let $0 < \delta < 1$, and Δ_{δ} and $\Xi_{\delta,k}$ as defined as in (1.4) and (1.7). Then*

$$\sqrt{\frac{2}{\Xi_{\delta,k}}} \leq \sup_{\alpha_j \in [\delta, 1]} \min_{|u_{(1)}|=1} \frac{\|V_{k,n}^T u\|_2}{\sqrt{n}} \leq 2 \left[\Delta_{\delta}^{k-1} + \Delta_{\delta}^{-(k-1)} \right]^{-1}. \quad (3.18)$$

The next theorem will be used later to construct an indefinite system for which MINRES converges extremely slow.

Theorem 3.5 *Let $\alpha < 0 < \beta$, and let α_j 's be the translated Chebyshev zero nodes t_{jn}^{tr} on $[\alpha, \beta]$, as defined in (2.11). Then*

$$\min_{|u_{(1)}|=1} \frac{\|V_{k,n}^T u\|_2}{\sqrt{n}} = \Phi_{k,\delta} \stackrel{\text{def}}{=} \left[1 + 2 \sum_{j=1}^{k-1} (\cos j\theta_{\delta})^2 \right]^{-1/2}, \quad (3.19)$$

where $\theta_{\delta} \stackrel{\text{def}}{=} \arccos \frac{1-\delta}{1+\delta}$, $\delta = \min\{|\alpha|, \beta\} / \max\{|\alpha|, \beta\}$. For $g \in \mathbb{C}^n$, we have

$$\frac{\sqrt{n} \min_j |g_{(j)}|}{\|g\|_2} \sqrt{\frac{1}{\Phi_{k,\delta}}} \leq \min_{|u_{(1)}|=1} \frac{\|\text{diag}(g) V_{k,n}^T u\|_2}{\|g\|_2} \leq \frac{\sqrt{n} \max_j |g_{(j)}|}{\|g\|_2} \sqrt{\frac{1}{\Phi_{k,\delta}}}. \quad (3.20)$$

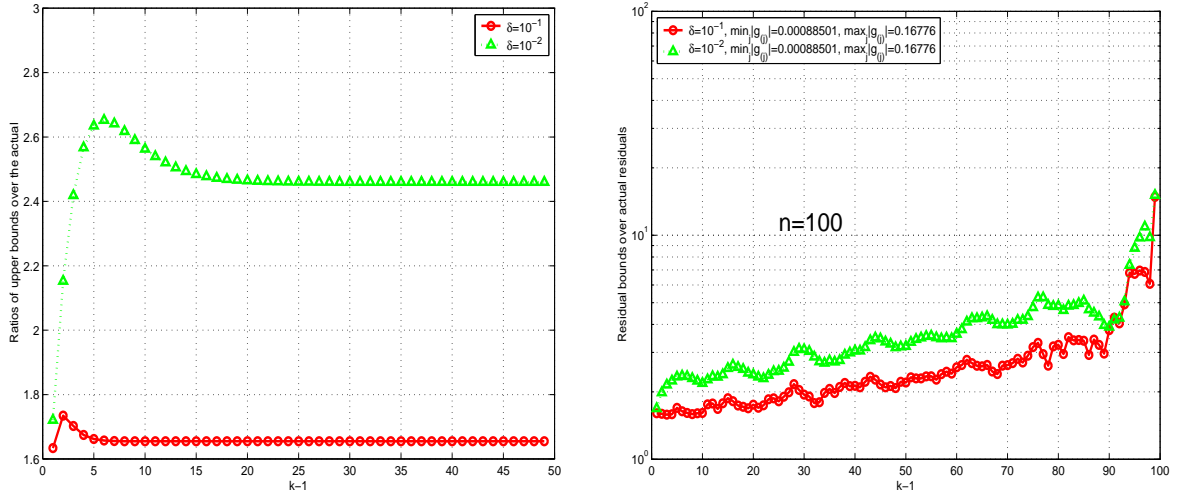


Figure 3.1: Ratios of $2 \left(\Delta_\delta^{k-1} + \Delta_\delta^{-(k-1)} \right)^{-1}$ over $\min_{|u_{(1)}|=1} \|\text{diag}(g) V_{k,n}^T u\|_2 / \|g\|_2$ for $V_{k,n}$ with translated Chebyshev zero nodes on $[\delta, 1]$; **Left:** g is the vector of all ones; **Right:** unit random g .

Proof: It is very much like the proof of Theorem 3.4, except noticing $|a_{0j}| = |T_j(\tau)| = |\cos j\theta_\delta|$. Hence

$$e_1^T (V_{k,n} V_{k,n}^T)^{-1} e_1 = \frac{1}{n} \left[1 + 2 \sum_{j=1}^{k-1} (\cos j\theta_\delta)^2 \right],$$

as needed. ■

4 Sharpness in rate of convergence for CG

Let $0 < \delta \leq 1$. For positive definite A with $\kappa \equiv \kappa(A) \leq 1/\delta$, we have

$$\min_{x \in \mathcal{K}_k} \frac{\|b - Ax\|_p}{\|b\|_p} \leq 2 \left[\Delta_\delta^k + \Delta_\delta^{-k} \right]^{-1}, \quad (4.1)$$

where Δ_δ is defined by (1.4), $p = A^{-1}$ for CG or 2 for MINRES. For CG, it is a consequence of (1.3), and for MINRES, it follows from (3.3) and Theorem 3.3. Theorem 4.1 below answers positively to (1.5), and a similar question for MINRES as well.

Theorem 4.1 *Let positive definite A have eigen-decomposition (3.1) with eigenvalues $\lambda_j = t_{jn}^{\text{tr}}$ on $[\alpha, \beta]$ for $j = 1, 2, \dots, n$, where $0 < \alpha$. Then*

$$\kappa(A) = \frac{1 + \delta + (1 - \delta) \cos \frac{\pi}{2n}}{1 + \delta + (1 - \delta) \cos \frac{\pi}{2n}} = \frac{1}{\delta} - \frac{1 - \delta^2}{4\delta^2} \frac{\pi^2}{4n^2} + \mathcal{O}\left(\frac{1}{n^4}\right) \quad (4.2)$$

$$\leq \frac{1}{\delta}, \quad (4.3)$$

where $\delta = \alpha/\beta$. For $b \in \mathbb{C}^n$,

$$\frac{\sqrt{n} \min_j |g_{(j)}|}{\|g\|_2} \sqrt{\frac{2}{\Xi_{\delta, k+1}}} \leq \min_{x \in \mathcal{K}_k} \frac{\|b - Ax\|_p}{\|b\|_p} \leq \frac{\sqrt{n} \max_j |g_{(j)}|}{\|g\|_2} \sqrt{\frac{2}{\Xi_{\delta, k+1}}}, \quad (4.4)$$

for $1 \leq k < n$, where

$$\begin{aligned} p = A^{-1} \text{ and } g = \Lambda^{-1/2} U^* b \text{ for CG, or} \\ p = 2 \text{ and } g = U^* b \text{ for MINRES.} \end{aligned}$$

(4.4) is an equality if g parallels to the vector of all ones.

Proof: Straightforward calculations yield (4.2) – (4.3). (4.4) is a consequence of (3.2) and (3.13), or (3.3) and (3.13). \blacksquare

Positive definite linear as in the theorem above with g parallel to the vector of all ones can be easily constructed for testing examples for any given b . We outline a procedure here to make g parallel to the vector of all ones for CG. Pick a unitary $U \in \mathbb{C}^{n \times n}$ such that

$$U^* b = \eta (\lambda_1^{1/2} \lambda_2^{1/2} \dots \lambda_n^{1/2})^T,$$

for some $\eta \neq 0$. Such U exists, and in fact it can be taken to be a Householder matrix $I_n - 2ww^*$ as follows: pick η such that

$$|\eta| = \|b\|_2 \left/ \sqrt{\sum_{j=1}^n \lambda_j} \right., \quad \text{and} \quad b^* \eta (\lambda_1^{1/2} \lambda_2^{1/2} \dots \lambda_n^{1/2})^T \text{ is real,}$$

and then set $w = z/\|z\|_2$, where $z = b - \eta (\lambda_1^{1/2} \lambda_2^{1/2} \dots \lambda_n^{1/2})^T$. Finally set $A = U \Lambda U^*$. Then $g = \eta (1 \ 1 \ \dots \ 1)^T$.

Corollary 4.1 *Let $0 < \delta \leq 1$, and denote $\kappa \equiv \kappa(A)$. Subject to positive definite A ,*

$$\sup_{\kappa \leq \delta^{-1}} \max_{1 \leq k \leq n-1} 2 \left[\Delta_\delta^k + \Delta_\delta^{-k} \right]^{-1} \left/ \min_{x \in \mathcal{K}_k} \frac{\|b - Ax\|_p}{\|b\|_p} \right. \leq \max_{1 \leq k \leq n-1} \frac{\sqrt{2\Xi_{\delta,k+1}}}{\Delta_\delta^k + \Delta_\delta^{-k}}, \quad (4.5)$$

where $p = A^{-1}$ or 2.

Proof: It is a consequence of (4.1) and (4.4) for g being the vector of all ones. \blacksquare

5 Extremely slow convergence examples for MINRES

MINRES was originally proposed for indefinite symmetric (Hermitian) linear systems [15]. One may view GMRES [19] as its generalization to a generally non-symmetric linear systems. Both are often used today in solving large and sparse linear systems in practice, but they may be slow in convergence. In what follows, we shall construct two artificial examples for which MINRES converges extremely slowly or does not converge at all. As our proofs above suggest, it suffices for us to look for those nodes α_j so that the associated

$$\min_{|u_{(1)}|=1} \frac{\|V_{k+1,n}^T u\|_2}{\sqrt{n}} \quad (5.1)$$

either goes to zero extremely slowly as k increases or does not decrease at all. The construction of the first example is based on Theorem 3.5. We need to make sure no $\alpha_j \neq 0$, or equivalently no $\cos \theta_{j,n} = \cos \theta_\delta$, and at the same time to make $\max_j |\alpha_j| / \min_j |\alpha_j|$ not too big. To meet both objectives, we shall pick δ as follows, and then make $\alpha < 0 < \beta$ to achieve that δ .

- If $n + 1$ is even, then $\theta_{(n+1)/2n} = \pi/2$. Pick δ so that $\theta_\delta = \pi/2 - \pi/2n$, i.e.,

$$\delta = \frac{1 - \cos \theta_\delta}{1 + \cos \theta_\delta} = 1 - \frac{\pi}{n} + \mathcal{O}\left(\frac{1}{n^2}\right),$$

and thus

$$\begin{aligned} \frac{\min_j |\alpha_j|}{\max\{|\alpha|, \beta\}} &= \frac{1 + \delta}{2} \cos \theta_\delta = \frac{\cos \theta_\delta}{1 + \cos \theta_\delta} = \frac{\sin(\pi/2n)}{1 + \sin(\pi/2n)}, \\ \frac{\max_j |\alpha_j|}{\max\{|\alpha|, \beta\}} &= \frac{1 - \delta}{2} + \frac{1 + \delta}{2} \cos \theta_{1n} = \frac{\sin(\pi/2n) + \cos(\pi/2n)}{1 + \sin(\pi/2n)}, \\ \frac{\max_j |\alpha_j|}{\min_j |\alpha_j|} &= \frac{\sin(\pi/2n) + \cos(\pi/2n)}{\sin(\pi/2n)} = 1 + \frac{2n}{\pi} - \frac{\pi}{6n} + \mathcal{O}\left(\frac{1}{n^3}\right); \end{aligned}$$

- If $n + 1$ is even, then $\theta_{n/2n} = \pi/2 - \pi/(2n)$. Pick $\delta = 1$, i.e., $\theta = \pi/2$, and

$$\begin{aligned} \frac{\min_j |\alpha_j|}{\max\{|\alpha|, \beta\}} &= \frac{1 + \delta}{2} \sin \frac{\pi}{2n} = \sin \frac{\pi}{2n}, \\ \frac{\max_j |\alpha_j|}{\max\{|\alpha|, \beta\}} &= \frac{1 - \delta}{2} + \frac{1 + \delta}{2} \cos \theta_{1n} = \cos(\pi/2n), \\ \frac{\max_j |\alpha_j|}{\min_j |\alpha_j|} &= \frac{\cos(\pi/2n)}{\sin(\pi/2n)} = \frac{2n}{\pi} - \frac{\pi}{6n} + \mathcal{O}\left(\frac{1}{n^3}\right). \end{aligned}$$

In both cases, $\kappa = \mathcal{O}(n)$, but (5.1) barely moves as k increases, according to Theorem 3.5.

The second example is $Ax = b$ with A being unitary. In practice this is not a problem at all because $x = A^*b$ gives the solution right away. But nevertheless it is a linear system for which MINRES (GMRES) makes no progress at all. As mentioned before, it is sufficient for us to give α_j for which (5.1) is independent of k . The α_j 's are

$$\text{the } n\text{th roots of unit: } \alpha_j = e^{i2j\pi/n}.$$

Then $V_{k+1,n}^T$ has orthogonal columns, and thus $\min_{|u_{(1)}|=1} \frac{\|V_{k+1,n}^T u\|_2}{\sqrt{n}} = 1$ for all k .

The two examples are rather extreme, especially the second one, and thus they are unlikely representatives of linear systems from applications. But conceivably the first example could be an excellent test problem for solvers of indefinite linear systems.

6 Sharpness in rate of convergence for Lanczos algorithm

Lanczos algorithm for finding some eigenvalues and corresponding eigenvectors of a Hermitian matrix A may be compactly described as follows. First apply Lanczos process [16] to get

$$AQ_k = Q_k T_k + f_k e_k^T, \quad Q_k = (q_1 \ q_2 \ \cdots \ q_k), \quad (6.1)$$

where Q_k has orthonormal columns, $q_1 = b/\|b\|_2$ and b is a pre-chosen vector, f_k (a vector of dimension n) satisfies $Q_k^* f_k = 0$. Then the eigenvalue problem for T_k is solved. Let (μ, z) be

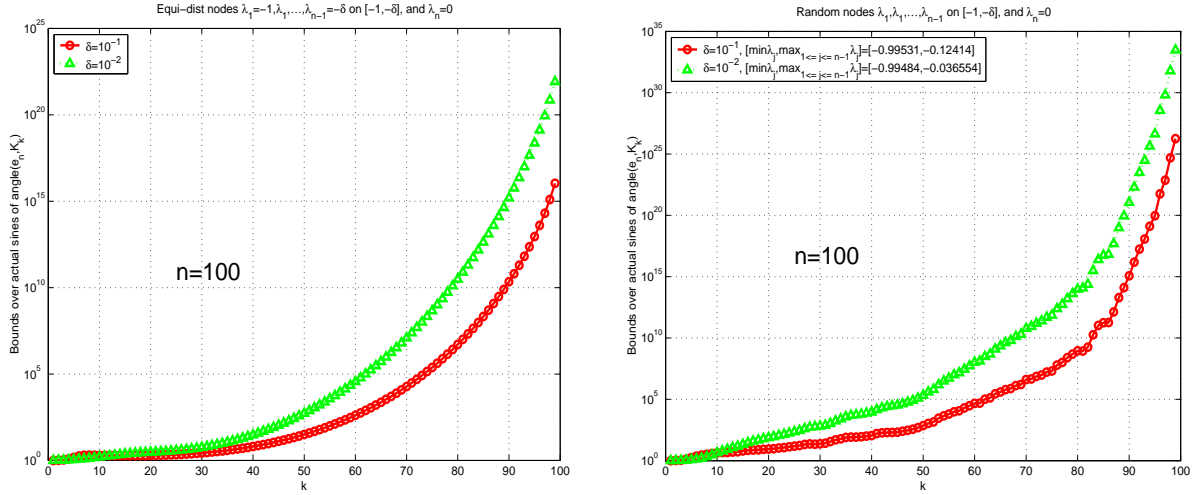


Figure 6.1: Lanczos Algorithm for $Ax = \lambda x$ with $A = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ on b the vector of all ones – Ratios of bounds (6.4) over the actual $\sin \angle(e_n, \mathcal{K}_k)$ for $\lambda_n = 0$, and equidistant or random distributed $\{\lambda_j\}_{j=1}^{n-1}$.

an eigenpair of T_k , i.e., $T_k z = \mu z$. An approximate eigenpair $(\mu, Q_k z)$, so-called *Ritz pair*, for A is obtained. It can be seen that

$$\text{span}\{q_1, q_2, \dots, q_k\} = \mathcal{K}_k(A, b).$$

Assume that A admits eigen-decomposition (3.1). Then $U_{(:,j)}$ is the eigenvector of A associated with eigenvalue λ_j . For the sake of presentation, assume

$$\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n. \quad (6.2)$$

Naturally, we ask how good an eigenvalue of T_k approximates A 's, and how far $U_{(:,j)}$ is from $\mathcal{K}_k(A, b)$. A well-developed theory for this is due to Kaniel [9] and Saad [17], and if more detailed information on A 's eigenvalue distribution is available, better bounds can be derived, too [16]. From the distribution point of view in the limiting sense as k and n both goes to ∞ but with fixed ratio k/n , it was studied which eigenvalues are found and what are their associated convergence speeds in [4, 10].

Consider again $A = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ with either randomly or equidistantly distributed $\{\lambda_j\}_{j=1}^{n-1}$ on $[-1, -\delta]$ and $\lambda_n = 0$. If Lanczos algorithm is applied to A on vector b of all ones, Figure 6.1 plots the ratios of bounds by (6.4) and the actual sines of $\angle(e_n, \mathcal{K}_k)$. It shows the disproportions between the bounds and the actual sines. The main contribution of this section is to show they are however sharp in general, despite Figure 6.1.

6.1 Eigenvector Convergence

Let us look at how close $U_{(:,j)}$ is to $\mathcal{K}_k(A, b)$. This can be turned into the following minimization problem

$$\begin{aligned} \min_{\phi_{k-1}} \|U_{(:,j)} - \phi_{k-1}(A)b\|_2 &= \min_{\phi_{k-1}} \|e_j - \phi_{k-1}(\Lambda)U^*b\|_2 \\ &= \min_{u_{(1)}=1} \|(e_j \text{diag}(g)V_{k,n}^T)u\|_2, \end{aligned} \quad (6.3)$$

where ϕ_{k-1} denote a polynomial of degree no bigger than $k-1$, $g = U^*b$, and $\alpha_j = \lambda_j$ for $V_{k,n}$. (6.3) does not exactly say if there is a Ritz vector approximates $U_{(:,j)}$ (well). For which the reader is referred to [8], where it is proved that under suitable separation conditions if $U_{(:,j)}$ is close to $\mathcal{K}_k(A, b)$, then there is a Ritz vector that approximates $U_{(:,j)}$ well.

Theorem 6.1 Assume λ_j are ordered as in (6.2), and let $V_{k,n}$ be defined with $\alpha_j = \lambda_j$. Then

$$\min_{u_{(1)}=1} \|(e_j \text{ diag}(g)V_{k,n}^T)u\|_2 \leq \frac{\frac{\gamma}{\varrho} \|g_{(1:j-1)}\|_2}{\sqrt{\frac{\gamma^2}{\varrho^2} \|g_{(1:j-1)}\|_2^2 + \frac{|g_{(j)}|^2}{4} \left[\Delta_\delta^{k-1-(n-j)} + \Delta_\delta^{-(k-1)+(n-j)} \right]^2}}, \quad (6.4)$$

where $\delta = \frac{\lambda_j - \lambda_{j-1}}{\lambda_j - \lambda_1}$, $\gamma = \prod_{i=j+1}^n (\lambda_i - \lambda_1)$, and $\varrho = \prod_{i=j+1}^n (\lambda_i - \lambda_j)$.

Proof: For ω and τ in (2.6) with $[\alpha, \beta] = [\lambda_1, \lambda_{j-1}]$, $|T_{k-1-(n-j)}(\lambda_i/\omega + \tau)| \leq 1$ for $1 \leq i \leq j-1$. Let $v \in \mathbb{C}^{k+1}$ with $v_{(1)} = 1$ and $v_{(i)} = \eta c_{i-2}$ for $2 \leq i \leq k+1$, where c_i are coefficients of t^i in

$$\phi_{k-1}(t) = \prod_{i=j+1}^n (t - \lambda_i) \times T_{k-1-(n-j)}(t/\omega + \tau),$$

$\eta \in \mathbb{C}$ to be determined such that $\eta g_{(j)} \zeta = -|\eta g_{(j)} \zeta|$, and $\zeta = \phi_{k-1}(\lambda_j)$. Then

$$\begin{aligned} \min_{u_{(1)}=1} \|(e_j \text{ diag}(g)V_{k,n}^T)u\|_2 &\leq \|(e_j \text{ diag}(g)V_{k,n}^T)v\|_2 \\ &\leq \left[\eta^2 \gamma^2 \|g_{(1:j-1)}\|_2^2 + (1 - |g_{(j)} \eta \zeta|)^2 \right]^{1/2}. \end{aligned}$$

Now it is clear that $|\eta|$ should be chosen to minimize the last quantity above, which gives

$$|\eta| = \frac{|g_{(j)} \zeta|}{\gamma^2 \|g_{(1:j-1)}\|_2^2 + |g_{(j)} \zeta|^2}$$

and

$$\min_{u_{(1)}=1} \|(e_j \text{ diag}(g)V_{k,n}^T)u\|_2 \leq \frac{\frac{\gamma}{\varrho} \|g_{(1:j-1)}\|_2}{\sqrt{\gamma^2 \|g_{(1:j-1)}\|_2^2 + |g_{(j)} \zeta|^2}}. \quad (6.5)$$

Now by (2.6),

$$\frac{\lambda_j}{\omega} + \tau = \frac{2\lambda_j}{\lambda_{j-1} - \lambda_1} - \frac{\lambda_{j-1} + \lambda_1}{\lambda_{j-1} - \lambda_1} = \frac{1 + \delta}{1 - \delta},$$

and thus by (2.5), we have (6.4). ■

REMARK 6.1 (6.4) is equivalent to an existing bound of Kaniel and Saad [16, p.270]. Let

$$\psi = \angle(U_{(:,j)}, \mathcal{K}_k), \quad t = \|g_{(1:j-1)}\|_2 / |g_{(j)}|, \quad M = \frac{1}{2} \left[\Delta_\delta^{k-1-(n-j)} + \Delta_\delta^{-(k-1)+(n-j)} \right], \quad \epsilon = \frac{t}{M}.$$

(6.4) can be rewritten as $\sin \psi \leq \epsilon / \sqrt{1 + \epsilon^2}$ which implies $\cos \psi \geq 1 / \sqrt{1 + \epsilon^2}$ and consequently $\tan \psi \leq \epsilon$, a bound of Kaniel and Saad. On the other hand $\tan \psi \leq \epsilon$ implies $(\sin \psi)^2 \leq \epsilon^2 [1 - (\sin \psi)^2]$ to get $\sin \psi \leq \epsilon / \sqrt{1 + \epsilon^2}$ that is (6.4).

Lemma 6.1 Let $\alpha < \beta < 0$, and let $V_{k,n}$ have nodes $\alpha_j = t_{j,n-1}^{\text{tr}}$ for $1 \leq j \leq n-1$ and $\alpha_n = 0$. If $g_{(j)} = \rho$ for $1 \leq j \leq n-1$, then

$$\min_{|u_{(1)}|=1} \|(e_n \quad \text{diag}(g)V_{k,n}^T)u\|_2 = \left[1 + \frac{|g_{(n)}|^2}{|\rho|^2} \frac{1}{2(n-1)} \Xi_{\delta,k} \right]^{-1/2}.$$

Proof: Notice that \mathbf{T}_n in (2.13) depends on n only. In this proof, we need \mathbf{T}_{n-1} . Let upper triangular R_k be defined as in (2.12) with n replaced by k . Denote $\eta = g_{(n)}$, and set

$$Z = (e_n \quad \text{diag}(g)V_{k,n}^T) \equiv \begin{pmatrix} 0 & \rho V_{k,n-1}^T \\ 1 & \eta e_1^T \end{pmatrix}$$

from which it can be seen that it has full column rank. By Theorem 3.1, we need to compute $[e_1^T(Z^*Z)^{-1}e_1]^{-1/2}$. We have

$$Z \begin{pmatrix} 1 \\ R_k \end{pmatrix} = \begin{pmatrix} 0 & \rho (\mathbf{T}_{n-1})_{(:,1:k)} \\ 1 & \eta e_1^T R_k \end{pmatrix}, \quad Z = \begin{pmatrix} 0 & \rho (\mathbf{T}_{n-1})_{(:,1:k)} \\ 1 & \eta e_1^T R_k \end{pmatrix} \begin{pmatrix} 1 & \\ & R_k^{-1} \end{pmatrix}$$

to get

$$\begin{aligned} Z^*Z &= \begin{pmatrix} 1 & \\ & R_k^{-*} \end{pmatrix} \begin{pmatrix} 1 & \eta e_1^T R_k \\ \eta^* R_k^* e_1 & |\rho|^2 D + |\eta|^2 R_k^* e_1 e_1^T R_k \end{pmatrix} \begin{pmatrix} 1 & \\ & R_k^{-1} \end{pmatrix} \\ &= \begin{pmatrix} 1 & \\ & R_k^{-*} \end{pmatrix} \begin{pmatrix} 1 & \\ \eta^* R_k^T e_1 & I \end{pmatrix} \begin{pmatrix} 1 & \\ & |\rho|^2 D \end{pmatrix} \begin{pmatrix} 1 & \eta e_1^T R_k \\ & I \end{pmatrix} \begin{pmatrix} 1 & \\ & R_k^{-1} \end{pmatrix}, \end{aligned}$$

where $D = [(\mathbf{T}_{n-1})_{(:,1:k)}]^* (\mathbf{T}_{n-1})_{(:,1:k)} = (\mathbf{T}_{n-1}^T \mathbf{T}_{n-1})_{(1:k,1:k)} = \frac{n-1}{2} \text{diag}(2, 1, 1, \dots, 1)$ by (2.14). Therefore

$$\begin{aligned} e_1^T (Z^*Z)^{-1} e_1 &= e_1^T \begin{pmatrix} 1 & -\eta e_1^T R_k \\ & I \end{pmatrix} \begin{pmatrix} 1 & \\ & |\rho|^{-2} D^{-1} \end{pmatrix} \begin{pmatrix} 1 & \\ -\eta^* R_k^T e_1 & I \end{pmatrix} e_1 \\ &= 1 + |\eta|^2 e_1^T R_k |\rho|^{-2} D^{-1} R_k^T e_1 \\ &= 1 + \frac{|\eta|^2}{|\rho|^2 (n-1)} \left(a_{00}^2 + 2 \sum_{j=1}^{k-1} a_{0j}^2 \right) \\ &= 1 + \frac{|\eta|^2}{|\rho|^2} \frac{1}{2(n-1)} \Xi_{\delta,k}, \end{aligned}$$

similarly to at the end of proof of Theorem 3.4. ■

Theorem 6.2 Let α_j 's be as given in Lemma 6.1. Suppose A is a Hermitian matrix whose eigenvalues are $\lambda_j = \alpha_j$ and admits eigen-decomposition (3.1) and $g = U^*b$. Apply Lanczos algorithm on A with $q_1 = b$ as in (6.1). Then

$$\min_{x \in \mathcal{K}_k} \|U_{(:,n)} - x\|_2 \leq \left[1 + \frac{|g_{(n)}|^2}{\max_{1 \leq j \leq n-1} |g_{(j)}|^2} \frac{1}{2(n-1)} \Xi_{\delta,k} \right]^{-1/2}, \quad (6.6)$$

$$\min_{x \in \mathcal{K}_k} \|U_{(:,n)} - x\|_2 \geq \left[1 + \frac{|g_{(n)}|^2}{\min_{1 \leq j \leq n-1} |g_{(j)}|^2} \frac{1}{2(n-1)} \Xi_{\delta,k} \right]^{-1/2}. \quad (6.7)$$

Both are equalities if $|g_{(1)}| = |g_{(2)}| = \dots = |g_{(n-1)}|$.

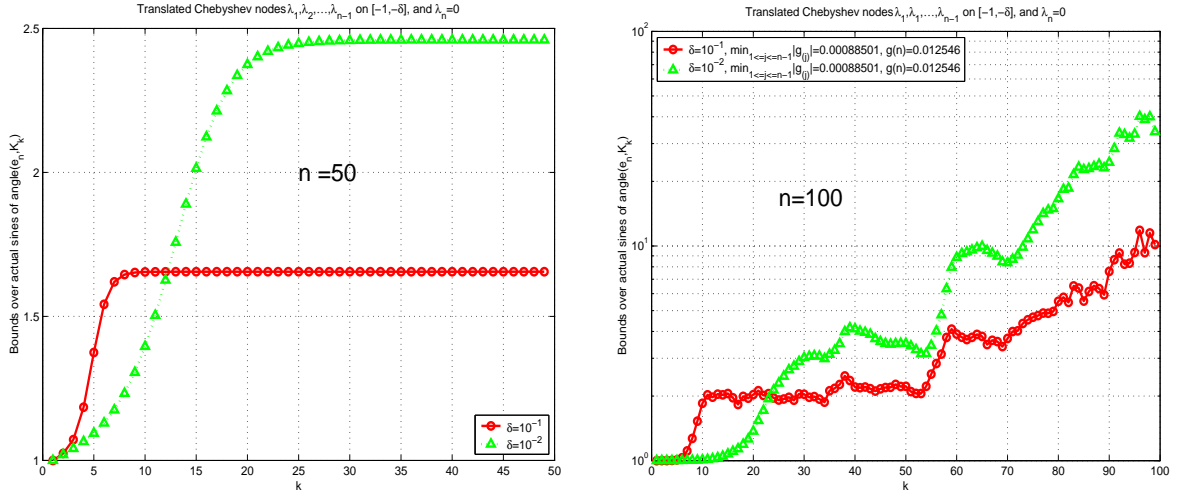


Figure 6.2: Ratios of upper bounds by (6.4) for $j = n$ over $\sin \angle(U_{(:,n)}, \mathcal{K}_k)$. Eigenvalues are $\lambda_j = t_{j,n-1}^{\text{tr}}$ for $1 \leq j \leq n-1$ on $[-1, -\delta]$ and $\lambda_n = 0$. **Left:** g is the vector of all ones; **Right:** g is random.

Proof: They are consequences of Lemma 6.1 and

$$\min_{u_{(1)}=1} \|(e_n \text{diag}(\tilde{g})V_{k,n}^T)u\|_2 \leq \min_{u_{(1)}=1} \|(e_n \text{diag}(g)V_{k,n}^T)u\|_2 \leq \min_{u_{(1)}=1} \|(e_n \text{diag}(\hat{g})V_{k,n}^T)u\|_2,$$

where $\tilde{g}_{(j)} = \min_{1 \leq i \leq n-1} |g_{(j)}|$ and $\hat{g}_{(j)} = \max_{1 \leq i \leq n-1} |g_{(j)}|$ for $1 \leq j \leq n-1$, and $\tilde{g}_{(n)} = \hat{g}_{(n)} = g_{(n)}$. ■

In order to see how good the existing bound by Theorem 6.1 for $j = n$ is, we plot in Figure 6.2

1. the ratio of the upper bound over the lower bound (i.e., the right-hand side of (6.4) for $j = n$ over that of (6.7)) when g is the vector of all ones for which (6.7) is an equality;
2. the ratio of the upper bound (6.4) for $j = n$ over $\min_{x \in \mathcal{K}_k} \|U_{(:,n)} - x\|_2 \equiv \sin \angle(U_{(:,n)}, \mathcal{K}_k)$ when g is a random vector and $\|g\|_2 = 1$.

In the first case when g is the vector of all ones, the ratio quickly approaches a constant because for $j = n$

$$\frac{\text{RHS of (6.4)}}{\text{RHS of (6.7)}} \rightarrow \frac{\sqrt{2} \Delta_\delta}{\sqrt{\Delta_\delta^2 - 1}} = \frac{1 + \sqrt{\delta}}{\sqrt{2} \sqrt[4]{\delta}} \quad \text{as } k \rightarrow \infty$$

which, together with (3.17), explain the left plot in Figure 6.2. With random g , however, the ratios behaves irregularly but still grow slowly proportionally to $|g_{(n)}|^{-1}$.

Theorem 6.2 only shows that (6.4) for $j = n$ is quite sharp in general. The situation for $j \neq n$ appears to be very complicated, and it is not clear how to best approach the situation.

6.2 Eigenvalue Convergence

A is Hermitian; so is $T_k = Q_k^* A Q_k$. We expect the largest eigenvalue μ_k of T_k best approximates λ_n . We have

$$\mu_k = \max_z \frac{z^* T_k z}{z^* z} = \max_z \frac{z^* Q_k^* A Q_k z}{z^* Q_k^* Q_k z} = \lambda_n + \max_{u \in \mathcal{K}_k} \frac{u^* (A - \lambda_n I) u}{u^* u}$$

$$= \lambda_n + \max_{\phi_{k-1}} \frac{[\phi_{k-1}(A - \lambda_n I)b]^*(A - \lambda_n I)[\phi_{k-1}(A - \lambda_n I)b]}{[\phi_{k-1}(A - \lambda_n I)b]^*[\phi_{k-1}(A - \lambda_n I)b]},$$

since $\mathcal{K}_k(A, b) = \mathcal{K}_k(A - \lambda_n I, b)$. Substitute $A = U\Lambda U^*$ to get

$$0 \geq \mu_k - \lambda_n = - \min_u \frac{\|\text{diag}((\lambda_n I - \Lambda)^{1/2} g) V_{k,n}^T u\|_2^2}{\|\text{diag}(g) V_{k,n}^T u\|_2^2}, \quad (6.8)$$

where $g = U^*b$, and $\alpha_j = \lambda_j - \lambda_n$ for the nodes of $V_{k,n}$. Theorem 6.3 below rephrases an existing result of Kaniel and Saad [16].

Theorem 6.3 *Assume λ_j are ordered as in (6.2), and let $V_{k,n}$ have nodes $\alpha_j = \lambda_j - \lambda_n \leq 0$, and $\Omega = -\text{diag}(\alpha_1, \alpha_2, \dots, \alpha_n)$. Then*

$$\min_u \frac{\|\Omega^{1/2} \text{diag}(g) V_{k,n}^T u\|_2^2}{\|\text{diag}(g) V_{k,n}^T u\|_2^2} \leq 4(\lambda_n - \lambda_1) \frac{\|g_{(1:n-1)}\|_2^2}{|g_{(n)}|^2} \left[\Delta_\delta^{k-1} + \Delta_\delta^{-(k-1)} \right]^2, \quad (6.9)$$

where $\delta = \frac{\lambda_n - \lambda_{n-1}}{\lambda_n - \lambda_1}$.

Proof: For ω and τ in (2.6) with $[\alpha, \beta] = [\alpha_1, \alpha_{n-1}]$, $|T_{k-1}(\alpha_j/\omega + \tau)| \leq 1$ for $1 \leq j \leq n-1$. Let $v \in \mathbb{C}^k$ with $v_{(j)} = a_{j-1} T_{k-1}(\omega, \tau)$ as defined in (2.9). Then

$$\begin{aligned} \min_u \frac{\|\Omega^{1/2} \text{diag}(g) V_{k,n}^T u\|_2^2}{\|\text{diag}(g) V_{k,n}^T u\|_2^2} &\leq \frac{\|\Omega^{1/2} \text{diag}(g) V_{k,n}^T v\|_2^2}{\|\text{diag}(g) V_{k,n}^T v\|_2^2} \\ &\leq (\lambda_n - \lambda_1) \frac{\|g_{(1:n-1)}\|_2^2}{|g_{(n)} v_{(1)}|^2} \\ &= (\lambda_n - \lambda_1) \frac{\|g_{(1:n-1)}\|_2^2}{|g_{(n)}|^2} \frac{1}{|T_{k-1}(\tau)|^2}. \end{aligned} \quad (6.10)$$

Now by (2.6)

$$\tau = \frac{\alpha_1 + \alpha_{n-1}}{\alpha_1 - \alpha_{n-1}} = \frac{1 + \delta}{1 - \delta},$$

and (2.5), we have (6.9). ■

Lemma 6.2 *Let $\alpha < \beta < 0$, and let $V_{k,n}$ have nodes $\alpha_j = t_{j,n-1}^{\text{tr}}$ for $1 \leq j \leq n-1$ and $\alpha_n = 0$. If $g_{(j)} = \rho$ for $1 \leq j \leq n-1$, then*

$$\min_u \frac{\|(\text{diag}(g) V_{k,n}^T u)_{(1:n-1)}\|_2^2}{\|\text{diag}(g) V_{k,n}^T u\|_2^2} = \left[1 + \frac{|g_{(n)}|^2}{|\rho|^2} \frac{1}{2(n-1)} \Xi_{\delta,k} \right]^{-1},$$

where $\delta = \beta/\alpha$.

Proof: Notice $V_{k,n} = (V_{k,n-1} \ e_1)$ to get $\|\text{diag}(g) V_{k,n}^T u\|_2^2 = |\rho|^2 \|(V_{k,n}^T u)_{(1:n-1)}\|_2^2 + |g_{(n)} u_{(1)}|^2$, and thus

$$\frac{\|(\text{diag}(g) V_{k,n}^T u)_{(1:n-1)}\|_2^2}{\|\text{diag}(g) V_{k,n}^T u\|_2^2} = \left[1 + \frac{|g_{(n)} u_{(1)}|^2}{|\rho|^2 \|V_{k,n-1}^T u\|_2^2} \right]^{-1}$$

Therefore by Theorem 3.4

$$\begin{aligned}
\min_u \frac{\|(\text{diag}(g) V_{k,n}^T u)_{(1:n-1)}\|_2^2}{\|\text{diag}(g) V_{k,n}^T u\|_2^2} &= \left[1 + \max_u \frac{|g(n)u(1)|^2}{|\rho|^2 \|V_{k,n-1}^T u\|_2^2} \right]^{-1} \\
&= \left[1 + \frac{|g(n)|^2}{|\rho|^2} \frac{1}{\min_{|u(1)|=1} \|V_{k,n-1}^T u\|_2^2} \right]^{-1} \\
&= \left[1 + \frac{|g(n)|^2}{|\rho|^2} \frac{1}{2(n-1) \Xi_{\delta,k}} \right]^{-1},
\end{aligned}$$

as expected. ■

Theorem 6.4 *Let α_j 's as given in Lemma 6.2. Suppose A is a Hermitian matrix whose eigenvalues are $\lambda_j = \alpha_j$ and admits eigen-decomposition (3.1) and $g = U^*b$. Apply Lanczos algorithm on A with $q_1 = b$ as in (6.1). Set μ_k to be the k th Lanczos approximation to λ_n . Then*

$$\lambda_n - \mu_k \leq \Gamma \left[1 + \frac{|g(n)|^2}{\max_{1 \leq j \leq n-1} |g(j)|^2} \frac{1}{2(n-1)} \Xi_{\delta,k} \right]^{-1}, \quad (6.11)$$

$$\lambda_n - \mu_k \geq \gamma \left[1 + \frac{|g(n)|^2}{\min_{1 \leq j \leq n-1} |g(j)|^2} \frac{1}{2(n-1)} \Xi_{\delta,k} \right]^{-1}, \quad (6.12)$$

where $\theta \equiv \theta_{1:n-1} = \frac{\pi}{2(n-1)}$, $\delta = \beta/\alpha$, and

$$\Gamma = |\alpha| \left[\frac{1 + \cos \theta}{2} + \delta \frac{1 - \cos \theta}{2} \right], \quad \gamma = |\alpha| \left[\frac{1 - \cos \theta}{2} + \delta \frac{1 + \cos \theta}{2} \right].$$

Proof: Let $\Omega = \lambda_n I - \Lambda$. It can be verified that $\gamma \leq \|\Omega\|_2 \leq \Gamma$. Now

$$\begin{aligned}
\frac{\|\Omega^{1/2} \text{diag}(g) V_{k,n}^T u\|_2^2}{\|\text{diag}(g) V_{k,n}^T u\|_2^2} &\geq \gamma \frac{\|(\text{diag}(g) V_{k,n}^T u)_{(1:n-1)}\|_2^2}{\|\text{diag}(g) V_{k,n}^T u\|_2^2} \\
&= \gamma \left[1 + \frac{|g(n)u(1)|^2}{\|(\text{diag}(g) V_{k,n}^T u)_{(1:n-1)}\|_2^2} \right]^{-1} \\
&\geq \gamma \left[1 + \frac{|g(n)u(1)|^2}{\min_{1 \leq j \leq n-1} |g(j)|^2 \|V_{k,n-1}^T u\|_2^2} \right]^{-1}, \\
\frac{\|\Omega^{1/2} \text{diag}(g) V_{k,n}^T u\|_2^2}{\|\text{diag}(g) V_{k,n}^T u\|_2^2} &\leq \Gamma \left[1 + \frac{|g(n)u(1)|^2}{\max_{1 \leq j \leq n-1} |g(j)|^2 \|V_{k,n-1}^T u\|_2^2} \right]^{-1},
\end{aligned}$$

The rest of the proof for (6.12) is the same as that for Lemma 6.2 ■

Theorem 6.4 shows that the existing result (6.9) tells the correct speed of convergence for λ_n on the matrix. Kaniel and Saad obtained similar bounds on approximating other λ_j by Ritz values [16]. Their sharpness in general remains to be studied.

REMARK 6.2 For Theorems 6.2 and 6.4, one implies the other with slightly weakened inequalities in the theorems. In fact we have

$$(\lambda_n - \lambda_{n-1})^2 \epsilon^2 \leq \lambda_n - \mu_k \leq \frac{(\lambda_n - \lambda_1)^2}{\lambda_n - \lambda_{n-1}} \epsilon^2. \quad (6.13)$$

where $\epsilon = \sin \angle(U_{(:,n)}, \mathcal{K}_k)$. To see this, let $\|U_{(:,n)} - y\|_2 = \min_{x \in \mathcal{K}_k} \|U_{(:,n)} - x\|_2 = \epsilon$. Then $U_{(:,n)} - y$ and y are perpendicular and $\|y\|_2 = \sqrt{1 - \epsilon^2}$. Write $y = \xi U_{(:,n)} + U_{(:,1:n-1)}c$. It can be seen that $|\xi| = 1 - \epsilon^2$ and $\|c\|_2 = \epsilon\sqrt{1 - \epsilon^2}$. Therefore

$$r \stackrel{\text{def}}{=} Ay - \lambda_n y = (A - \lambda_n I)U_{(:,1:n-1)}c$$

and gives $\|r\|_2 \leq (\lambda_n - \lambda_1)\epsilon\sqrt{1 - \epsilon^2}$, and thus [11]

$$\lambda_n - \mu_k \leq \frac{(\|r\|_2 / \|y\|_2)^2}{\lambda_n - \lambda_{n-1}} \leq \frac{(\lambda_n - \lambda_1)^2}{\lambda_n - \lambda_{n-1}} \epsilon^2,$$

the second inequality in (6.13). Let u be the corresponding Ritz vector to μ_k . By [12, Theorem 2.1],

$$\epsilon \leq \sin \angle(U_{(:,n)}, u) \leq \sqrt{\frac{\lambda_n - \mu_k}{\lambda_n - \lambda_{n-1}}}$$

which leads to the first inequality in (6.13).

7 Conclusions

It is often observed that the existing error bounds for solutions by CG (and MINRES) for positive definite linear systems and symmetric Lanczos algorithm for symmetric eigenvalue problems are very good in indicating the accuracy of the computed solutions for the first few iterations but after that the bounds overestimates the actual errors too much to be of much use. Is this always the case? Through studying minimization problem (3.4):

$$\min_{|u_{(1)}|=1} \|\text{diag}(g) V_{k,n}^T u\|_2$$

which also leads to solutions to related minimization problems (6.3) and (6.8), we have devised examples for CG (and MINRES) and symmetric Lanczos algorithm by which the computed solutions have errors that are comparable to the existing error bounds at all iteration steps. This implies that the existing bounds can not be improved in general unless further information upon the problems to be solved becomes available.

For symmetric Lanczos algorithm, we only have an example that shows the existing bounds for the approximations to the largest (smallest) eigenvalue and its associated eigenvectors are sharp, modulo modest factors. The situation for approximations to any other eigenvalues and associated vectors can be very complicated and we suspect that the existing bounds would probably not sharp, even after modulo modest constant factors.

Concerning the sharpness of the error bound for CG, Meinardus [14] who established the bound himself showed the error bound could actually be achieved but he only did so for¹

¹This corresponds to $k = n$ in (3.4).

$k = n - 1$ in (1.3), leaving out what might happen for $1 \leq k \leq n - 2$ unanswered. Meinardus's example corresponds to V_n with nodes being the extreme points of a translated Chebyshev polynomial. Numerical tests show Meinardus's example does not achieve the error bound for $k \neq n - 1$, but has approximation errors comparable to the bound for all other k , nonetheless. We have a proof for this, but it is too complicated to fit in here and so we decide to publish the proof along with other results elsewhere. In the meantime, our results are sufficient to accomplish the goal we set out to achieve at the beginning of this paper.

The foundation of this paper is built upon an explicit evaluation of minimization problem (3.4) for translated Chebyshev zero nodes. This successful evaluation turns out not just for translated Chebyshev zero nodes. It can be done for any V_n whose nodes are the zero of the n th translated orthogonal polynomial from any orthogonal polynomial system because of the existence of a QR-like decomposition like (2.13) as the result of Gaussian quadrature formula. An immediate implication of this is that one can construct various examples for which CG residuals and errors in approximations by Lanczos algorithm can be expressed explicitly in terms of the constant terms of the associated translated orthogonal polynomials. Results as such along with others will be reported elsewhere, too.

References

- [1] B. BECKERMANN, *The condition number of real Vandermonde, Krylov and positive definite Hankel matrices*, Numerische Mathematik, 85 (2000), pp. 553–577.
- [2] P. BORWEIN AND T. ERDÉLYI, *Polynomials and Polynomial Inequalities*, vol. 161 of Graduate Texts in Mathematics, Springer, New York, 1995.
- [3] J. DEMMEL, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.
- [4] T. A. DRISCOLL, K.-C. TOH, AND L. N. TREFETHEN, *From potential theory to matrix iterations in six steps*, SIAM Review, 40 (1998), pp. 547–578.
- [5] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore, Maryland, 3rd ed., 1996.
- [6] A. GREENBAUM, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.
- [7] M. HANKE, *Superlinear convergence rates for the Lanczos method applied to elliptic operators*, Numerische Mathematik, 77 (1997), pp. 487–499.
- [8] Z. JIA AND G. W. STEWART, *An analysis of the rayleigh-ritz method for approximating eigenspaces*, Mathematics of Computation, 70 (2001), pp. 637–647.
- [9] S. KANIEL, *Estimates for some computational techniques in linear algebra*, Mathematics of Computation, 20 (1966), pp. 369–378.
- [10] A. B. J. KUIJLAARS, *Which eigenvalues are found by the Lanczos method?*, SIAM Journal on Matrix Analysis and Applications, 22 (2000), pp. 306–321.
- [11] C.-K. LI AND R.-C. LI, *A note on eigenvalues of perturbed Hermitian matrices*, Linear Algebra and its Application, 395 (2005), pp. 183–190.
- [12] R.-C. LI, *Accuracy of computed eigenvectors via optimizing a rayleigh quotient*, BIT, 44 (2004), pp. 585–593.
- [13] ———, *Asymptotically optimal lower bounds for the condition number of a real Vandermonde matrix*, Technical Report 2004-05, Department of Mathematics, University of Kentucky, 2004. Available at <http://www.ms.uky.edu/~math/MAreport/PDF/rc-05.pdf>.

- [14] G. MEINARDUS, *Über eine Verallgemeinerung einer Ungleichung von L. V. Kantorowitsch*, Numerische Mathematik, 5 (1963), pp. 14–23.
- [15] C. C. PAIGE AND M. A. SAUNDERS, *Solution of sparse indefinite systems of linear equations*, SIAM Journal on Numerical Analysis, 12 (1975), pp. 617–629.
- [16] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, 1998. This SIAM edition is an unabridged, corrected reproduction of the work first published by Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1980.
- [17] Y. SAAD, *On the rates of convergence of the Lanczos and the block-Lanczos methods*, SIAM Journal on Numerical Analysis, 15 (1980), pp. 687–706.
- [18] Y. SAAD, *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, 2nd ed., 2003.
- [19] Y. SAAD AND M. SCHULTZ, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM Journal on Scientific and Statistical Computing, 7 (1986), pp. 856–869.
- [20] G. L. G. SLEIJPEN AND A. VAN DER SLUIS, *Further results on the convergence behavior of conjugate-gradients and Ritz values*, Linear Algebra and its Applications, 246 (1996), pp. 233–378.
- [21] G. W. STEWART AND J.-G. SUN, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.
- [22] A. VAN DER SLUIS AND H. A. VAN DER VORST, *The convergence behavior of Ritz values in the presence of close eigenvalues*, Linear Algebra and its Applications, 88/89 (1987), pp. 651–694.