MATHEMATICS OF COMPUTATION
Volume 79, Number 269, January 2010, Pages 419–435
S 0025-5718(09)02258-3
Article electronically published on May 14, 2009

SHARPNESS IN RATES OF CONVERGENCE FOR THE SYMMETRIC LANCZOS METHOD

REN-CANG LI

Abstract. The Lanczos method is often used to solve a large and sparse symmetric matrix eigenvalue problem. There is a well-established convergence theory that produces bounds to predict the rates of convergence good for a few extreme eigenpairs. These bounds suggest at least linear convergence in terms of the number of Lanczos steps, assuming there are gaps between individual eigenvalues. In practice, often superlinear convergence is observed. The question is "*do the existing bounds tell the correct convergence rate in general?*". An affirmative answer is given here for the two extreme eigenvalues by examples whose Lanczos approximations have errors comparable to the error bounds for all Lanczos steps.

1. Introduction

The Lanczos method is widely used for finding a small number of eigenvalues and their associated eigenvectors of a large symmetric matrix because it requires only matrix-vector products to extract enough information to compute desired solutions. There is a well-established convergence theory to go with the method in terms of error bounds indicating how fast the computed approximations converge to a few extreme eigenpairs. These bounds usually underestimate the rate of convergence, however. In practice, often the observed convergence is (much) faster than these error bounds suggest [7, 24, 27]. This paper investigates the attainability of these bounds in general.

However, in finite precision without full orthogonalization, the Lanczos method can behave very differently from what it is supposed to be in theory [4, 6]. Nonetheless the existing theoretic bounds which assume exact arithmetic are still very suggestive as to what we may expect numerically. In this paper we assume exact arithmetic.

By default, all vectors are column vectors. Given an $N \times N$ Hermitian matrix A and a vector b of dimension N, the Lanczos process [20] may be compactly described as follows:

$$AX_k = X_k H_k + f_k e_k^{\mathrm{T}}, \tag{1.1}$$

Received by the editor July 31, 2006 and, in revised form, January 14, 2008 and January 6, 2009.

2000 *Mathematics Subject Classification.* Primary 65F10.

Key words and phrases. Lanczos method, Krylov subspace, rate of convergence, Chebyshev polynomial.

This work was supported in part by the National Science Foundation under Grant No. DMS-0702335 and DMS-0810506.

©2009 American Mathematical Society

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

where X_k is $N \times k$ and has orthonormal columns with its first column being a scalar multiple of b, f_k (a vector of dimension N) satisfies $X_k^* f_k = 0$, e_k is a k-vector with all entries zero except its last entry which is 1, and H_k is $k \times k$, real symmetric, and tridiagonal. The eigenvalue problem for H_k is then solved. Let (μ, z) be an eigenpair of H_k, i.e., $H_k z = \mu z$. An approximate eigenpair $(\mu, X_k z)$, so-called a *Ritz pair* of the *Ritz value* μ and its associated *Ritz vector* $X_k z$, is obtained for A. Without loss of generality, we shall consider only the case when H_k is irreducible; namely, none of its off-diagonal entries is zero. It can be seen that

$$\text{the column space of } X_k = \mathcal{K}_k(A, b) \stackrel{\text{def}}{=} \mathsf{span}\{b, Ab, \ldots, A^{k-1}b\}, \tag{1.2}$$

the kth *Krylov subspace* of A on b. Often we write $\mathcal{K}_k \equiv \mathcal{K}_k(A, b)$ for short when A and b are evident from the context. Assume that A admits the following eigendecomposition:

$$A = Q\Lambda Q^*,\ Q^*Q = I_N,\ \Lambda = \mathsf{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N). \tag{1.3}$$

Then Q's jth column $Q_{(:,j)}$ is the eigenvector of A associated with the eigenvalue λ_j. For the sake of presentation, assume

$$\lambda_1 \le \lambda_2 \le \cdots \le \lambda_N. \tag{1.4}$$

Naturally, we ask how well does an eigenvalue of H_k approximate A's eigenvalue, and how far is $Q_{(:,j)}$ from $\mathcal{K}_k(A, b)$. A well-developed theory for this is due to Kaniel [10] and Saad [21], and if more detailed information on A's eigenvalue distribution is available, better bounds can be derived, too [20].

Consider $A = \mathsf{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ with either randomly or equidistantly distributed $\{\lambda_j\}_{j=1}^{N-1}$ on $[-1, -\delta]$ and $\lambda_N = 0$ whose associated eigenvector is e_N. If the Lanczos algorithm is applied to A on a vector b of all ones, Figure 1.1 plots

$$\frac{\sqrt{N-1}}{\sqrt{(N-1) + |T_{k-1}(\delta_N)|^2}} \Big/ \sin\angle(e_N, \mathcal{K}_k), \tag{1.5}$$

which is the ratio of a bound due to Kaniel [10] and Saad [21] for this case (see Remark 3.2) over the actual sine of the angle $\angle(e_N, \mathcal{K}_k)$, where

$$\kappa_N = \frac{\lambda_N - \lambda_1}{\lambda_N - \lambda_{N-1}}, \quad \delta_N = \frac{\kappa_N + 1}{\kappa_N - 1},$$

and T_{N-1} is the $(N-1)$st Chebyshev polynomial of the first kind. This figure indicates that the bound of Kaniel and Saad can dramatically overestimate the actual sine of the angle as k varies.

To the best of my knowledge, there is no study in the past regarding the sharpness of the existing error bounds for the symmetric Lanczos method. Perhaps this is due in part to the fact that these bounds were established with a technique basically the same as the one for obtaining the error bounds for the conjugate gradient method (CG) [3, 5, 22, 24, 26]. The latter were argued to be (locally) sharp [1, 5] and more recently (globally) sharp[1] [16, 18]. Consequently the existing error bounds

[1]This concept of local and global sharpness is coined by [18] based on the consideration that the sharpness claim in, e.g., [1, 5], is in fact in the sense that for each iteration step k there is a linear system $Ax = b$ (depending on k) on which the kth CG residual attains the bound. It turns out that for such $Ax = b$, CG computes the exact solution in the very next iteration [10]! In [16, 18], however, it is demonstrated that there are linear systems $Ax = b$ on which CG residuals are comparable to the existing bounds for all iteration steps.

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

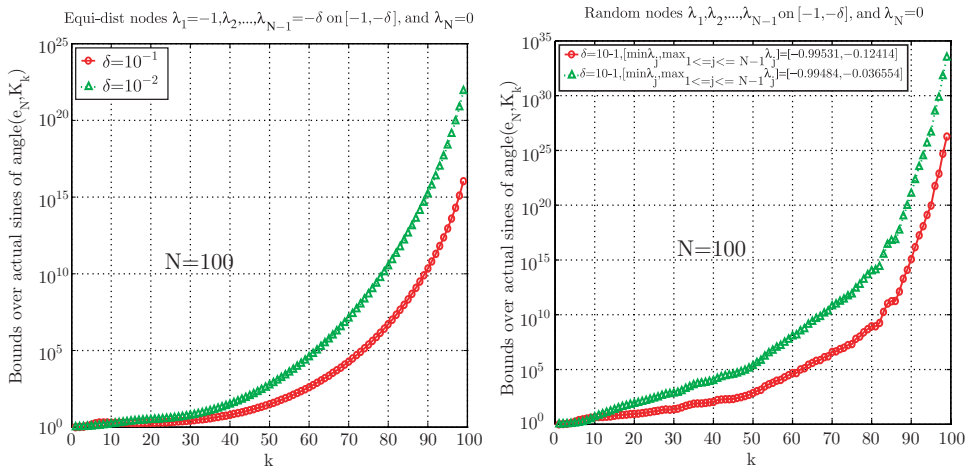

FIGURE 1.1. The Lanczos Algorithm for $Ax = \lambda x$ with $A = \mathsf{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ on b, the vector of all ones, where $\{\lambda_j\}_{j=1}^{N-1}$ is either equidistantly (*left plot*) or randomly (*right plot*) distributed.

for the Lanczos method could also be sharp at least for the first Ritz value, thanks to Sleijpen and van der Sluis [24, Theorems 6.1 and 6.2].

The main contribution of this paper is to show that the existing error bounds for the Lanczos method *are* indeed sharp in general, despite Figure 1.1. The same conclusion was also reached in the unpublished technical report [16], where examples were constructed with the Chebyshev zero nodes. Here with the help of the Chebyshev extreme nodes, we are able to devise examples for which the existing error bounds are much closer to the actual sines.

This paper strives to produce difficult problems for the Lanczos method, but it does so from a different perspective from Scott [23], where efforts were made to select a perverse starting vector b to delay the convergence until the last step. Since the theory of Kaniel and Saad guarantees fast and noticeable convergence provided the starting vector has a nontrivial component in the direction of the eigenvectors associated with the extreme eigenvalues which also have nontrivial gaps from the rest of the eigenvalues, any perverse starting vector of Scott's choice must have a negligible component in the direction of the desired eigenvectors. On the contrary, this paper and [16] assume the nontrivial components in all eigenvector directions and seek certain eigenvalue distributions so as to almost achieve the existing error bounds.

It is worth mentioning that in the potential-theoretic approach, Kuijlaars [12, 13] studied which eigenvalues are found first given how A's eigenvalues are distributed as $N \to \infty$, and what are their associated convergence rates as k goes to ∞ while k/N stays fixed.

The rest of this paper is organized as follows. Section 2 presents some preliminary material that will be used frequently later. Section 3 investigates the sharpness of the existing error bounds for eigenvectors associated with the two extreme eigenvalues, while Section 4 is concerned with eigenvalues. Finally, concluding remarks are given in Section 5.

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

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 the jth column of an identity matrix with a compatible dimension in the context. The superscript * denotes conjugate transpose while $^{\mathrm{T}}$ denotes transpose only. We shall also adopt a 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 a vector u and a matrix X, $u_{(j)}$ is u's jth entry, $X_{(i,j)}$ is X's (i,j)th entry, $\mathsf{diag}(u)$ is the diagonal matrix with $(\mathsf{diag}(u))_{(j,j)}=u_{(j)}$; X's submatrices $X_{(k:\ell,i:j)}$, $X_{(k:\ell,:)}$, and $X_{(:,i:j)}$ consist 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. The generic norm $\|\cdot\|_2$ is the usual ℓ_2 norm of a vector or the spectral norm of a matrix.

2. Preliminaries

The mth Chebyshev polynomial of the 1st kind is

$$T_m(t)=\cos(m\arccos t) \qquad \text{for } |t|\le 1, \tag{2.1}$$

$$=\frac{1}{2}\left(t+\sqrt{t^2-1}\right)^m+\frac{1}{2}\left(t-\sqrt{t^2-1}\right)^m \qquad \text{for } |t|\ge 1. \tag{2.2}$$

It frequently shows up in numerical analysis and computations because of its numerous nice properties; for example, $|T_m(t)|\le 1$ for $|t|\le 1$ and $|T_m(t)|$ grows extremely fast[2] for $|t|>1$. We will also need [18]

$$\left|T_m\left(\frac{1+t}{1-t}\right)\right|=\left|T_m\left(\frac{t+1}{t-1}\right)\right|=\frac{1}{2}\left[\Delta_t^m+\Delta_t^{-m}\right] \quad \text{for } 1\ne t>0, \tag{2.3}$$

where

$$\Delta_t\stackrel{\text{def}}{=}\frac{\sqrt{t}+1}{|\sqrt{t}-1|} \quad \text{for } t>0. \tag{2.4}$$

$T_m(t)$ has $m+1$ extreme points in $[-1,1]$, the so-called *mth Chebyshev extreme nodes*:

$$\tau_{jm}=\cos\vartheta_{jm},\ \vartheta_{jm}=\frac{j}{m}\pi,\ 0\le j\le m, \tag{2.5}$$

at which $|T_m(\tau_{jm})|=1$. Given $\alpha<\beta$, set

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

Throughout the rest of this paper, ω and τ are always defined this way when the interval $[\alpha,\beta]$ is specified; otherwise they can be any two numbers. The linear transformation

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

[2]In fact, a result due to Chebyshev himself says that if $p(t)$ is a polynomial of degree no greater than m and $|p(t)|\le 1$ for $-1\le t\le 1$, then $|p(t)|\le|T_m(t)|$ for any t outside $[-1,1]$ [2, p.65].

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

maps $z \in [\alpha, \beta]$ one-to-one and onto $t \in [-1, 1]$. With its inverse transformation $x(t) = \omega(t - \tau)$, we define the so-called *mth translated Chebyshev extreme nodes* on $[\alpha, \beta]$ as

$$\tau_{jm}^{\mathsf{tr}} = \omega(\tau_{jm} - \tau),\ 0 \le j \le m. \tag{2.8}$$

It can be verified that $\tau_{0m} = \beta$ and $\tau_{mm} = \alpha$.

3. Eigenvector convergence

Let us look at how close $Q_{(:,j)}$ is to $\mathcal{K}_k(A, b)$. It can be seen that

$$\sin \angle(Q_{(:,j)}, \mathcal{K}_k) = \min_{x \in \mathcal{K}_k} \|Q_{(:,j)} - x\|_2. \tag{3.1}$$

Given any number ν, this can be turned into the following minimization problem:

$$\begin{aligned} \min_{x \in \mathcal{K}_k} \|Q_{(:,j)} - x\|_2 &= \min_{\phi_{k-1}} \|Q_{(:,j)} - \phi_{k-1}(A)b\|_2 \\ &= \min_{\psi_{k-1}} \|Q_{(:,j)} - \psi_{k-1}(A - \nu I)b\|_2 \\ &= \min_{\psi_{k-1}} \|e_j - \psi_{k-1}(\Lambda - \nu I)Q^* b\|_2 \\ &= \min_{u_{(1)}=1} \|(e_j \quad \mathsf{diag}(g)V_{k,N}^{\mathrm{T}})u\|_2, \end{aligned} \tag{3.2}$$

where ϕ_{k-1} and ψ_{k-1} denote polynomials of degree at most $k-1$, $u \in \mathbb{C}^{k+1}$ with its first entry $u_{(1)}$ forced to be 1 always,

$$g = Q^* b, \tag{3.3}$$

and

$$V_{k,N} \stackrel{\text{def}}{=} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_N \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^{k-1} & \alpha_2^{k-1} & \cdots & \alpha_N^{k-1} \end{pmatrix}, \tag{3.4}$$

a $k \times N$ rectangular Vandermonde matrix with $\alpha_i = \lambda_i - \nu$ $(1 \le i \le N)$. That $Q_{(:,j)}$ is close to $\mathcal{K}_k(A, b)$ as measured by $\sin \angle(Q_{(:,j)}, \mathcal{K}_k)$ does not necessarily imply that there is a Ritz vector that approximates $Q_{(:,j)}$ well. For that the reader is referred to [9], where it is proved that, under suitable separation conditions, if $Q_{(:,j)}$ is close to $\mathcal{K}_k(A, b)$, then there is a Ritz vector that approximates $Q_{(:,j)}$ well.

For $k \ge N$, $\mathcal{K}_k$ is either the entire space $\mathbb{C}^N$ $(\mathbb{R}^N)$ or is A's invariant subspace. The former case implies

$$0 = \min_{x \in \mathcal{K}_k} \|Q_{(:,j)} - x\|_2 = \min_{u_{(1)}=1} \|(e_j \quad \mathsf{diag}(g)V_{k,N}^{\mathrm{T}})u\|_2,$$

and the latter case implies

$$\min_{x \in \mathcal{K}_k} \|Q_{(:,j)} - x\|_2 = \begin{cases} 0, & \text{if } g_{(j)} \ne 0, \\ 1, & \text{if } g_{(j)} = 1. \end{cases}$$

So it suffices to restrict $1 \le k \le N - 1$ from now on.

Equation (3.2) points to a new direction to analyze the convergence behavior of the Lanczos algorithm, i.e., by studying the minimization problem on its right-hand side. This will be the approach we will take from now on. Inequality (3.5) in the next theorem turns out to be equivalent to an existing bound but expressed differently, as explained in Remark 3.2. We present it here for completeness. The

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

sharpness of the inequality will be investigated afterwards (only for the case of $j = N$).

Theorem 3.1. *Let the Hermitian matrix A have eigendecomposition* (1.3) *with $\{\lambda_i\}_{i=0}^N$ ordered as in* (1.4)*, and let $\mathcal{K}_k \equiv \mathcal{K}_k(A, b)$ as in* (1.2) *and $g = Q^*b$ as in* (3.3)*. Then for $1 \le k \le N-1$,*

$$\sin\angle(Q_{(:,j)}, \mathcal{K}_k) \le \frac{\epsilon_j}{\sqrt{1+\epsilon_j^2}}, \tag{3.5}$$

where (set $\gamma_N = \varsigma_N = 1$ by convention $\prod_{j=N+1}^N(\cdots) \equiv 1$)

$$\kappa_j = \frac{\lambda_j - \lambda_1}{\lambda_j - \lambda_{j-1}},\ \delta_j = \frac{\kappa_j + 1}{\kappa_j - 1},\ \gamma_j = \prod_{i=j+1}^{N}(\lambda_i - \lambda_1),\ \varsigma_j = \prod_{i=j+1}^{N}(\lambda_i - \lambda_j),$$

$$\chi_j = \frac{\gamma_j}{\varsigma_j} \cdot \frac{\|g_{(1:j-1)}\|_2}{|g_{(j)}|},\ \epsilon_j = \frac{\chi_j}{|T_{k-1-(N-j)}(\delta_j)|}.$$

Proof. It suffices to bound the right-hand side of (3.2) with $\nu = 0$. For ω and τ in (2.6) with $[\alpha, \beta] = [\lambda_1, \lambda_{j-1}]$, $|T_{k-1-(N-j)}(\lambda_i/\omega + \tau)| \le 1$ for $1 \le i \le j-1$. Let $v \in \mathbb{C}^{k+1}$ with $v_{(1)} = 1$ and $v_{(i)} = \xi\, c_{i-2}$ for $2 \le i \le k+1$, where the c_i are coefficients of t^i in

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

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

$$\begin{aligned}\min_{u_{(1)}=1} \|(e_j \quad \mathsf{diag}(g)V_{k,N}^{\mathrm{T}})u\|_2 &\le \|(e_j \quad \mathsf{diag}(g)V_{k,N}^{\mathrm{T}})v\|_2 \\ &\le \left[|\xi|^2\gamma_j^2\|g_{(1:j-1)}\|_2^2 + (1 - |g_{(j)}\xi\,\zeta|)^2\right]^{1/2}.\end{aligned}$$

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

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

and

$$\min_{u_{(1)}=1} \|(e_j \quad \mathsf{diag}(g)V_{k,N}^{\mathrm{T}})u\|_2 \le \frac{\gamma_j\|g_{(1:j-1)}\|_2}{\sqrt{\gamma_j^2\|g_{(1:j-1)}\|_2^2 + |g_{(j)}\zeta|^2}}. \tag{3.6}$$

Now by (2.6),

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

and thus $|\zeta| = \varsigma_j\, |T_{k-1-(N-j)}(\delta_j)|$, and we have (3.5). □

Remark 3.1. It is known that the eigenvalues at both ends are often the first few to emerge from an application of the Lanczos method. But this is not reflected by Theorem 3.1 because for small j and huge N, γ_j and ς_j not only complicates the bound by (3.5) but also may significantly offset the effectiveness of $|T_{k-1-(N-j)}(\delta_j)|$. A remedy for generating better bounds for small j is by applying the theorem to $-A$ upon noticing $\mathcal{K}_k(A, b) \equiv \mathcal{K}_k(-A, b)$. Thus any conclusion on approximating

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

the largest eigenvalues by the Lanczos algorithm has a counterpart for the smallest eigenvalues. Owing to this property, we in this paper will focus only on approximating the largest eigenvalues and their associated eigenvectors.

Remark 3.2. Inequality (3.5) is equivalent to an existing bound of Kaniel and Saad [20, p. 270] because $\sin\theta \le \epsilon/\sqrt{1+\epsilon^2}$ is equivalent to $\tan\theta \le \epsilon$ for $0 \le \theta < \pi/2$.

In the rest of this section, we will investigate the attainability of the bound by (3.5) for $j = N$ only. Before we present our main result, Theorem 3.2, for the section, we shall introduce some notation and establish two lemmas. Throughout the rest of this paper, we set

$$n = N - 1, \tag{3.7}$$

and define

$$\Psi_{t,k} = \begin{cases} \sum_{i=0}^{k-1}{}' \, |T_i(t)|^2, & \text{for } 1 \le k \le n-1, \\ \sum_{i=0}^{k-1}{}'' \, |T_i(t)|^2, & \text{for } k = n, \end{cases} \tag{3.8}$$

where $\sum_i'$ means the first term is halved, while for $\sum_i''$ both the first and last terms are halved.

In its present general form, the next lemma was proved in [16, 18]. It was also implied by the proof of [8, Theorem 2.1]. See also [19].

Lemma 3.1. *If Z has full column rank, then*

$$\min_{|u_{(1)}|=1} \|Zu\|_2 = \left[e_1^{\mathrm{T}}(Z^*Z)^{-1}e_1\right]^{-1/2}. \tag{3.9}$$

Lemma 3.2. *Let $\alpha < \beta < 0$, and let $V_{k,N}$ have nodes $\alpha_{i+1} = \tau^{\mathsf{tr}}_{i\,n-1}$ $(0 \le i \le n-1)$ on $[\alpha, \beta]$, and $\alpha_N = 0$. $\sigma \in \mathbb{C}$ is a given number and nonzero.*

(1) *For*

$$g_{(i)} \overset{\text{def}}{=} \begin{cases} \sigma/\sqrt{2}, & \text{for } i \in \{1, n\}, \\ \sigma, & \text{for } 2 \le i \le n-1, \end{cases} \tag{3.10}$$

we have for $1 \le k \le n$,

$$\min_{|u_{(1)}|=1} \left\|(e_N \quad \mathsf{diag}(g)V_{k,N}^{\mathrm{T}})u\right\|_2 = \left[1 + \frac{|g_{(N)}|^2}{\|g_{(1:n)}\|_2^2}\, 2\Psi_{\tau,k}\right]^{-1/2}, \tag{3.11}$$

where $\tau = -(\alpha+\beta)/(\beta-\alpha)$ is given by (2.6).

(2) *For*

$$g_{(i)} \overset{\text{def}}{=} \begin{cases} \sigma\sqrt{1/|\tau^{\mathsf{tr}}_{i-1\,n-1}|}, & \text{for } i \in \{1, n\}, \\ \sigma\sqrt{2/|\tau^{\mathsf{tr}}_{i-1\,n-1}|}, & \text{for } 2 \le i \le n-1, \end{cases} \tag{3.12}$$

we have for $1 \le k \le n$,

$$\min_{|u_{(1)}|=1} \left\|(e_N \quad \mathsf{diag}(g)V_{k,N}^{\mathrm{T}})u\right\|_2 = \left[1 + \frac{|g_{(N)}|^2}{\|g_{(1:n)}\|_2^2}\, \rho_{k-1}^{-2}\, |T_{k-1}(\tau)|^2\right]^{-1/2}, \tag{3.13}$$

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

where $\kappa = \alpha/\beta > 1$, *and*

$$\frac{1}{2} < \frac{1}{2}\left(1 + \frac{2\Delta_\kappa^{n-1}}{\Delta_\kappa^{2(n-1)}+1}\right) \le \rho_{k-1}^2 = \frac{1}{2}\left(1 + \frac{\Delta_\kappa^{2(k-1)} + \Delta_\kappa^{2[(n-1)-(k-1)]}}{\Delta_\kappa^{2(n-1)}+1}\right) \le 1. \tag{3.14}$$

Proof. Set

$$Z \equiv \begin{pmatrix} e_N & \mathsf{diag}(g)V_{k,N}^{\mathrm{T}}\end{pmatrix} = \begin{pmatrix} 0 & \mathsf{diag}(g_{(1:n)})V_{k,n}^{\mathrm{T}} \\ 1 & g_{(N)}\, e_1^{\mathrm{T}}\end{pmatrix},$$

since $\alpha_N = 0$. It can be seen that Z has full column rank if $g_{(i)} \neq 0$ for $1 \le i \le n$. Then by Lemma 3.1, we need to compute $\left[e_1^{\mathrm{T}}(Z^*Z)^{-1}e_1\right]^{-1/2}$. We have

$$\begin{aligned} Z^*Z &= \begin{pmatrix} 1 & g_{(N)}\, e_1^{\mathrm{T}} \\ g_{(N)}^* e_1 & |g_{(N)}|^2 e_1 e_1^{\mathrm{T}} + \bar{V}_{k,n} G_n V_{k,n}^{\mathrm{T}}\end{pmatrix} \\ &= \begin{pmatrix} 1 & \\ g_{(N)}^* e_1 & I_n\end{pmatrix}\begin{pmatrix} 1 & \\ & \bar{V}_{k,n} G_n V_{k,n}^{\mathrm{T}}\end{pmatrix}\begin{pmatrix} 1 & g_{(N)}\, e_1^{\mathrm{T}} \\ & I_n\end{pmatrix}, \end{aligned}$$

where $\bar{V}_{k,n}$ is the complex conjugate of $V_{k,n}$ and $G_n = [\mathsf{diag}(g_{(1:n)})]^*\mathsf{diag}(g_{(1:n)})$. Therefore

$$\begin{aligned} e_1^{\mathrm{T}}(Z^*Z)^{-1}e_1 &= e_1^{\mathrm{T}}\begin{pmatrix} 1 & -g_{(N)}\, e_1^{\mathrm{T}} \\ & I\end{pmatrix}\begin{pmatrix} 1 & \\ & \left[V_{k,n}G_nV_{k,n}^{\mathrm{T}}\right]^{-1}\end{pmatrix}\begin{pmatrix} 1 & \\ -g_{(N)}^* e_1 & I\end{pmatrix} e_1 \\ &= 1 + |g_{(N)}|^2 e_1^{\mathrm{T}}\left[V_{k,n}G_nV_{k,n}^{\mathrm{T}}\right]^{-1} e_1. \end{aligned} \tag{3.15}$$

For g as in (3.10), $G_n = |\sigma|^2\,\mathsf{diag}(2^{-1}, 1, 1, \ldots, 1, 2^{-1})$, and Li [17, Theorem 5.3] yields that[3]

$$\left[e_1^{\mathrm{T}}\left(V_{k,n}G_nV_{k,n}^{\mathrm{T}}\right)^{-1} e_1\right]^{-1/2} = \|g_{(1:n)}\|_2\,(2\Psi_{\tau,k})^{-1/2}, \tag{3.16}$$

which, together with (3.15), leads to (3.11). For g as in (3.12),

$$G_n = |\sigma|^2\mathsf{diag}(1/|\tau_{0\,n-1}^{\mathsf{tr}}|, 2/|\tau_{1\,n-1}^{\mathsf{tr}}|, 2/|\tau_{2\,n-1}^{\mathsf{tr}}|, \ldots, 2/|\tau_{n-2\,n-1}^{\mathsf{tr}}|, 1/|\tau_{n-1\,n-1}^{\mathsf{tr}}|),$$

and Li [17, Theorem 5.4] yields that[4]

$$\left[e_1^{\mathrm{T}}\left(V_{k,n}G_nV_{k,n}^{\mathrm{T}}\right)^{-1} e_1\right]^{-1/2} = \|g_{(1:n)}\|_2\,\rho_{k-1}\,|T_{k-1}(\tau)|^{-1}, \tag{3.17}$$

which, together with (3.15), leads to (3.13). □

Theorem 3.2. *Suppose A is Hermitian with its first n eigenvalues $\{\lambda_j\}_{j=1}^n$ being the translated Chebyshev extreme nodes in $[\alpha, \beta] = [\lambda_1, \lambda_n]$ and its last eigenvalue $\lambda_N > \lambda_n$, and suppose A admits an eigendecomposition* (1.3) *and $g = Q^*b$. Let $\tau_{i\,n-1}^{\mathsf{tr}}$ $(0 \le i \le n-1)$ be the translated Chebyshev extreme nodes in*

$$[\alpha, \beta] = [\lambda_1 - \lambda_N, \lambda_n - \lambda_N].$$

[3]Li [17, Theorem 5.3] is really for $\sigma = 1$. But since $|\sigma|^2$ can be factored out in G_n, one still has (3.16). This comment also applies to (3.17).

[4]Li [17, Theorem 5.4] was stated for all $\tau_{i-1\,n-1}^{\mathsf{tr}} > 0$, but it is not hard to see that the theorem holds for all $\tau_{i-1\,n-1}^{\mathsf{tr}} < 0$ because $V_{k,n}$ with all nodes negative can be turned into one with all nodes positive by pre-multiplying the diagonal matrix $\mathsf{diag}(1, -1, 1, -1, \ldots)$.

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Apply the Lanczos algorithm with A on b as in (1.1). *We have*

(3.18)
$$\left[1+\frac{|g_{(N)}|^2}{(n-1)c_1^2}\,2\Psi_{\delta_N,k}\right]^{-1/2} \le \sin\angle(Q_{(:,N)},\mathcal{K}_k) \le \left[1+\frac{|g_{(N)}|^2}{(n-1)c_2^2}\,2\Psi_{\delta_N,k}\right]^{-1/2},$$

where δ_N is as in Theorem 3.1,

$$\begin{aligned} c_1 &= \min\left\{\sqrt{2}|g_{(1)}|, \sqrt{2}|g_{(n)}|, \min_{1<i<n}|g_{(i)}|\right\},\\ c_2 &= \max\left\{\sqrt{2}|g_{(1)}|, \sqrt{2}|g_{(n)}|, \max_{1<i<n}|g_{(i)}|\right\}. \end{aligned}$$

Also

(3.19)
$$\begin{aligned} &\left[1+\frac{|g_{(N)}|^2}{\zeta c_3^2}\,\rho_{k-1}^{-2}\,|T_{k-1}(\delta_N)|^2\right]^{-1/2}\\ &\quad\le \sin\angle(Q_{(:,N)},\mathcal{K}_k)\\ &\quad\le \left[1+\frac{|g_{(N)}|^2}{\zeta c_4^2}\,\rho_{k-1}^{-2}\,|T_{k-1}(\delta_N)|^2\right]^{-1/2}, \end{aligned}$$

where $\zeta=\sum''_{0\le i\le n-1} 2/|\tau^{\mathsf{tr}}_{i\,n-1}|$, ρ_{k-1} as in (3.14) *with $\kappa\equiv\kappa_N=(\lambda_N-\lambda_1)/(\lambda_N-\lambda_n)$, and*

$$\begin{aligned} c_3 &= \min\left\{\sqrt{|\tau^{\mathsf{tr}}_{0\,n-1}|}\,|g_{(1)}|,\ \sqrt{|\tau^{\mathsf{tr}}_{n-1\,n-1}|}\,|g_{(n)}|,\ 2^{-1/2}\min_{1<i<n}\sqrt{|\tau^{\mathsf{tr}}_{i-1\,n-1}|}\,|g_{(i)}|\right\},\\ c_4 &= \max\left\{\sqrt{|\tau^{\mathsf{tr}}_{0\,n-1}|}\,|g_{(1)}|,\ \sqrt{|\tau^{\mathsf{tr}}_{n-1\,n-1}|}\,|g_{(n)}|,\ 2^{-1/2}\max_{1<i<n}\sqrt{|\tau^{\mathsf{tr}}_{i-1\,n-1}|}\,|g_{(i)}|\right\}. \end{aligned}$$

Proof. In (3.2), take $\nu=\lambda_N$, and $V_{k,N}$ with nodes $\alpha_i=\lambda_i-\lambda_N$ $(1\le i\le N)$. Then

$$\sin\angle(Q_{(:,N)},\mathcal{K}_k)=\min_{u_{(1)}=1}\|(e_N \quad \mathsf{diag}(g)V_{k,N}^{\mathrm{T}})u\|_2.$$

Note that $\{\alpha_i\}_{i=1}^N$ are the same as the ones in Lemma 3.2 and $\tau=-(\beta+\alpha)/(\beta-\alpha)=\delta_N$. It can be seen that

(3.20)
$$\begin{aligned} \min_{u_{(1)}=1}\|(e_N \quad \mathsf{diag}(\widetilde{g})V_{k,N}^{\mathrm{T}})u\|_2 &\le \min_{u_{(1)}=1}\|(e_N \quad \mathsf{diag}(g)V_{k,N}^{\mathrm{T}})u\|_2\\ &\le \min_{u_{(1)}=1}\|(e_N \quad \mathsf{diag}(\widehat{g})V_{k,N}^{\mathrm{T}})u\|_2, \end{aligned}$$

where

(3.21)
$$\widetilde{g}_{(i)}=\begin{cases} c_1/\sqrt{2}, & \text{for } i\in\{1,n\},\\ c_1, & \text{for } 2\le i\le n-1,\end{cases} \quad \widehat{g}_{(i)}=\begin{cases} c_2/\sqrt{2}, & \text{for } i\in\{1,n\},\\ c_2, & \text{for } 2\le i\le n-1,\end{cases}$$

and $\widetilde{g}_{(N)}=\widehat{g}_{(N)}=g_{(N)}$. Item 1 of Lemma 3.2 and (3.20) give (3.18), upon noticing that $(n-1)c_1^2=\|\widetilde{g}_{(1:n)}\|_2^2$ and $(n-1)c_2^2=\|\widehat{g}_{(1:n)}\|_2^2$. The inequalities in (3.20) also

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

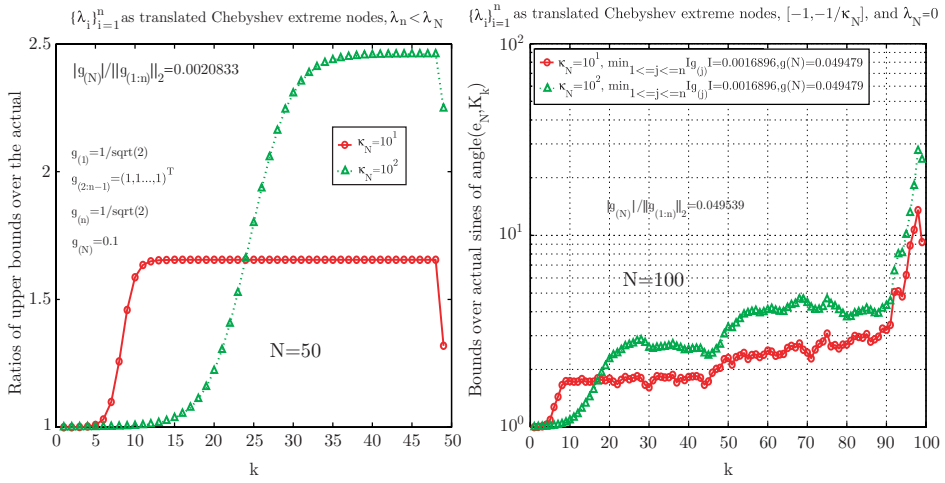

FIGURE 3.1. Ratios of the upper bound by (3.5) for $j = N$ over $\sin\angle(Q_{(:,N)}, \mathcal{K}_k)$. **Left:** $g = (2^{-1/2}, 1, 1, \ldots, 1, 2^{-1/2}, 0.1)^{\mathrm{T}}$; **Right:** g is random and $\|g\|_2 = 1$.

hold for

$$\widetilde{g}_{(i)} = \begin{cases} c_3 \sqrt{\dfrac{1}{|\tau^{\mathsf{tr}}_{i-1\,n-1}|}}, & \text{for } i \in \{1, n\}, \\ c_3 \sqrt{\dfrac{2}{|\tau^{\mathsf{tr}}_{i-1\,n-1}|}}, & \text{for } 2 \le i \le n-1, \end{cases}$$

$$\widehat{g}_{(i)} = \begin{cases} c_4 \sqrt{\dfrac{1}{|\tau^{\mathsf{tr}}_{i-1\,n-1}|}}, & \text{for } i \in \{1, n\}, \\ c_4 \sqrt{\dfrac{2}{|\tau^{\mathsf{tr}}_{i-1\,n-1}|}}, & \text{for } 2 \le i \le n-1, \end{cases} \tag{3.22}$$

and again $\widetilde{g}_{(N)} = \widehat{g}_{(N)} = g_{(N)}$. Item 2 of Lemma 3.2 and (3.20) give (3.19), upon noticing that $\zeta c_3^2 = \|\widetilde{g}_{(1:n)}\|_2^2$ and $\zeta c_4^2 = \|\widehat{g}_{(1:n)}\|_2^2$. □

Theorem 3.2 leads to two examples that can demonstrate that the existing bound

$$\sin\angle(Q_{(:,N)}, \mathcal{K}_k) \le \left[1 + \frac{|g_{(N)}|^2}{\|g_{(1:n)}\|_2^2}\, |T_{k-1}(\delta_N)|^2\right]^{-1/2} \tag{3.23}$$

by Theorem 3.1 for $j = N$ is rather sharp in general.

Example 3.1. Let A be as described in Theorem 3.2 such that $c_1 = c_2$; namely, g takes the form of (3.10). Then (3.18) becomes an equality

$$\sin\angle(Q_{(:,N)}, \mathcal{K}_k) = \left[1 + \frac{|g_{(N)}|^2}{\|g_{(1:n)}\|_2^2}\, 2\Psi_{\delta_N,k}\right]^{-1/2}. \tag{3.24}$$

To compare the right-hand side of (3.23) and that of (3.24), we notice that

$$|T_i(\delta_N)| = \frac{1}{2}\left[\Delta_{\kappa_N}^{i} + \Delta_{\kappa_N}^{-i}\right] \tag{3.25}$$

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

by (2.3). It can be seen that

$$1 \le \frac{\text{RHS of (3.23)}}{\text{RHS of (3.24)}} \le \sqrt{\frac{2\Psi_{\delta_N,k}}{|T_{k-1}(\delta_N)|^2}},$$

and at the same time if $|g_{(N)}| > 0$ and $\kappa_N > 1$,

$$\frac{\text{RHS of (3.23)}}{\text{RHS of (3.24)}} \sim \sqrt{\frac{2\Psi_{\delta_N,k}}{|T_{k-1}(\delta_N)|^2}}$$

as $N > k \to \infty$. Here and in what follows, the notation $a_{k,N} \sim b_{k,N}$ means $a_{k,N}/b_{k,N} \to 1$ as $N > k \to \infty$. Now for $1 < k \le n-1$, by (3.25) and writing Δ for Δ_{κ_N} for short,

$$\begin{aligned}\Psi_{\delta_N,k} &= \frac{1}{2} + \sum_{i=1}^{k-1} \frac{1}{4}\left[\Delta^{2i} + 2 + \Delta^{-2i}\right] \\ &= \frac{1}{4}\frac{\Delta^{2k}-1}{\Delta^2-1} + \frac{1}{2}(k-1) + \frac{1}{4}\frac{\Delta^{-2k}-1}{\Delta^{-2}-1},\end{aligned}$$

and for $k = n$,

$$\begin{aligned}\Psi_{\delta_N,n} &= \frac{1}{2} + \sum_{i=1}^{n-1} \frac{1}{4}\left[\Delta^{2i} + 2 + \Delta^{-2i}\right] - \frac{1}{8}\left[\Delta^{2(n-1)} + 2 + \Delta^{-2(n-1)}\right] \\ &= \frac{1}{4}\frac{\Delta^{2n}-1}{\Delta^2-1} + \frac{1}{2}(k-1) + \frac{1}{4}\frac{\Delta^{-2n}-1}{\Delta^{-2}-1} - \frac{1}{8}\left[\Delta^{2(n-1)} + 2 + \Delta^{-2(n-1)}\right].\end{aligned}$$

Therefore for $N > k \to \infty$,

(3.26)

$$\frac{\text{RHS of (3.23)}}{\text{RHS of (3.24)}} \sim \sqrt{\frac{2\Psi_{\delta_N,k}}{|T_{k-1}(\delta_N)|^2}} \sim \begin{cases} \sqrt{\frac{2\Delta^2}{\Delta^2-1}} = \frac{1+\sqrt{\kappa_N}}{\sqrt{2}\sqrt[4]{\kappa_N}}, & \text{for } k \le n-1, \\ \sqrt{\frac{\Delta^2+1}{\Delta^2-1}} = \frac{\sqrt{1+\kappa_N}}{\sqrt{2}\sqrt[4]{\kappa_N}}, & \text{for } k = n. \end{cases}$$

The left plot in Figure 3.1 is for the leftmost ratio for $\kappa_N = 10$ and 10^2 and $N = 50$. Our asymptotical analysis in (3.26) shows up in the plot:

$$\frac{1+\sqrt{\kappa_N}}{\sqrt{2}\sqrt[4]{\kappa_N}} = \begin{cases} 1.6551, & \text{for } \kappa_N = 10, \\ 2.4597, & \text{for } \kappa_N = 10^2, \end{cases} \qquad \frac{\sqrt{1+\kappa_N}}{\sqrt{2}\sqrt[4]{\kappa_N}} = \begin{cases} 1.3188, & \text{for } \kappa_N = 10, \\ 2.2472, & \text{for } \kappa_N = 10^2. \end{cases}$$

The right plot in Figure 3.1 is for a random unit vector g. It, too, indicates that the existing bound by (3.5) for $j = N$ is fairly tight.

Example 3.2. Let A be as described in Theorem 3.2 such that $c_3 = c_4$; namely, g takes the form of (3.12). Then (3.19) becomes an equality

$$\sin\angle(Q_{(:,N)}, \mathcal{K}_k) = \left[1 + \frac{|g_{(N)}|^2}{\|g_{(1:n)}\|_2^2}\,\rho_{k-1}^{-2}\,|T_{k-1}(\delta_N)|^2\right]^{-1/2}. \tag{3.27}$$

It can be seen that

$$1 \le \frac{\text{RHS of (3.23)}}{\text{RHS of (3.27)}} \le \rho_{k-1}^{-1},$$

and at the same time if $|g_{(N)}| > 0$ and $\kappa_N > 1$,

$$\frac{\text{RHS of (3.23)}}{\text{RHS of (3.27)}} \sim \rho_{k-1}^{-1}$$

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

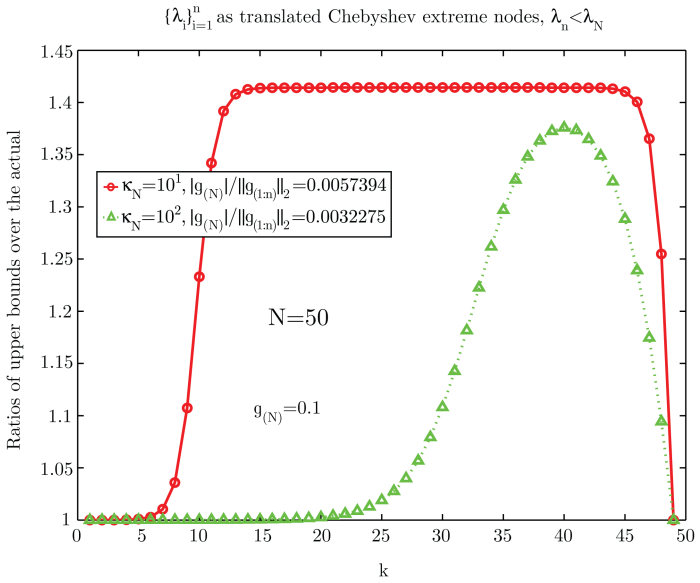

FIGURE 3.2. Ratios of the upper bound by (3.5) for $j = N$ over $\sin\angle(Q_{(:,N)}, \mathcal{K}_k)$ for g as in (3.12) with $\sigma = 1$.

as $N > k \to \infty$. But $1 \le \rho_{k-1}^{-1} \le \sqrt{2}$, which means the ratio is bounded uniformly by $\sqrt{2}$. Figure 3.2 plots the ratio with $\sigma = 1$ and $g_{(N)} = 0.1$.

4. EIGENVALUE CONVERGENCE

A is Hermitian; so is $H_k = X_k^* A X_k$. We expect the largest eigenvalue μ_k of H_k best approximates λ_N. We have

$$\begin{aligned}\mu_k &= \max_z \frac{z^* H_k z}{z^* z} = \max_z \frac{z^* X_k^* A X_k z}{z^* X_k^* X_k z} = \lambda_N + \max_{u\in\mathcal{K}_k} \frac{u^*(A-\lambda_N I)u}{u^* u} \\ &= \lambda_N + \max_{\phi_{k-1}} \frac{[\phi_{k-1}(A-\lambda_N I)b]^*(A-\lambda_N I)[\phi_{k-1}(A-\lambda_N I)b]}{[\phi_{k-1}(A-\lambda_N I)b]^*[\phi_{k-1}(A-\lambda_N I)b]},\end{aligned}$$

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

$$0 \ge \mu_k - \lambda_N = -\min_u \frac{\|\mathsf{diag}((\lambda_N I - \Lambda)^{1/2} g) V_{k,N}^{\mathrm{T}} u\|_2^2}{\|\mathsf{diag}(g) V_{k,N}^{\mathrm{T}} u\|_2^2}, \tag{4.1}$$

where $g = Q^* b$ as before, $\alpha_i = \lambda_i - \lambda_N$ $(1 \le i \le N)$ are the nodes for $V_{k,N}$, and u is the vector of coefficients of ϕ_{k-1}. Recall $n = N - 1$. For ω and τ in (2.6) with $[\alpha, \beta] = [\alpha_1, \alpha_n]$, $|T_{k-1}(\alpha_j/\omega + \tau)| \le 1$ for $1 \le i \le n$. Let $v \in \mathbb{C}^k$ with $v_{(i)}$ being

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

the coefficient of z^{i-1} in $T_{k-1}(z/\omega+\tau)$. We have

$$\begin{aligned}\min_u \frac{\|(\lambda_N I-\Lambda)^{1/2}\mathsf{diag}(g)\,V_{k,N}^{\mathrm{T}}u\|_2^2}{\|\mathsf{diag}(g)V_{k,N}^{\mathrm{T}}u\|_2^2} &\le \frac{\|(\lambda_N I-\Lambda)^{1/2}\mathsf{diag}(g)\,V_{k,N}^{\mathrm{T}}v\|_2^2}{\|\mathsf{diag}(g)V_{k,N}^{\mathrm{T}}v\|_2^2}\\ &\le (\lambda_N-\lambda_1)\frac{\|g_{(1:n)}\|_2^2}{|g_{(N)}v_{(1)}|^2}\\ &= (\lambda_N-\lambda_1)\frac{\|g_{(1:n)}\|_2^2}{|g_{(N)}|^2}\frac{1}{|T_{k-1}(\tau)|^2}. \qquad (4.2)\end{aligned}$$

Now by (2.6),

$$\tau=\frac{\alpha_1+\alpha_n}{\alpha_1-\alpha_n}=\frac{\kappa_N+1}{\kappa_N-1}=\delta_N,$$

and by (2.3) and (4.1), we have

$$0\le\lambda_N-\mu_k\le(\lambda_N-\lambda_1)\epsilon_N^2, \qquad (4.3)$$

where κ_N, δ_N, and ϵ_N are as defined in Theorem 3.1. Inequality (4.3) is an existing result of Kaniel and Saad [20]. Theorem 4.1 below presents a slightly sharper result due to an anonymous referee.

Theorem 4.1. *Let the Hermitian matrix A have the eigendecomposition* (1.3) *with $\{\lambda_i\}_{i=0}^N$ ordered as in* (1.4)*, and let $\mathcal{K}_k\equiv\mathcal{K}_k(A,b)$ as in* (1.2)*, $g=Q^*b$ as in* (3.3)*, $1\le k\le n$, and μ_k is the largest eigenvalue of H_k. Then*

$$0\le\lambda_N-\mu_k\le(\lambda_N-\lambda_1)\frac{\epsilon_N^2}{1+\epsilon_N^2}. \qquad (4.4)$$

Proof. Letting y_1 denote the closest unit vector in $\mathcal{K}_k$ to $Q_{(:,N)}$, apply Sun [25, (2.4)] to get

$$0\le\lambda_N-\mu_k\le\lambda_N-y_1^*Ay_1\le(\lambda_N-\lambda_1)\sin^2\angle(Q_{(:,N)},\mathcal{K}_k), \qquad (4.5)$$

which, combined with Theorem 3.1, leads to (4.4). □

We need the following lemma before presenting our main theorem in the section that shows the bound by (4.4) is very sharp.

Lemma 4.1. *Let $\alpha<\beta<0$, and let $V_{k,N}$ have nodes $\alpha_{i+1}=\tau^{\mathsf{tr}}_{i\,n-1}$ $(0\le i\le n-1)$ on $[\alpha,\beta]$, and $\alpha_N=0$.*

(1) *For g as in* (3.10)*,*

$$\min_u\frac{\|\mathsf{diag}(g_{(1:n)})\,V_{k,n}^{\mathrm{T}}u\|_2^2}{\|\mathsf{diag}(g)\,V_{k,N}^{\mathrm{T}}u\|_2^2}=\left[1+\frac{|g_{(N)}|^2}{\|g_{(1:n)}\|_2^2}\,2\Psi_{\tau,k}\right]^{-1}, \qquad (4.6)$$

where $\tau=-(\alpha+\beta)/(\beta-\alpha)$ is given by (2.6) *and $\Psi_{\tau,k}$ is defined as in* (3.8)*.*

(2) *For g as in* (3.12)*,*

$$\min_u\frac{\|\mathsf{diag}(g_{(1:n)})\,V_{k,n}^{\mathrm{T}}u\|_2^2}{\|\mathsf{diag}(g)\,V_{k,N}^{\mathrm{T}}u\|_2^2}=\left[1+\frac{|g_{(N)}|^2}{\|g_{(1:n)}\|_2^2}\,\rho_{k-1}^{-2}\,|T_{k-1}(\tau)|^2\right]^{-1}, \qquad (4.7)$$

where ρ_{k-1} is defined as in (3.14) *with $\kappa=\alpha/\beta>1$.*

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Proof. Notice that $V_{k,N} = (V_{k,n} \;\; e_1)$ to get

$$\|\mathsf{diag}(g)\, V_{k,N}^{\mathrm{T}} u\|_2^2 = \|\mathsf{diag}(g_{(1:n)})\, V_{k,n}^{\mathrm{T}} u\|_2^2 + |g_{(N)} u_{(1)}|^2,$$

and thus

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

Therefore

$$\begin{aligned} \min_u \frac{\|\mathsf{diag}(g_{(1:n)})\, V_{k,n}^{\mathrm{T}} u\|_2^2}{\|\mathsf{diag}(g)\, V_{k,N}^{\mathrm{T}} u\|_2^2} &= \left[1 + \max_u \frac{|g_{(N)} u_{(1)}|^2}{\|\mathsf{diag}(g_{(1:n)})\, V_{k,n}^{\mathrm{T}} u\|_2^2}\right]^{-1} \\ &= \left[1 + |g_{(N)}|^2 \frac{1}{\min_{|u_{(1)}|=1} \|\mathsf{diag}(g_{(1:n)})\, V_{k,n}^{\mathrm{T}} u\|_2^2}\right]^{-1}. \end{aligned} \tag{4.8}$$

For g as in (3.10), $\mathsf{diag}(g_{(1:n)}) = \sigma\, \mathsf{diag}(2^{-1/2}, 1, 1, \ldots, 1, 2^{-1/2})$, and Li [17, Theorem 5.3] yields that

$$\min_{|u_{(1)}|=1} \|\mathsf{diag}(g_{(1:n)})\, V_{k,n}^{\mathrm{T}} u\|_2^2 = (n-1)|\sigma|^2\, (2\Psi_{\tau,k})^{-1} = \|g_{(1:n)}\|_2^2\, (2\Psi_{\tau,k})^{-1},$$

which, together with (4.8), leads to (4.6). For g as in (3.12),

$$\mathsf{diag}(g_{(1:n)}) = \sigma\, \mathsf{diag}\left(\sqrt{\frac{1}{|\tau^{\mathsf{tr}}_{0\,n-1}|}}, \sqrt{\frac{2}{|\tau^{\mathsf{tr}}_{1\,n-1}|}}, \sqrt{\frac{2}{|\tau^{\mathsf{tr}}_{2\,n-1}|}}, \ldots, \sqrt{\frac{2}{|\tau^{\mathsf{tr}}_{n-2\,n-1}|}}, \sqrt{\frac{1}{|\tau^{\mathsf{tr}}_{n-1\,n-1}|}}\right)$$

and Li [17, Theorem 5.4] yields that

$$\min_{|u_{(1)}|=1} \|\mathsf{diag}(g_{(1:n)})\, V_{k,n}^{\mathrm{T}} u\|_2^2 = \|g_{(1:n)}\|_2^2\, \rho_{k-1}^2\, |T_{k-1}(\tau)|^{-2},$$

which, together with (4.8), leads to (4.7). □

Theorem 4.2. *Assume the conditions of* Theorem 3.2*, and let* c_1*,* c_2*,* c_3*,* c_4*,* ζ*, and* δ_N *be the same as defined there. Let* μ_k *be the largest eigenvalue of* H_k*. Then for* $1 \le k \le n$*,*

$$\begin{aligned} &(\lambda_N - \lambda_n)\left[1 + \frac{|g_{(N)}|^2}{(n-1)c_1^2}\, 2\Psi_{\delta_N,k}\right]^{-1/2} \\ &\quad\le \lambda_N - \mu_k \\ &\quad\le (\lambda_N - \lambda_1)\left[1 + \frac{|g_{(N)}|^2}{(n-1)c_2^2}\, 2\Psi_{\delta_N,k}\right]^{-1/2}, \end{aligned} \tag{4.9}$$

$$\begin{aligned} &(\lambda_N - \lambda_n)\left[1 + \frac{|g_{(N)}|^2}{\zeta c_3^2}\, \rho_{k-1}^{-2}\, |T_{k-1}(\delta_N)|^2\right]^{-1/2} \\ &\quad\le \lambda_N - \mu_k \\ &\quad\le (\lambda_N - \lambda_1)\left[1 + \frac{|g_{(N)}|^2}{\zeta c_4^2}\, \rho_{k-1}^{-2}\, |T_{k-1}(\delta_N)|^2\right]^{-1/2}. \end{aligned} \tag{4.10}$$

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

Proof. Let $\Gamma = \lambda_N I - \Lambda$. It can be verified that $\lambda_N - \lambda_n \le \|\Gamma\|_2 \le \lambda_N - \lambda_1$. Thus

$$(\lambda_N - \lambda_n)\frac{\|\mathsf{diag}(g_{(1:n)})\, V_{k,n}^{\mathrm{T}} u\|_2^2}{\|\mathsf{diag}(g)\, V_{k,N}^{\mathrm{T}} u\|_2^2} \le \frac{\|\Gamma^{1/2}\mathsf{diag}(g) V_{k,N}^{\mathrm{T}} u\|_2^2}{\|\mathsf{diag}(g) V_{k,N}^{\mathrm{T}} u\|_2^2} \le (\lambda_N - \lambda_1)\frac{\|\mathsf{diag}(g_{(1:n)})\, V_{k,n}^{\mathrm{T}} u\|_2^2}{\|\mathsf{diag}(g)\, V_{k,N}^{\mathrm{T}} u\|_2^2}. \tag{4.11}$$

It can be seen that, because of (4.8),

$$\min_u \frac{\|\mathsf{diag}(\widetilde{g}_{(1:n)})\, V_{k,n}^{\mathrm{T}} u\|_2^2}{\|\mathsf{diag}(\widetilde{g})\, V_{k,N}^{\mathrm{T}} u\|_2^2} \le \min_u \frac{\|\mathsf{diag}(g_{(1:n)})\, V_{k,n}^{\mathrm{T}} u\|_2^2}{\|\mathsf{diag}(g)\, V_{k,N}^{\mathrm{T}} u\|_2^2} \le \min_u \frac{\|\mathsf{diag}(\widehat{g}_{(1:n)})\, V_{k,n}^{\mathrm{T}} u\|_2^2}{\|\mathsf{diag}(\widehat{g})\, V_{k,N}^{\mathrm{T}} u\|_2^2}, \tag{4.12}$$

where $\widetilde{g}$ and $\widehat{g}$ are either given by (3.21) or by (3.22) and $\widetilde{g}_{(N)} = \widehat{g}_{(N)} = g_{(N)}$ always. As in the proof of Theorem 3.2, $\tau = \delta_N$ for $[\alpha, \beta] = [\lambda_1 - \lambda_N, \lambda_n - \lambda_N]$. Item 1 of Lemma 4.1, (4.11), and (4.12) give (4.9), upon noticing that $(n-1)c_1^2 = \|\widetilde{g}_{(1:n)}\|_2^2$ and $(n-1)c_2^2 = \|\widehat{g}_{(1:n)}\|_2^2$. Item 2 of Lemma 4.1, (4.11), and (4.12) give (4.10), upon noticing that $\zeta c_3^2 = \|\widetilde{g}_{(1:n)}\|_2^2$ and $\zeta c_4^2 = \|\widehat{g}_{(1:n)}\|_2^2$. □

Similarly to our analysis in Examples 3.1 and 3.2, Theorem 4.2 will also lead to examples for which the existing result (4.3) tells the correct rate of convergence to λ_N. The details are omitted. Kaniel and Saad obtained similar bounds on approximating other λ_j by Ritz values [20]. Their sharpness in general remains to be studied.

Remark 4.1. Between Theorems 3.2 and 4.2, one implies the other with slightly *weakened* inequalities. In fact we have

$$(\lambda_N - \lambda_n)\varepsilon^2 \le \lambda_N - \mu_k \le (\lambda_N - \lambda_1)\varepsilon^2, \tag{4.13}$$

where $\varepsilon = \sin\angle(Q_{(:,N)}, \mathcal{K}_k)$. The second inequality in (4.13) is a consequence of (4.5). To see the first inequality, let y be the corresponding Ritz vector to μ_k. By [15, Theorem 2.1] or [11, Theorem 4],

$$\varepsilon \le \sin\angle(Q_{(:,N)}, y) \le \sqrt{\frac{\lambda_N - \mu_k}{\lambda_N - \lambda_n}},$$

which leads to the first inequality in (4.13). We note in passing that the subspace version of (4.5) can be found in [14].

Remark 4.2. That the existing error bound for μ_k tells the correct rate of convergence to λ_N also follows from two equivalence theorems between CG convergence and the convergence of the first Ritz value, due to Sleijpen and van der Sluis [24, Theorems 6.1 and 6.2], and a recent result of the author's [18, Theorem 2.1]. In fact the Hermitian matrix H constructed according to [24, Theorem 6.1] from the matrix in [18, Theorem 2.1] relates to the Hermitian matrix A in Example 3.2 by $H = \mu A + \nu I$ for two real numbers μ and ν. However, the implied bounds by [24, Theorem 6.1] are weaker than those in Theorem 4.2. The details are omitted.

5. Concluding remarks

It is often observed that the existing error bounds for the 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 overestimate the actual errors often too much to be of much use. *Is this always the*

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

case? We have devised examples for the symmetric Lanczos algorithm to demonstrate that the computed solutions have errors that are comparable to the existing error bounds at all iteration steps for the two extreme eigenpairs. This implies that the existing bounds cannot be improved in general unless further information upon the problems becomes available.

We only succeed in dealing with the largest and smallest eigenvalues and their associated eigenvectors by showing that the existing bounds are sharp, modulo modest factors. The situation for approximations to any other eigenvalues and their associated eigenvectors can be very complicated, and we suspect that the existing bounds would probably not be sharp, even after modulo modest constant factors.

The foundation of this paper is built upon an explicit evaluation of certain minimization problems for the translated Chebyshev extreme nodes, similarly to [16] where the translated Chebyshev zero nodes were used. This idea is extendable to evaluate the minimization problems that are the same in form but involve the translated zero nodes of an orthogonal polynomial; see [17, Section 6].

In passing, we also obtained a slightly improved error bound in (4.4) over (4.3), an existing result of Kaniel and Saad [20].

Acknowledgment

The author wishes to thank an anonymous referee for his/her constructive comments that significantly improved the paper. He especially is indebted to the referee for the observation $\lambda_N - \mu_k \le (\lambda_N - \lambda_1)\sin^2\angle(Q_{(:,N)}, \mathcal{K}_k)$ that led to (4.4). The referee also pointed out the possibility of combining the results from [18, 24], as we commented in Remark 4.2.

References

[1] O. Axelsson, *Iterative solution methods*, Cambridge University Press, Cambridge, 1994. MR1276069 (95f:65005)

[2] E. W. Cheney, *Introduction to approximation theory*, 2nd ed., Chelsea Publishing Company, New York, 1982. MR1656150 (99f:41001)

[3] J. Demmel, *Applied numerical linear algebra*, SIAM, Philadelphia, PA, 1997. MR1463942 (98m:65001)

[4] Anne Greenbaum, *Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences*, Linear Algebra Appl. **113** (1989), 7–63. MR978581 (90e:65044)

[5] ———, *Iterative methods for solving linear systems*, SIAM, Philadelphia, 1997. MR1474725 (98j:65023)

[6] Anne Greenbaum and Z. Strakos, *Predicting the behavior of finite precision Lanczos and conjugate gradient computations*, SIAM J. Matrix Anal. Appl. **13** (1992), no. 1, 121–137. MR1146656 (92j:65043)

[7] Martin Hanke, *Superlinear convergence rates for the Lanczos method applied to elliptic operators*, Numer. Math. **77** (1997), no. 4, 487–499. MR1473393 (98m:65182)

[8] I. C. F. Ipsen, *Expressions and bounds for the GMRES residual*, BIT **40** (2000), no. 3, 524–535. MR1780406 (2001g:65032)

[9] Zhongxiao Jia and G. W. Stewart, *An analysis of the Rayleigh-Ritz method for approximating eigenspaces*, Math. Comp. **70** (2001), 637–647. MR1697647 (2001g:65040)

[10] S. Kaniel, *Estimates for some computational techniques in linear algebra*, Math. Comp. **20** (1966), no. 95, 369–378. MR0234618 (38:2934)

[11] J. Kovač-Striko and K. Veselić, *Some remarks on the spectra of Hermitian matrices*, Linear Algebra Appl. **145** (1991), 221–229. MR1080687 (92b:15018)

[12] A. B. J. Kuijlaars, *Which eigenvalues are found by the Lanczos method?*, SIAM J. Matrix Anal. Appl. **22** (2000), no. 1, 306–321. MR1779731 (2001g:65042)

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use

[13] Arno B. J. Kuijlaars, *Convergence analysis of Krylov subspace iterations with methods from potential theory*, SIAM Rev. **48** (2006), no. 1, 3–40. MR2219308 (2007a:65054)
[14] Ren-Cang Li, *On eigenvalues of a Rayleigh quotient matrix*, Linear Algebra Appl. **169** (1992), 249–255. MR1158372 (93a:15013)
[15] ———, *Accuracy of computed eigenvectors via optimizing a Rayleigh quotient*, BIT **44** (2004), no. 3, 585–593. MR2106018 (2005i:65055)
[16] ———, *Sharpness in rates of convergence for CG and symmetric Lanczos methods*, Technical Report 2005-01, Department of Mathematics, University of Kentucky, 2005, Avaliable at `http://www.ms.uky.edu/~math/MAreport/`.
[17] ———, *Vandermonde matrices with Chebyshev nodes*, Linear Algebra Appl. **428** (2007), 1803–1832. MR2398120
[18] ———, *On Meinardus' examples for the conjugate gradient method*, Math. Comp. **77** (2008), no. 261, 335–352, Electronically published on September 17, 2007. MR2353956
[19] J. Liesen, M. Rozložník, and Z. Strakoš, *Least squares residuals and minimal residual methods*, SIAM J. Sci. Comput. **23** (2002), no. 5, 1503–1525. MR1885072 (2003a:65033)
[20] 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. MR570116 (81j:65063)
[21] Y. Saad, *On the rates of convergence of the Lanczos and the block-Lanczos methods*, SIAM J. Numer. Anal. **15** (1980), no. 5, 687–706. MR588755 (82g:65022)
[22] Yousef Saad, *Iterative methods for sparse linear systems*, 2nd ed., SIAM, Philadelphia, 2003. MR1990645 (2004h:65002)
[23] D. S. Scott, *How to make the Lanczos algorithm converge slowly*, Math. Comp. **33** (1979), no. 145, 239–247. MR514821 (80c:65091)
[24] G. L. G. Sleijpen and A. van der Sluis, *Further results on the convergence behavior of conjugate-gradients and Ritz values*, Linear Algebra Appl. **246** (1996), 233–378. MR1407670 (97j:65067)
[25] Ji-Guang Sun, *Eigenvalues of Rayleigh quotient matrices*, Numer. Math. **59** (1991), 603–614. MR1124130 (93a:15015)
[26] Lloyd N. Trefethen and David Bau, III, *Numerical linear algebra*, SIAM, Philadelphia, 1997. MR1444820 (98k:65002)
[27] A. van der Sluis and H. A. van der Vorst, *The convergence behavior of Ritz values in the presence of close eigenvalues*, Linear Algebra Appl. **88/89** (1987), 651–694. MR882466 (88f:65064)

Department of Mathematics, University of Texas at Arlington, P.O. Box 19408, Arlington, Texas 76019-0408.
E-mail address: `rcli@uta.edu`
URL: `http://www.uta.edu/faculty/rcli`

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use