Accuracy of Rayleigh-Ritz Approximations

Zhongming Teng
Linzhang Lu
Ren-Cang Li
Accuracy of Rayleigh-Ritz Approximations

Zhongming Teng* Linzhang Lu† Ren-Cang Li‡

January 12, 2015

Abstract

New bounds on the canonical angles between an invariant subspace of $A$ and an approximating subspace by the differences between Ritz values and the targeted eigenvalues are obtained. From this result, various bounds are readily available to estimate how accurate the Ritz vectors computed from the approximating subspace may be, based on information on approximation accuracies in the Ritz values. The result is helpful in understanding how Ritz vectors move towards eigenvectors while Ritz values are made to move towards eigenvalues.

Key words. invariant subspace, canonical angle, eigenvalue, eigenvector, Rayleigh-Ritz procedure, Ritz value, Ritz vector, majorization

AMS subject classifications. 15A42, 65F15

1 Introduction

The Rayleigh-Ritz procedure [11, p.234] is widely used to find approximate eigenpairs of a Hermitian matrix $A \in \mathbb{C}^{N \times N}$, given a subspace $X$ of $\mathbb{C}^N$ with $\dim(X) = \ell$. Let $X$ be an orthonormal basis matrix of $X$. The basic idea of the procedure goes as follows. Compute the eigen-decomposition of $X^HAX$:

$$X^HAX = W\tilde{\Lambda}W^H,$$

where $W = [w_1, \ldots, w_{\ell}] \in \mathbb{C}^{\ell \times \ell}$ is unitary, and then take each pair $(\tilde{\lambda}_i, Xw_i)$, called a Rayleigh-Ritz pair, as an approximate eigenpair of $A$. The number $\tilde{\lambda}_i$ is called a Ritz value and $Xw_i$ a corresponding Ritz vector.

In the case when $X$ is an approximate invariant subspace of $A$, each Rayleigh-Ritz pair should be a good approximate eigenpair of $A$. In the case when $X$ isn’t near an invariant subspace of $A$ but nearby $X$ there is an invariant subspace of $A$ (of dimension less than $\ell$), then some, but not all, of the Rayleigh-Ritz pair should be good approximate eigenpairs.

---

*School of Computer and Information, Fujian Agriculture and Forestry University, Fuzhou, Fujian, 350002, P. R. China. Email: peter979@163.com.

†School of Mathematical Science, Xiamen University & School of Mathematics and Computer Science, Guizhou Normal University, P. R. China. Email: lzl@xmu.edu.cn, llz@gznu.edu.cn. Supported in part by National Natural Science Foundation of China Grant No.11261012.

‡Department of Mathematics, University of Texas at Arlington, P.O. Box 19408, Arlington, TX 76019. Email: rcli@uta.edu. School of Mathematical Science, Xiamen University, Xiamen, Fujian, China. Supported in part by NSF grants DMS-1115834 and DMS-1317330, and a Research Gift Grant from Intel Corporation.
of $A$. The latter case is more common than the former in eigenvalue computations, where often a subspace is built so that nearby the subspace there is an invariant subspace of $A$, e.g., in the Lanczos method a Krylov subspace is built and usually the Krylov subspace as a whole is not close to any invariant subspace (of the same dimension) but more likely nearby the Krylov subspace there is an invariant subspace of $A$ (of a smaller dimension).

There are existing results to quantify how good the approximate eigenpairs are. Most results are bounds in terms of the norms of the residual

$$R(X) := AX - X(X^HAX)$$

in the case when $X$ is an approximate invariant subspace. The interested reader is referred to a short summary at the end of [2] for bounds of this kind. The result of [7] can be interpreted as a result of this kind, too. Note $R(X) = 0$ if $X$ is an exact invariant subspace.

Another set of results are also bounds but in terms of the canonical angles between $X$ and the invariant subspace which $X$ is supposed to approximate. In this regard, Knyazev and Argentati [5] presented the most comprehensive study so far. They obtained several beautiful results in terms of how the vector of eigenvalue differences between the exact eigenvalues and their approximations is weakly majorized by the canonical angles between/from the invariant subspace of interest and/to $X$. We will state some of their results to motivate what we will do in section 3. The results in [5] are basically about estimating the approximation accuracy of (some of) the Ritz values, given information on the approximate accuracy in $X$ to an invariant subspace of $A$. In this paper, we are interested in the converses to these results, i.e., estimating the approximate accuracy in $X$, given information on the approximation accuracy of (some of) the Ritz values. Our motivation is from eigenvector computations in Principal Component Analysis in image processing [10, 14], where eigenvectors may be computed by optimizing Rayleigh quotients with the conjugate gradient type methods. Recently in large scale electronic structure calculations, there is an increasing trend to compute a few extreme eigenvalues and corresponding eigenvectors through optimizing Rayleigh quotients. Our main result is helpful in our understanding the relations between approximation accuracies in Ritz values and vectors.

The rest of this paper is organized as follows. Section 2 collects some preliminaries on the concept of majorization and the canonical angles of two subspaces. In section 3, we first discuss existing results of Knyazev and Argentati [5] and then let the discussion lead to our main result – Theorem 3.2. Section 4 is devoted to the proof of Theorem 3.2. Concluding remarks are given in section 6.

Notation. $\mathbb{C}^{m \times n}$ is the set of all $m \times n$ complex matrices, $\mathbb{C}^n = \mathbb{C}^{n \times 1}$, and $\mathbb{C} = \mathbb{C}^1$. Similarly for $\mathbb{R}^{m \times n}$, $\mathbb{R}^n$, and $\mathbb{R}$, except all involved numbers are real. For $X \in \mathbb{C}^{n \times n}$, eig($X$) $\in \mathbb{C}^{1 \times n}$ is the vector of the eigenvalues of $X$. We shall also adopt MATLAB-like convention to access the entries of vectors and matrices. Let $i : j$ be the set of integers from $i$ to $j$ inclusive. For a vector $u$ and a matrix $X$, $u_{(j)}$ is $u$’s $j$th entry, $X_{(i,j)}$ is $X$’s $(i,j)$th entry; $X$’s submatrices $X_{(k:l,i:j)}$, $X_{(k:l,:)}$, and $X_{(:,i:j)}$ consist of intersections of row $k$ to row $\ell$ and column $i$ to column $j$, row $k$ to row $\ell$, and column $i$ to column $j$, respectively. For $X \in \mathbb{C}^{m \times n}$, $\mathcal{R}(X)$ is the column space of $X$, $X^H$ is its complex conjugate transpose. $I_n$ is the $n \times n$ identity matrix or simply $I$ if its dimension is clear from the context.
2 Preliminaries

2.1 Majorization

For $x = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{1 \times n}$, denote by $x^\dagger = [x_1^\dagger, x_2^\dagger, \ldots, x_n^\dagger]$, the row vector obtained by rearranging $x_i$ in descending order, i.e.,

$$x_1^\dagger \geq x_2^\dagger \geq \cdots \geq x_n^\dagger.$$

Given two vectors $x = [x_1, \ldots, x_n], y = [y_1, \ldots, y_n] \in \mathbb{R}^{1 \times n}$, we say that $x$ is weakly majorized by $y$, in symbols $x \prec_w y$, if [1, chapter II]

$$\sum_{i=1}^{k} x_i^\dagger \leq \sum_{i=1}^{k} y_i^\dagger, \quad \text{for } 1 \leq k \leq n. \quad (2.1)$$

If, in addition,

$$\sum_{i=1}^{n} x_i^\dagger = \sum_{i=1}^{n} y_i^\dagger;$$

we say that the vector $x$ is majorized by $y$, written $x \prec y$. Notation $x \circ y = [x_1 y_1, \ldots, x_n y_n]$ denotes the Hadamard product of $x$ and $y$. In particular, $x^{\circ 2} = x \circ x$.

Majorization provides a succinct way to express numerous useful inequalities involving two vectors $x$ and $y$. For example, suppose $x$ and $y$ are entrywise nonnegative, i.e., $x_i \geq 0$ and $y_i \geq 0$, and $x \prec_w y$. Then, besides those in (2.1) by the definition, we have [1, p.42]

$$\|x\|_p := \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p} \leq \|y\|_p := \left( \sum_{i=1}^{n} |y_i|^p \right)^{1/p} \quad \text{for } 1 \leq p \leq \infty. \quad (2.2)$$

For $p = \infty$, this is simply (2.1) for $k = 1$: $\max_i |x_i| \leq \max_i |y_i|$. For the purpose of error estimation in numerical computations, often the inequality (2.2) for $p = 2$ and $\infty$ suffices.

2.2 Angles between subspaces

Consider two subspaces $\mathcal{X}$ and $\mathcal{Y}$ of $\mathbb{C}^N$ and suppose

$$\ell := \dim(\mathcal{X}) \geq \dim(\mathcal{Y}) =: k. \quad (2.3)$$

Let $X \in \mathbb{C}^{N \times \ell}$ and $Y \in \mathbb{C}^{N \times k}$ be orthonormal basis matrices of $\mathcal{X}$ and $\mathcal{Y}$, respectively, i.e.,

$$X^H X = I_\ell, \ X = \mathcal{R}(X), \quad \text{and} \quad \ Y^H Y = I_k, \ Y = \mathcal{R}(Y),$$

and denote by $\sigma_j$ for $1 \leq j \leq k$ in ascending order, i.e., $\sigma_1 \leq \cdots \leq \sigma_k$, the singular values of $Y^H X$. The $k$ canonical angles $\theta_j(\mathcal{X}, \mathcal{Y})$ from $\mathcal{Y}$ to $\mathcal{X}$ are defined by

$$0 \leq \theta_j(\mathcal{X}, \mathcal{Y}) := \arccos \sigma_j \leq \frac{\pi}{2} \quad \text{for } 1 \leq j \leq k. \quad (2.4)$$

\footnote{If $\ell = k$, we may says that these angles are between $\mathcal{X}$ and $\mathcal{Y}$.}
They are in descending order, i.e., \( \theta_1(X, y) \geq \cdots \geq \theta_k(X, y) \). Set
\[
\Theta(X, y) = [\theta_1(X, y), \ldots, \theta_k(X, y)] \in \mathbb{R}^{1 \times k}.
\] (2.5)

It can be seen that angles so defined are independent of the orthonormal basis matrices \( X \) and \( Y \) (which are not unique). A different way to define these angles is through orthogonal projections onto \( X \) and \( Y \) [12].

When \( k = 1 \), i.e., \( Y \) is a vector, there is only one canonical angle from \( Y \) to \( X \) and so we will simply write \( \theta(X, y) \).

For any given function \( f(t) \) defined for \( t \in [0, \pi/2] \), we define
\[
f(\Theta(X, y)) := [f(\theta_1(X, y)), \ldots, f(\theta_k(X, y))].
\]

In what follows, we sometimes place a vector or matrix in one of or both arguments of \( \theta(\cdot, \cdot) \), \( \theta(\cdot, \cdot, \cdot) \), and \( \Theta(\cdot, \cdot) \) with the understanding that it is about the subspace spanned by the vector or the columns of the matrix argument.

**Proposition 2.1** ([9, Proposition 2.1]). Let \( X \) and \( Y \) be two subspaces in \( \mathbb{C}^N \) with \( \dim(X) = \ell \), \( \dim(Y) = k \) and \( \ell \geq k \).

1. For any \( \hat{X} \subseteq X \) with \( \dim(\hat{X}) = \dim(Y) = k \), we have \( \theta_j(X, y) \leq \theta_j(\hat{X}, Y) \) for \( 1 \leq j \leq k \).

2. There exist \( \hat{X} \subseteq X \) with \( \dim(\hat{X}) = k \leq \ell \) such that \( \theta_j(X, y) = \theta_j(\hat{X}, Y) \) for \( 1 \leq j \leq k \).

**Proposition 2.2.** Let \( X \) and \( Y \) be two subspaces in \( \mathbb{C}^N \) satisfying (2.3). Then
\[
\theta_i(X, y) = 0 \quad \text{for } k - m_0 + 1 \leq i \leq k,
\]
where \( m_0 = \dim(X \cap Y) \).

**Proof.** Let \( X \in \mathbb{C}^{N \times \ell} \) and \( Y \in \mathbb{C}^{N \times k} \) be orthonormal basis matrices of \( X \) and \( Y \), respectively. Without loss of generality, we may assume
\[
X^H Y = \begin{bmatrix}
\diag(\sigma_1, \sigma_2, \ldots, \sigma_k) \\
0_{(\ell-k) \times k}
\end{bmatrix} \in \mathbb{C}^{\ell \times k},
\]
where
\[
0 \leq \sigma_1 \leq \cdots \leq \sigma_r < \sigma_{r+1} = \cdots = \sigma_k = 1,
\]
where \( 1 \leq r \leq k \). The case \( r = k \) means no \( \sigma_i = 1 \). That is \( \theta_i(X, Y) = 0 \) for \( r + 1 \leq i \leq k \). We shall prove \( r = k - m_0 \). To this end, we write \( X = [x_1, x_2, \ldots, x_\ell] \) and \( Y = [y_1, y_2, \ldots, y_k] \). Since \( x_i^H y_i = 1 \) and \( x_i \) and \( y_i \) are unit vectors for \( r + 1 \leq i \leq k \), we have \( x_i = y_i \) for \( r + 1 \leq i \leq k \). We claim \( x_1, x_2, \ldots, x_r, y_1, y_2, \ldots, y_r \) form a basis of \( X + Y \). First they span \( X + Y \) because all \( x_i \) and all \( y_i \) are among them, and second they are linearly independent, because
\[
[X, y_1, y_2, \ldots, y_r]^H [X, y_1, y_2, \ldots, y_r] = \begin{bmatrix}
I_\ell & \Sigma^T \\
\Sigma & I_r
\end{bmatrix}
\]
is positive definite and thus nonsingular, where
\[
\Sigma = \begin{bmatrix}
\diag(\sigma_1, \cdots, \sigma_r) \\
0_{(\ell-r) \times r}
\end{bmatrix} \in \mathbb{R}^{\ell \times r}.
\]
Therefore \( \dim(X + Y) = \ell + r \). A well-known formula also gives \( \dim(X + Y) = \ell + k - m_0 \).
So \( r = k - m_0 \), as expected. \( \square \)
3 Main Result

For the rest of this paper, \( A \) is an \( N \times N \) Hermitian matrix, and has

\[
\begin{align*}
\text{eigenvalues:} & \quad \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N, \\
\text{orthonormal eigenvectors:} & \quad u_1, u_2, \ldots, u_N, \text{ and} \\
\text{eigen-decomposition:} & \quad A = U \Lambda U^H \quad \text{and} \quad U^H U = I_N.
\end{align*}
\]

(3.1)

As in the introduction, let \( X \) be a subspace of \( \mathbb{C}^N \) with \( \dim(X) = \ell \), intended to approximate an invariant subspace of \( A \) in the sense that either \( X \) as a whole is an approximate invariant subspace or nearby \( X \) there is an invariant subspace of \( A \) of dimension less than \( \ell \). \( X \in \mathbb{C}^{N \times \ell} \) is an orthonormal basis matrix of \( X \). Similarly, we introduce notations for \( X^H A X \):

\[
\begin{align*}
\text{eigenvalues (also Ritz values):} & \quad \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{\ell}, \\
\text{orthonormal eigenvectors:} & \quad w_1, w_2, \ldots, w_{\ell}, \text{ and} \\
\text{eigen-decomposition:} & \quad (X^H A X) W = W \Omega \quad \text{and} \quad W^H W = I_\ell, \\
\text{Ritz vectors:} & \quad \tilde{u}_j = X w_j \text{ for } 1 \leq j \leq \ell, \text{ and} \\
& \quad U = [\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_\ell].
\end{align*}
\]

(3.2)

Knyazev and Argentati [5] established several beautiful results that lead to bounds on \( \lambda_i - \tilde{\lambda}_i \) in terms of the canonical angles between/from an invariant subspace of \( A \) and/or \( X \). Some of them that are relevant to this paper are summarized in Theorem 3.1 below.

**Theorem 3.1 ([5]).** Suppose \( Y \) with \( \dim(Y) = k \) is an invariant subspace of \( A \), corresponding to its eigenvalues

\[ \lambda_{\pi_1} \leq \lambda_{\pi_2} \leq \cdots \leq \lambda_{\pi_k}, \]

where \( 1 \leq \pi_1 < \pi_2 < \cdots < \pi_k \leq N \) and suppose \( k \leq \ell \). Let \( m = \dim(X + Y) \), and let the eigenvalues of \( P_{X+Y} A |_{X+Y} \) be \( \omega_1 \leq \omega_2 \leq \cdots \leq \omega_m \), where \( P_{X+Y} \) is the orthogonal projector onto \( X + Y \) and \( P_{X+Y} A |_{X+Y} \) is the restriction of \( P_{X+Y} A \) onto \( X + Y \), and set

\[
\begin{align*}
\delta & := \omega_m - \omega_1 \quad (\text{which is no bigger than } \lambda_N - \lambda_1), \\
\Delta & := [\omega_m - \omega_1, \omega_m - \omega_2, \ldots, \omega_m - \omega_k], \\
\tilde{\Delta} & := [\omega_m - \omega_1, \omega_m - \omega_2, \ldots, \omega_m - \omega_k].
\end{align*}
\]

(3.3) (3.4) (3.5)

1. [5, Theorem 2.1] If \( k = \ell \), then

\[
[|\tilde{\lambda}_1 - \lambda_{\pi_1}|, \ldots, |\tilde{\lambda}_k - \lambda_{\pi_k}|] \prec_w \delta \sin^2 \Theta(X, Y).
\]

(3.6)

2. [5, Theorem 2.2] If \( k = \ell \) and if \( \pi_i = i \) for \( 1 \leq i \leq \ell \), then

\[
0 \leq [\tilde{\lambda}_1 - \lambda_1, \ldots, \tilde{\lambda}_k - \lambda_k] \prec_w \Delta \sin^2 \Theta(X, Y) \leq \delta \sin^2 \Theta(X, Y).
\]

(3.7) (3.8)
3. [5, Theorem 2.4] If \( k < \ell \) and if \( \pi_i = i \) for \( 1 \leq i \leq k \), then
\[
0 \leq [\tilde{\lambda}_1 - \lambda_1, \ldots, \tilde{\lambda}_k - \lambda_k] \prec \wedge \delta \sin^2 \Theta(X, y)
\]
\[
\prec \wedge \delta \sin^2 \Theta(X, y). \tag{3.9}
\]

The theory behind computing a few extreme eigenvalues and corresponding eigenvectors through optimizing Rayleigh quotients is the well-known Ky Fan trace minimization principle [3]
\[
\min_{Z^HZ = I_k} \text{trace}(Z^HAX) = \sum_{i=1}^{k} \lambda_i
\]
and its variations. Conceivably, as \( \text{trace}(X^HAX) \) approaches the targeted value \( \sum_{i=1}^{k} \lambda_i \), \( R(X) \) should be a good approximate invariant subspace. The question is how good it is. For that purpose, there is an existing well-known inequality in the case \( k = \ell \):
\[
(\lambda_{k+1} - \lambda_k) \| \sin \Theta(X, U_k) \|^2 \leq \sum_{i=1}^{k} (\tilde{\lambda}_i - \lambda_i), \tag{3.11}
\]
first appeared in [13] in 1974 and rediscovered a few times later (see [4, 6, 8]), where
\[
U_k = R(U_{(1:k)}). \tag{3.12}
\]
We caution the reader that here \( \sin \Theta(X, U_k) \) is a vector and thus
\[
\| \sin \Theta(X, U_k) \|^2 = \sum_{i=1}^{k} \sin^2 \theta_i(X, U_k).
\]

Given the various results collected in Theorem 3.1 and (3.11), it is natural to expect
\[
(\lambda_{k+1} - \lambda_k) \sin^2 \Theta(X, U_k) \prec \wedge [\tilde{\lambda}_1 - \lambda_1, \ldots, \tilde{\lambda}_k - \lambda_k]. \tag{3.13}
\]
Unfortunately, it may fail, as Example 3.1 shows.

**Example 3.1.** Consider
\[
A = \begin{bmatrix}
-1.0642 & -0.2490 \\
0.2490 & 1.2347 \\
1.2347 & 1.6035
\end{bmatrix}
\]
and
\[
X = \begin{bmatrix}
-0.1440 & 0.4026 \\
-0.9444 & -0.3116 \\
-0.2788 & 0.6496 \\
-0.0978 & 0.5646
\end{bmatrix}.
\]

Here \( X \) is only shown with 4 significant digits of \( X \). To make this example repeatable, within MATLAB we correct \( X \) by executing \( \text{qr}(X, 0) \) and reset \( X \) accordingly. It is computed (again only 4 significant digits are shown)
\[
\text{eig}(X^HAX) = [-0.2311, 0.9338], \quad \text{i.e.,} \quad \tilde{\lambda}_1 = -0.2311, \quad \tilde{\lambda}_2 = 0.9338.
\]

It is clear that
\[
\lambda_1 = -1.0642, \quad \lambda_2 = -0.2490, \quad \text{and} \quad U_{(1:2)} = \begin{bmatrix}
1 & 0 \\
0 & 1 \\
0 & 0 \\
0 & 0
\end{bmatrix}.
\]
We have \( \sin \Theta(\mathcal{X}, \mathcal{U}_2) = [0.9040, 0.1039] \) which consists of the singular values of \( X_{(3,4,:)} \). Finally
\[
(\lambda_3 - \lambda_2) \sin^2 \theta_1(\mathcal{X}, \mathcal{U}_2) = 1.2126 > \max_{1 \leq i \leq 2} (\tilde{\lambda}_i - \lambda_i) = 1.1828,
\]
an opposite of the inequality (3.13).

Although (3.13) may fail, in what follows we will establish inequalities that resemble (3.13) but only slightly weaker. Compared to (3.11), our results are more general and sharper. More discussions are at the end of this section.

**Theorem 3.2.** Let \( m = \dim(\mathcal{X} + \mathcal{U}_k) \), and let the eigenvalues of \( P_{\mathcal{X}+\mathcal{U}_k} A_{\mathcal{X}+\mathcal{U}_k} \) be
\[
\omega_1 \leq \omega_2 \leq \cdots \leq \omega_m.
\]
Define \( \mathcal{X}_j := \Re(X[w_1, w_2, \ldots, w_j]) \). If \( m > k \), then
\[
\sum_{i=1}^{\min\{m-k,j\}} (\omega_{k+i} - \omega_{j-i+1}) \sin^2 \theta_i(\mathcal{U}_k, \mathcal{X}_j) \leq \sum_{i=1}^{j} (\tilde{\lambda}_i - \lambda_i) \quad \text{for} \ 1 \leq j \leq k, \quad (3.14a)
\]
\[
(\omega_{k+1} - \omega_j) \sum_{i=1}^{j} \sin^2 \theta_i(\mathcal{U}_k, \mathcal{X}_j) \leq \sum_{i=1}^{j} (\tilde{\lambda}_i - \lambda_i) \quad \text{for} \ 1 \leq j \leq k. \quad (3.14b)
\]

The condition \( m > k \) in the theorem is not restrictive at all. In fact, \( m \geq k \) always, and if \( m = k \), then \( \mathcal{X} = \mathcal{U}_k \) and thus \( \Theta(\mathcal{X}, \mathcal{U}_k) = 0 \) and \( \tilde{\lambda}_i = \lambda_i \) for \( 1 \leq i \leq k \), a highly unlikely but otherwise trivial and welcomed case for which the inequalities (3.14a) and (3.14b) trivially hold as equalities.

Suppose \( m > k \). By the Cauchy interlacing inequalities [11],
\[
\lambda_i \leq \omega_i \leq \lambda_{N-m+i} \quad \text{for} \ 1 \leq i \leq m.
\]
Since \( \mathcal{U}_k \subset \mathcal{X} + \mathcal{U}_k \), \( \lambda_i = \omega_i \) for \( 1 \leq i \leq k \). Therefore Theorem 3.2 implies

**Corollary 3.1.** Assume the conditions of Theorem 3.2, we have
\[
\sum_{i=1}^{\min\{m-k,j\}} (\lambda_{k+i} - \lambda_{j-i+1}) \sin^2 \theta_i(\mathcal{U}_k, \mathcal{X}_j) \leq \sum_{i=1}^{j} (\tilde{\lambda}_i - \lambda_i) \quad \text{for} \ 1 \leq j \leq k, \quad (3.15a)
\]
\[
(\lambda_{k+1} - \lambda_j) \sum_{i=1}^{j} \sin^2 \theta_i(\mathcal{U}_k, \mathcal{X}_j) \leq \sum_{i=1}^{j} (\tilde{\lambda}_i - \lambda_i) \quad \text{for} \ 1 \leq j \leq k. \quad (3.15b)
\]

These inequalities in (3.15) differ from the ones in (3.14) only in replacing \( \omega_i \) by \( \lambda_i \).

**Corollary 3.2.** Assume the conditions of Theorem 3.2, we have
\[
\sum_{i=1}^{k} (\omega_{k+i} - \omega_{k-i+1}) \sin^2 \theta_i(\mathcal{U}_k, \mathcal{X}) \leq \sum_{i=1}^{k} (\tilde{\lambda}_i - \lambda_i), \quad (3.16a)
\]
\[
\sum_{i=1}^{k} (\lambda_{k+i} - \lambda_{k-i+1}) \sin^2 \theta_i(\mathcal{U}_k, \mathcal{X}) \leq \sum_{i=1}^{k} (\tilde{\lambda}_i - \lambda_i). \quad (3.16b)
\]
Proof. Let \( j = k \) in (3.14a), we have
\[
\sum_{i=1}^{k} (\tilde{\lambda}_i - \lambda_i) \geq \sum_{i=1}^{k} (\omega_{k+i} - \omega_{k-i+1}) \sin^2 \theta_i(\mathbf{u}_k, \mathbf{x}_k)
\]
\[
\geq \sum_{i=1}^{k} (\omega_{k+i} - \omega_{k-i+1}) \sin^2 \theta_i(\mathbf{u}_k, \mathbf{x}).
\]
The last inequality holds because of Proposition 2.1. This gives (3.16a). The inequality (3.16b) follows from (3.16a) due to the same reason as Corollary 3.1 follows from Theorem 3.2.

The inequality (3.16b) implies (3.11). Next we shall derive an inequality from (3.14b) that only slightly differs from (3.13) that was expected but not true. Let
\[
[\tilde{\lambda}_1 - \lambda_1, \ldots, \tilde{\lambda}_k - \lambda_k] = [\tilde{\lambda}_1 - \lambda_1, \ldots, \tilde{\lambda}_k - \lambda_k],
\]
where \( \{t_i\} \) for \( 1 \leq i \leq k \) is a permutation of \( \{1, 2, \ldots, k\} \). Notice that \( \tilde{\lambda}_i - \lambda_i \geq 0 \) for \( 1 \leq i \leq k \). By (3.14b), we have for \( 1 \leq j \leq k \)
\[
\sum_{i=1}^{j} (\tilde{\lambda}_{t_i} - \lambda_{t_i}) \geq \sum_{i=1}^{j} (\tilde{\lambda}_i - \lambda_i)
\]
\[
\geq (\omega_{k+1} - \omega_j) \sum_{i=1}^{j} \sin^2 \theta_i(\mathbf{u}_k, \mathbf{x}_j)
\]
\[
\geq (\lambda_{k+1} - \lambda_k) \sum_{i=1}^{j} \sin^2 \theta_i(\mathbf{u}_k, \mathbf{x}_j).
\]
(3.18)
This is very similar to (3.13). The only difference between them is the use of \( \theta_i(\mathbf{u}_k, \mathbf{x}_j) \) in (3.18) instead of \( \theta_i(\mathbf{u}_k, \mathbf{x}) \) in (3.13).

4 Proof of Theorem 3.2

We need the following lemmas.

**Lemma 4.1** ([8, Lemma 2.3]). Let \( \gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_m \geq 0 \). If \([\beta_1, \beta_2, \ldots, \beta_m]\) majorizes \([\alpha_1, \alpha_2, \ldots, \alpha_m]\), then
\[
\sum_{i=1}^{m} \gamma_i \beta_i^4 \leq \sum_{i=1}^{m} \gamma_i \alpha_i \leq \sum_{i=1}^{m} \gamma_i \beta_i^4.
\]

**Lemma 4.2** (Schur’s Theorem [1, p.35]). For Hermitian matrix \( H \in \mathbb{C}^{n \times n} \),
\[
\text{diag}(H) \prec \text{eig}(H),
\]
where \( \text{diag}(H) \in \mathbb{R}^{1 \times n} \) is the vector of the diagonal entries of \( H \).

**Proof** of Theorem 3.2. Without loss of generality, we first assume \( A \) is positive definite. Otherwise we shift \( A \) to \( A - \mu I \) for any \( \mu < \lambda_1 \). Doing so will shift all eigenvalues of \( A \) and Ritz values by the same \( \mu \) and thus keep all the differences \( \lambda_i - \tilde{\lambda}_j, \lambda_i - \lambda_j \) unchanged.
We may also assume \(X + \mathcal{U}_k = \mathbb{C}^N\) since we can simply replace \(A\) by \(P_{X + \mathcal{U}_k}A|_{X + \mathcal{U}_k}\). Doing so changes no Ritz values and Ritz vectors, but changes \(N\) to \(m\) and reduce the set \(\{\lambda_i\}_{i=1}^N\) to \(\{\omega_i\}_{i=1}^m\). However, \(\omega_i = \lambda_i\) for \(1 \leq i \leq k\) because \(\mathcal{U}_k \subseteq X + \mathcal{U}_k\).

Assume \(A\) is positive definite and \(X + \mathcal{U}_k = \mathbb{C}^N\).

Recall (3.1) nd (3.2) and partition \(W\) and \(U\) as

\[
W = \begin{bmatrix} W_1 & W_2 \end{bmatrix}, \quad U = \begin{bmatrix} U_1 & U_2 & U_3 \end{bmatrix}, \quad A = \begin{bmatrix} A_1 & 0 & 0 \\ 0 & A_2 & 0 \\ 0 & 0 & A_3 \end{bmatrix}.
\]

Let \(\tilde{X} = XW_1\), and write

\[
U^H \tilde{X} = U_{N-k}^{k-j} \begin{bmatrix} \tilde{X}_1^{j} \\ \tilde{X}_2^{j} \\ \tilde{X}_3^{j} \end{bmatrix}, \quad \tilde{X}_{1,2} = \begin{bmatrix} \tilde{X}_1^{j} \\ \tilde{X}_2^{j} \end{bmatrix}, \quad A_{1,2} = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}.
\]

It can be verified that \(\tilde{X}_{1,2}^{H} \tilde{X}_{1,2}^{j} + \tilde{X}_3^{H} \tilde{X}_3^{j} = I_j\) and thus

\[
eig(\tilde{X}_3^{H} \tilde{X}_3) = \eig(I_j - \tilde{X}_{1,2}^{H} \tilde{X}_{1,2}) = [\sin^2 \theta_i(\mathcal{U}_k, X_j)\) for \(1 \leq i \leq j]\), \quad (4.1)
\]

\[
eig(I_j - \tilde{X}_{1,2}^{H} \tilde{X}_{1,2}) = [\sin^2 \theta_j(\mathcal{U}_k, X_j)\) for \(1 \leq i \leq j, 1, \ldots, 1\)\], \quad (4.2)
\]

\[
eig(\tilde{X}_3^{H} \tilde{X}_3) = [\sin^2 \theta_i(\mathcal{U}_k, X_j)\) for \(1 \leq i \leq \min\{N - k, j\}, 0, \ldots, 0\)]\], \quad (4.3)
\]

We have \(\tilde{X}^{H} A \tilde{X} = \tilde{X}_{1,2}^{H} A_{1,2} \tilde{X}_{1,2} + \tilde{X}_3^{H} A_3 \tilde{X}_3 = \tilde{X}_{1,2}^{H} A_{1,2} \tilde{X}_{1,2} + \tilde{X}_3^{H} A_3 \tilde{X}_3\). Therefore

\[
\sum_{i=1}^{j} \lambda_i = \trace(\tilde{X}^{H} A \tilde{X})
\]

\[
= \trace(\tilde{X}_{1,2}^{H} A_{1,2} \tilde{X}_{1,2}) + \trace(\tilde{X}_3^{H} A_3 \tilde{X}_3)
\]

\[
= \trace(\tilde{X}_{1,2}^{H} A_{1,2} \tilde{X}_{1,2}) + \trace(\tilde{X}_3^{H} A_3 \tilde{X}_3)
\]

which yields

\[
\sum_{i=1}^{j} (\tilde{X}_i - \lambda_i) = -\trace([I_k - \tilde{X}_{1,2}^{H} \tilde{X}_{1,2}] A_{1,2}) + \sum_{i=j+1}^{k} \lambda_i + \trace(\tilde{X}_3^{H} A_3 \tilde{X}_3).
\]

(4.4)

Since \(\diag(I_k - \tilde{X}_{1,2}^{H} \tilde{X}_{1,2}) \prec \eig(I_k - \tilde{X}_{1,2}^{H} \tilde{X}_{1,2})\) and \(\diag(\tilde{X}_3^{H} \tilde{X}_3) \prec \eig(\tilde{X}_3^{H} \tilde{X}_3)\) by Lemma 4.2, we have by Lemma 4.1, (4.2), and (4.3)

\[
\trace([I_k - \tilde{X}_{1,2}^{H} \tilde{X}_{1,2}] A_{1,2}) = \diag(I_k - \tilde{X}_{1,2}^{H} \tilde{X}_{1,2}) \cdot [\lambda_1, \lambda_2, \ldots, \lambda_k]^H
\]

\[
\leq \sum_{i=1}^{j} \lambda_i \sin^2 \theta_{j-i+1}(U_k, X_j) + \sum_{i=j+1}^{k} \lambda_i
\]

\[
= 9
\]
\[ \sum_{i=1}^{j} \lambda_{j-i+1} \sin^2 \theta_i(U_k, X_j) + \sum_{i=j+1}^{k} \lambda_i, \]

\[ \text{trace}(\tilde{X}_3 \tilde{X}_3^H A_3) = \text{diag}(\tilde{X}_3 \tilde{X}_3^H) \cdot [\lambda_{k+1}, \lambda_{k+2}, \ldots, \lambda_N]^H \]

\[ \geq \sum_{i=1}^{\min\{N-k,j\}} \lambda_{k+i} \sin^2 \theta_i(U_k, X_j). \]

It follows from (4.4) that

\[ \sum_{i=1}^{j} (\tilde{\lambda}_i - \lambda_i) \geq - \sum_{i=1}^{j} \lambda_{j-i+1} \sin^2 \theta_i(U_k, X_j) + \sum_{i=1}^{\min\{N-k,j\}} \lambda_{k+i} \sin^2 \theta_i(U_k, X_j) \]

\[ = \sum_{i=1}^{\min\{N-k,j\}} (\lambda_{k+i} - \lambda_{j-i+1}) \sin^2 \theta_i(U_k, X_j). \]

The last equality holds because of \( \sin^2 \theta_i(U_k, X_j) = 0 \) if \( k \geq j \geq i > N - k \) by Proposition 2.2. This is (3.14a). For (3.14b), we notice that \( \omega_{k+i} - \omega_{j-i+1} \geq \omega_{k+1} - \omega_j \) for \( 1 \leq i \leq j \leq k \).

\section{5 Numerical examples}

In this section, we present some numerical examples to illustrate our results. In particular, we will demonstrate that our lower bounds are preferred to (3.11), especially in case of tiny \( \lambda_{k+1} - \lambda_k \) or a tight cluster. In the examples below, without loss of generality, we take

\[ A = \text{diag}(\lambda_1, \ldots, \lambda_N). \]

Thus the eigenvector matrix \( U = I_N \). Let \( N - k \geq k \) for simplicity.

We shall focus on comparing our bounds (3.15a), (3.15b), (3.16b), (3.18) and the existing result (3.11). For this reason, we will measure the following errors, for \( j = 1, \ldots, k \),

\[ \varepsilon_{1,j} = \sum_{i=1}^{j} (\tilde{\lambda}_i - \lambda_i) \leq \varepsilon_{2,j} = \sum_{i=1}^{j} (\tilde{\lambda}_i - \lambda_t), \quad (5.1a) \]

\[ \varepsilon_{3,j} = \sum_{i=1}^{j} (\lambda_{k+i} - \lambda_{j-i+1}) \sin^2 \theta_i(U_k, X_j), \quad (5.1b) \]

\[ \varepsilon_{4,j} = (\lambda_{k+1} - \lambda_j) \sum_{i=1}^{j} \sin^2 \theta_i(U_k, X_j), \quad (5.1c) \]

\[ \varepsilon_{5,j} = (\lambda_{k+1} - \lambda_k) \sum_{i=1}^{j} \sin^2 \theta_i(U_k, X_j), \quad (5.1d) \]

where \( \{t_i \mid 1 \leq i \leq k\} \) is the permutation of \( \{1, 2, \ldots, k\} \) as determined by (3.17). To emphasize our comparison between our results and the well-known (3.11), we also introduce

\[ \varepsilon_6 = (\lambda_{k+1} - \lambda_k) \sum_{i=1}^{k} \sin^2 \theta_i(U_k, X) \quad (5.2) \]
In this example, we make up a matrix such that

\[ U \text{ eigenvalues by the angles from} \]

We seek to bound the difference between the first five smallest Ritz values and the targeted \( X \) there is a subspace of \( X \) in Section 1, usually, \( X \) use the Lanczos method [11] to generate the subspace All “Example 5.2.

Example 5.1. In this example, we make up a matrix such that \( \lambda_{k+1} - \lambda_k \) is small. Let \( N = 10, \ell = k = 5 \),

\[
\lambda_1 = 0.0826, \ \lambda_2 = 0.4116, \ \lambda_3 = 0.5118, \ \lambda_4 = 0.5518, \ \lambda_5 = 0.5835, \\
\lambda_6 = 0.5836, \ \lambda_7 = 0.6026, \ \lambda_8 = 0.7196, \ \lambda_9 = 0.7505, \ \lambda_{10} = 0.9962,
\]

and \( X = \mathcal{R}(X) \), where \( X \) is an \( N \times k \) orthonormal matrix generated by \( \text{qr}((\text{rand}(N, k), 0) \) in MATLAB. For this example, \( \lambda_{k+1} - \lambda_k = 10^{-4} \), potentially bad news for \( \varepsilon_6 = \varepsilon_{4,k} = \varepsilon_{5,k} \) as lower bounds for \( \varepsilon_{1,k} = \varepsilon_{2,k} \). Table 5.1 reports the lower bounds \( \varepsilon_{3,j}, \varepsilon_{4,j} \) and \( \varepsilon_{5,j} \) as defined in (5.1b), (5.1c), (5.1d), and (5.2) on the differences \( \varepsilon_{1,j} \) and \( \varepsilon_{2,j} \) as defined in (5.1a) between Ritz values and the targeted eigenvalues. Our bound \( \varepsilon_{3,j} \) for \( j = 1, \ldots, k \) clearly stands out as the best. In particular, \( \varepsilon_{3,5} \) is three magnitudes sharper than \( \varepsilon_6 = \varepsilon_{4,5} = \varepsilon_{5,5} \). Overall \( \varepsilon_{4,j} \) is very good, too, except for \( j = 5 \).

\[
\varepsilon_{3,1} = \varepsilon_{4,1}, \ \varepsilon_6 \leq \varepsilon_{4,k} = \varepsilon_{5,k} \quad (\text{becoming equality when } \ell = k).
\]

All \( \varepsilon_{3,j}, \varepsilon_{4,j} \), and \( \varepsilon_{5,j} \) are lower bounds of \( \varepsilon_{1,j} \) and thus also of \( \varepsilon_{2,j} \) for \( j = 1, \ldots, k \). Consequently, \( \varepsilon_6 \) is, as a classical one, also a lower bound of \( \varepsilon_{1,k} \) and \( \varepsilon_{2,k} \).

Example 5.2. We take \( N = 600 \) and

\[
\lambda_1 = -3, \ \lambda_2 = -2.5, \ \lambda_3 = -2, \ \lambda_i = \frac{i - 4}{N}, \ i = 4, \ldots, N.
\]

There are two eigenvalue clusters: \( \{\lambda_1, \lambda_2, \lambda_3\} \) and \( \{\lambda_4, \ldots, \lambda_N\} \). In this example, we use the Lanczos method [11] to generate the subspace \( X \). The 10-step Lanczos process generates a Krylov subspace \( X = \text{span}\{v_0, A v_0, \ldots, A^9 v_0\} \), where \( v_0 \) is an initial vector obtained by the MATLAB built-in function \( \text{rand}(N, 1) \). So \( \ell = \dim(X) = 10 \). As explained in Section 1, usually, \( X \) as a whole is not close to any invariant subspace, but more likely there is a subspace of \( X \) which is close to an invariant subspace. Again, we set \( k = 5 \) and seek to bound the difference between the first five smallest Ritz values and the targeted eigenvalues by the angles from \( U_k \) to \( X \).

We computed \( \varepsilon_{1,j}, \varepsilon_{2,j}, \varepsilon_{3,j}, \varepsilon_{4,j}, \) and \( \varepsilon_{5,j} \) for \( j = 1, \ldots, k = 5 \) in Table 5.2. We also computed \( \varepsilon_6 = 0.30 \cdot 10^{-2} \) which is smaller than \( \varepsilon_{4,k} = \varepsilon_{5,k} \) (but not by much), as expected since \( \ell > k \). Table 5.2 suggests that our bounds \( \varepsilon_{3,j} \) and \( \varepsilon_{4,j} \) are very sharp for \( j = 1, 2, 3 \) for the example. In fact, the first three eigenvalues are very well-approximated by the corresponding Ritz values. But \( \varepsilon_{5,j} \) is not so good because of the factor \( \lambda_{k+1} - \lambda_k = 1/600 \).

All \( \varepsilon_{3,j}, \varepsilon_{4,j} \) and \( \varepsilon_{5,j} \) for \( j = 4, 5 \) are about the same.

\[
\begin{array}{cccccc}
\hline
j & \varepsilon_{1,j} & \varepsilon_{2,j} & \varepsilon_{3,j} & \varepsilon_{4,j} & \varepsilon_{5,j} \\
\hline
1 & 0.30 & 0.30 & 0.14 \cdot 10^{-1} & 0.14 \cdot 10^{-1} & 0.11 \cdot 10^{-5} \\
2 & 0.39 & 0.57 & 0.47 \cdot 10^{-1} & 0.47 \cdot 10^{-1} & 0.10 \cdot 10^{-4} \\
3 & 0.47 & 0.67 & 0.11 & 0.78 \cdot 10^{-1} & 0.41 \cdot 10^{-4} \\
4 & 0.54 & 0.75 & 0.16 & 0.56 \cdot 10^{-1} & 0.66 \cdot 10^{-4} \\
5 & 0.82 & 0.82 & 0.23 & 0.10 \cdot 10^{-3} & 0.10 \cdot 10^{-3} \\
\hline
\end{array}
\]
Table 5.2: $\varepsilon_{1,j}$, $\varepsilon_{2,j}$, $\varepsilon_{3,j}$, $\varepsilon_{4,j}$ and $\varepsilon_{5,j}$ of Example 5.2

<table>
<thead>
<tr>
<th>$j$</th>
<th>$\varepsilon_{1,j}$</th>
<th>$\varepsilon_{2,j}$</th>
<th>$\varepsilon_{3,j}$</th>
<th>$\varepsilon_{4,j}$</th>
<th>$\varepsilon_{5,j}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$0.26 \cdot 10^{-10}$</td>
<td>$0.14$</td>
<td>$0.22 \cdot 10^{-10}$</td>
<td>$0.22 \cdot 10^{-10}$</td>
<td>$0.12 \cdot 10^{-13}$</td>
</tr>
<tr>
<td>2</td>
<td>$0.91 \cdot 10^{-9}$</td>
<td>$0.16$</td>
<td>$0.72 \cdot 10^{-9}$</td>
<td>$0.72 \cdot 10^{-9}$</td>
<td>$0.48 \cdot 10^{-12}$</td>
</tr>
<tr>
<td>3</td>
<td>$0.18 \cdot 10^{-6}$</td>
<td>$0.16$</td>
<td>$0.14 \cdot 10^{-6}$</td>
<td>$0.14 \cdot 10^{-6}$</td>
<td>$0.12 \cdot 10^{-9}$</td>
</tr>
<tr>
<td>4</td>
<td>$0.20 \cdot 10^{-1}$</td>
<td>$0.16$</td>
<td>$0.28 \cdot 10^{-2}$</td>
<td>$0.28 \cdot 10^{-2}$</td>
<td>$0.14 \cdot 10^{-2}$</td>
</tr>
<tr>
<td>5</td>
<td>$0.16$</td>
<td>$0.16$</td>
<td>$0.58 \cdot 10^{-2}$</td>
<td>$0.31 \cdot 10^{-2}$</td>
<td>$0.31 \cdot 10^{-2}$</td>
</tr>
</tbody>
</table>

6 Concluding Remarks

Knyazev and Argentati [5] established several elegant majorization results on the differences between Ritz values and the eigenvalues of interest of a (large scale) Hermitian matrix $A$ by the canonical angles between/from an invariant subspace of $A$ and/to a subspace built usually for the purpose of computing the eigenvalues of interest and their corresponding eigenvectors. With these majorization results, bounds are readily available to tell how accurate (some of) the Ritz values may be, given information on the canonical angles. Our main results can be considered as converses to their results in the sense that our results are about bounding the canonical angles by the differences between Ritz values and the eigenvalues of interest, and bounds are readily available from our results to tell how accurate (some of) the Ritz vectors may be, given information on approximation accuracies in the Ritz values.

Recently, there is an increasing trend of using optimization techniques for solving today’s ever-growing large scale symmetric eigenvalue problems through optimizing Rayleigh quotient matrices because these optimization techniques turn to be more memory and computationally efficient – less memory and more matrix-matrix multiplications. The idea is to force Ritz values move towards eigenvalues. The result here is helpful in understanding how Ritz vectors move towards eigenvectors while Ritz values are made to move towards eigenvalues.

So far we have made the smallest eigenvalues the focus of our investigation. In the case when the largest eigenvalues are the interested ones, one can simply consider $-A$ instead.

References


