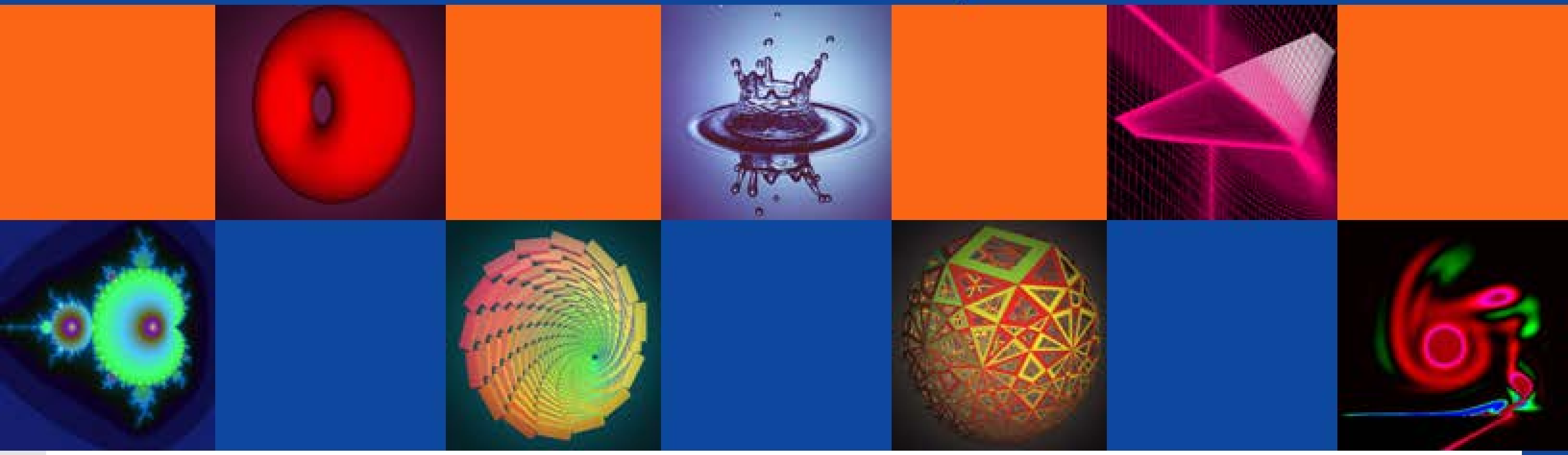
The University of Texas ARLINGTON

MATHEMATICS

Preprint Series



Backward Perturbation Analysis And Residual-based Error Bounds for the Linear Response Eigenvalue Problem

Lei-Hong Zhang

Wen-Wei Lin

Ren-Cang Li

Technical Report 2014-03

Backward Perturbation Analysis and Residual-based Error Bounds For the Linear Response Eigenvalue Problem

Lei-Hong Zhang[*] Wen-Wei Lin[†] Ren-Cang Li [‡]

February 4, 2014

Abstract

The numerical solution of a large scale linear response eigenvalue problem is often accomplished by computing a pair of deflating subspaces associated with the interested part of the spectrum. This paper is concerned with the backward perturbation analysis for a given pair of approximate deflating subspaces or an approximate eigen-quaternary. Various optimal backward perturbation bounds are obtained, as well as bounds for approximate eigenvalues computed through the pair of approximate deflating subspaces or approximate eigen-quaternary. These results are reminiscent of many existing classical ones for the standard eigenvalue problem.

Key words. Linear response eigenvalue problem, eigenvalue approximation, Rayleigh-Ritz approximation, backward perturbation, error bound, deflating subspace

AMS subject classifications. 65L15, 65F15, 81Q15, 15A18, 15A42

1 Introduction

In this paper, we are concerned with a backward perturbation analysis and residual-based error bounds for the *Linear Response Eigenvalue Problem* (LREP):

$$Hz := \begin{bmatrix} & K \\ M & \end{bmatrix} \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix} =: \lambda \boldsymbol{z}, \tag{1.1}$$

[*]Department of Applied Mathematics, Shanghai University of Finance and Economics, 777 Guoding Road, Shanghai 200433, People's Republic of China. Email: `longzlh@163.com`. Supported in part by the National Natural Science Foundation of China NSFC-11101257, NSFC-11371102, and the Basic Academic Discipline Program, the 11th five year plan of 211 Project for Shanghai University of Finance and Economics. Part of this work is done while this author is a visiting scholar at the Department of Mathematics, University of Texas at Arlington from February 2013 to January 2014.

[†]Department of Applied Mathematics, National Chiao Tung University, No.1001 University Road, Hsinchu 30013, Taiwan. Email: `wwlin@math.nctu.edu.tw`. Supported in part by National Center of Theoretical Science, Taiwan and ST Yau Center, Chiao Tung University, Taiwan.

[‡]Department of Mathematics, University of Texas at Arlington, P.O. Box 19408, Arlington, TX 76019-0408, USA. Email: `rcli@uta.edu`. Supported in part by NSF grants DMS-1115834 and DMS-1317330, by a Research Gift Grant from Intel Corporation, and by National Center of Theoretical Science, Taiwan while this author visited in December 2013.

where K and M are both n-by-n real symmetric and positive definite. The matrix H in (1.1) is a special Hamiltonian matrix whose eigenvalues are real [1] and come in pairs $\{\lambda, -\lambda\}$. Therefore, we can order the $2n$ eigenvalues of (1.1) as

$$-\lambda_n \leq \cdots \leq -\lambda_1 < \lambda_1 \leq \cdots \leq \lambda_n. \tag{1.2}$$

LREP (1.1) is mathematically equivalent to the so-called *random phase approximation* (RPA) eigenvalue problem in computational quantum chemistry and physics:

$$\begin{bmatrix} A & B \\ -B & -A \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix}, \tag{1.3}$$

where $A, B \in \mathbb{R}^{n\times n}$ are both symmetric matrices and $\begin{bmatrix} A & B \\ B & A \end{bmatrix}$ is positive definite. The equivalent relationship is established through the orthogonal matrix $J = \frac{1}{\sqrt{2}}\begin{bmatrix} I_n & I_n \\ I_n & -I_n \end{bmatrix}$ and the similarity transformation (see, e.g., [1, 2])

$$J^{\mathrm{T}} \begin{bmatrix} A & B \\ -B & -A \end{bmatrix} J = \begin{bmatrix} & A-B \\ A+B & \end{bmatrix} =: \begin{bmatrix} & K \\ M & \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix} := J^{\mathrm{T}} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix}. \tag{1.4}$$

RPA is one of the most widely used methods in studying the excitation states (energies) of physical systems in the study of collective motion of many-particle systems [1, 30, 31] which has applications in silicon nanoparticles, nanoscale materials, analysis of interstellar clouds [1, 2], among others. The heart of RPA calculations is to compute a few eigenpairs associated with the smallest *positive* eigenvalues, which, by the equivalent relationship (1.4) , are the eigenpairs associated with the eigenvalues $\lambda_1 \leq \cdots \leq \lambda_k$ of (1.1).

As the dimension n is usually very large, LREP is generally solved by iterative methods. Roughly speaking, any large scale eigenvalue computation is about approximating certain invariant subspaces associated with the interested part of the spectrum, and the interested eigenvalues are then extracted from projecting the problem by approximate invariant subspaces into much smaller eigenvalue problems. In the case of the linear response eigenvalue problem, it is the *pair of deflating subspaces* associated with the first few smallest positive eigenvalues that needs to be computed [2].

For two k-dimensional subspaces $\mathcal{U}$ and $\mathcal{V}$ in $\mathbb{R}^n$, we call $\{\mathcal{U}, \mathcal{V}\}$ a *pair of deflating subspaces* of $\{K, M\}$ if

$$K\mathcal{U} \subseteq \mathcal{V} \quad \text{and} \quad M\mathcal{V} \subseteq \mathcal{U}. \tag{1.5}$$

This notion of the pair of deflating subspaces is a generalization of the concept of the invariant subspace (or, eigenspace) in the standard eigenvalue problem upon considering the special structure in LREP (1.1) [1]. Whenever such a pair of deflating subspaces is available, we can project LREP (1.1) into a much smaller problem in the form of (1.1), an LREP by its own, whose spectrum are a part of that of H (see more discussions in section 2 and [1, 2]). Based on this fact, several efficient algorithms, including the Locally Optimal Block Preconditioned 4D Conjugate Gradient Method (LOBP4DCG) [2], the block Chebyshev-Davidson method [29], as well as the generalized Lanczos method [28, 32, 33], have been proposed. Each of these algorithms generates a sequence of approximate deflating subspace pairs that hopefully converge to or contain subspaces near the pair of deflating subspaces. The goal of this paper is to perform a backward perturbation analysis and to establish error bounds on the accuracy (in eigenvalue/eigenspace approximations) using proper residuals associated with any given approximate deflating subspace pair.

A related study is presented in [34]. The main difference among the results in [34] and those in the present paper is that the error bounds on eigenvalue/eigenspace approximations in [34] are characterized by the canonical angles between the approximate deflating subspace pair and the exact pair, whereas the error bounds in this paper use certain computable residuals. These two types of error bounds are well-established in the standard eigenvalue problem (see, e.g., [20, 22]), and both types are useful for analyzing the convergence and designing stopping criteria for iterative algorithms.

The rest of the paper is organized as follows. In section 2, we will state some basic properties about LREP for use later. Section 3 gives a backward perturbation analysis for a given pair of approximate deflating subspaces or an approximate eigen-quaternary, optimizes backward perturbation errors, and shows the near optimality of the so-called Rayleigh quotient pair. Section 4 derives several error bounds in terms of residuals on eigenvalue approximations. In section 5, we review related results for the standard eigenvalue problem as a comparison. Finally in section 6, we present our concluding remarks.

Notation. $\mathbb{K}^{n\times m}$ is the set of all $n \times m$ matrices whose entries belong to the number field $\mathbb{K}$, $\mathbb{K}^n = \mathbb{K}^{n\times 1}$, and $\mathbb{K} = \mathbb{K}^1$, where $\mathbb{K} = \mathbb{R}$ (the set of real numbers) or $\mathbb{C}$ (the set of complex numbers). I_n (or simply I if its dimension is clear from the context) denotes the $n \times n$ identity matrix. All vectors are column vectors and are in boldface. For a matrix Z,

1. Z^{T} and Z^{H} denote its transpose and the conjugate transpose, respectively,
2. $\mathcal{R}(Z)$ is Z's column space, spanned by its column vectors,
3. $Z^{\dagger}$ stands for the Moore-Penrose inverse and $P_Z = ZZ^{\dagger}$ is the orthogonal projection onto $\mathcal{R}(Z)$ and $P_Z^{\perp} = I - P_Z = Z^{\dagger}Z$ [22],
4. $\|Z\|_2$, $\|Z\|_{\mathrm{F}}$, and $\|Z\|_{\mathrm{ui}}$ are the spectral norm, the Frobenius norm, and a general unitarily invariant norm, respectively,
5. Z's submatrices $Z_{(k:\ell,i:j)}$, $Z_{(k:\ell,:)}$, and $Z_{(:,i:j)}$ 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,
6. when Z is a square matrix, its trace is $\mathrm{trace}(Z)$ and its eigenvalue set is $\mathrm{eig}(Z)$.

The assignments in (1.1) will be assumed, namely H is always defined that way for given K, $M \in \mathbb{R}^{n\times n}$ which are assumed by default to be symmetric positive semi-definite and one of which is definite, unless explicitly stated differently.

2 Preliminaries

Many theoretical properties of LREP have been established in [1, 2]. In Theorem 2.1, we present certain decompositions on K and M, necessary for our developments later in this paper. The reader is referred to [1, section 2] for proofs and more.

Theorem 2.1. *Suppose that K is semidefinite and M is definite. Then the following statements are true:*

(i) There exists a nonsingular $\Phi \in \mathbb{R}^{n\times n}$ such that

$$K = \Psi \Lambda^2 \Psi^{\mathrm{T}} \quad \textit{and} \quad M = \Phi\Phi^{\mathrm{T}}, \tag{2.1}$$

where $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$, λ_i are as in (1.2), and $\Psi = \Phi^{-\mathrm{T}}$.

(ii) If K is also definite, then all $\lambda_i > 0$ and H is diagonalizable:

$$H \begin{bmatrix} \Psi\Lambda & \Psi\Lambda \\ \Phi & -\Phi \end{bmatrix} = \begin{bmatrix} \Psi\Lambda & \Psi\Lambda \\ \Phi & -\Phi \end{bmatrix} \begin{bmatrix} \Lambda & \\ & -\Lambda \end{bmatrix}. \tag{2.2}$$

As we have introduced in section 1, for two given k-dimensional subspaces $\mathcal{U} \subseteq \mathbb{R}^n$ and $\mathcal{V} \subseteq \mathbb{R}^n$, the pair $\{\mathcal{U}, \mathcal{V}\}$ is called a *pair of deflating subspaces* of $\{K, M\}$ if

$$K\mathcal{U} \subseteq \mathcal{V} \quad \text{and} \quad M\mathcal{V} \subseteq \mathcal{U} \tag{1.5}$$

hold. This definition is essentially the same as the existing ones for the product eigenvalue problem [3, 7, 18, 19]. Let $U \in \mathbb{R}^{n\times k}$ and $V \in \mathbb{R}^{n\times k}$ be the basis matrices for $\mathcal{U}$ and $\mathcal{V}$, respectively. Alternatively, (1.5) can be restated as that there exist $K_{\mathrm{R}} \in \mathbb{R}^{k\times k}$ and $M_{\mathrm{R}} \in \mathbb{R}^{k\times k}$ such that

$$KU = VK_{\mathrm{R}} \quad \text{and} \quad MV = UM_{\mathrm{R}} \tag{2.3}$$

and vice versa, or equivalently,

$$H \begin{bmatrix} V & \\ & U \end{bmatrix} = \begin{bmatrix} V & \\ & U \end{bmatrix} H_{\mathrm{R}} \quad \text{with} \quad H_{\mathrm{R}} := \begin{bmatrix} & K_{\mathrm{R}} \\ M_{\mathrm{R}} & \end{bmatrix},$$

i.e., $\mathcal{V} \oplus \mathcal{U}$ is an invariant subspace of H [1, Theorem 2.4]. We call $\{U, V, K_{\mathrm{R}}, M_{\mathrm{R}}\}$ an *eigen-quaternary* of $\{K, M\}$.

Whenever a pair of deflating subspaces $\{\mathcal{R}(U), \mathcal{R}(V)\}$ is at hand, a part of the eigenpairs of H can be obtained via solving the smaller eigenvalue problem [1, Theorem 2.5]: if

$$H_{\mathrm{R}}\hat{\boldsymbol{z}} := \begin{bmatrix} & K_{\mathrm{R}} \\ M_{\mathrm{R}} & \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{y}} \\ \hat{\boldsymbol{x}} \end{bmatrix} = \lambda \begin{bmatrix} \hat{\boldsymbol{y}} \\ \hat{\boldsymbol{x}} \end{bmatrix} =: \lambda\hat{\boldsymbol{z}}, \tag{2.4}$$

then $(\lambda, \begin{bmatrix} V\hat{\boldsymbol{y}} \\ U\hat{\boldsymbol{x}} \end{bmatrix})$ is an eigenpair of H. The matrix H_{R} is the restriction of H onto $\mathcal{V} \oplus \mathcal{U}$ with respect to the basis matrices $V \oplus U$. Moreover, the eigenvalues of H_{R} are uniquely determined by the pair of deflating subspaces $\{\mathcal{U}, \mathcal{V}\}$; in the other word, different choices of the basis matrices for $\{\mathcal{U},\mathcal{V}\}$ result in the same eigenvalues. In fact, if $\widehat{U}=UD_1\in\mathbb{R}^{n\times k}$ and $\widehat{V}=VD_2\in\mathbb{R}^{n\times k}$ are new basis matrices for $\mathcal{R}(U)$ and $\mathcal{R}(V)$, respectively, and

$$K\widehat{U}=\widehat{V}\widehat{K}_{\mathrm{R}} \quad\text{and}\quad M\widehat{V}=\widehat{U}\widehat{M}_{\mathrm{R}}, \tag{2.5}$$

then $\widehat{K}_{\mathrm{R}}=D_2^{-1}K_{\mathrm{R}}D_1$ and $\widehat{M}_{\mathrm{R}}=D_1^{-1}K_{\mathrm{R}}D_2$ by comparing (2.3) to (2.5) after substituting in $\widehat{U}=UD_1$ and $\widehat{V}=VD_2$. Thus

$$\widehat{H}_{\mathrm{R}}:=\begin{bmatrix} & \widehat{K}_{\mathrm{R}}\\ \widehat{M}_{\mathrm{R}} & \end{bmatrix}=\begin{bmatrix}D_2 & \\ & D_1\end{bmatrix}^{-1}H_{\mathrm{R}}\begin{bmatrix}D_2 & \\ & D_1\end{bmatrix}. \tag{2.6}$$

Evidently $\widehat{H}_{\mathrm{R}}$ and H_{R} must have the same eigenvalues.

Two particular choices of $\{K_{\mathrm{R}},M_{\mathrm{R}}\}$ to satisfy (2.3) are

$$K_{\mathrm{R}}=(U^{\mathrm{T}}V)^{-1}U^{\mathrm{T}}KU,\quad M_{\mathrm{R}}=(V^{\mathrm{T}}U)^{-1}V^{\mathrm{T}}MV; \tag{2.7a}$$

$$K_{\mathrm{R}}=(V^{\mathrm{T}}V)^{-1}V^{\mathrm{T}}KU,\quad M_{\mathrm{R}}=(U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}MV. \tag{2.7b}$$

In (2.7a), it is assumed $U^{\mathrm{T}}V$ is invertible, which is guaranteed if one of K and M is definite [1, Lemma 2.7]. By what we just proved, the associated H_{R} with either (2.7a) or (2.7b) must have the same eigenvalues.

In practical computations, however, $\{\mathcal{U},\mathcal{V}\}$ is usually a pair of approximate deflating spaces, i.e., no $\{K_{\mathrm{R}},M_{\mathrm{R}}\}$ that satisfies (2.3) exists. Dependent on how good $\{\mathcal{U},\mathcal{V}\}$ is as a pair of approximate deflating spaces, the equations in (2.3) is satisfied approximately to an appropriate level for some $\{K_{\mathrm{R}},M_{\mathrm{R}}\}$ like the ones given in (2.7). In particular, $\{K_{\mathrm{R}},M_{\mathrm{R}}\}$ by (2.7a) relates to the structure-preserving projection H_{SR} of H in [2, (2.2)] that plays an important role numerically there. To highlight this particular pair, we will call $\{K_{\mathrm{R}},M_{\mathrm{R}}\}$ by (2.7a) a *Rayleigh quotient pair* of LREP (1.1) associated with $\{\mathcal{R}(U),\mathcal{R}(V)\}$ and introduce

$$K_{\mathrm{RQ}}:=(U^{\mathrm{T}}V)^{-1}U^{\mathrm{T}}KU,\quad M_{\mathrm{RQ}}:=(V^{\mathrm{T}}U)^{-1}V^{\mathrm{T}}MV \tag{2.8}$$

for the ease of future references. Both K_{RQ} and M_{RQ} vary with different selections of U and V as the basis matrices of $\mathcal{R}(U)$ and $\mathcal{R}(V)$, respectively. But the eigenvalues of the induced

$$H_{\mathrm{RQ}}=\begin{bmatrix} & K_{\mathrm{RQ}}\\ M_{\mathrm{RQ}} & \end{bmatrix}. \tag{2.9}$$

do not. In fact, with new basis matrices $\widehat{U}=UD_1$ and $\widehat{V}=VD_2$ and, accordingly, new $\widehat{K}_{\mathrm{RQ}}$ and $\widehat{M}_{\mathrm{RQ}}$, $\widehat{H}_{\mathrm{RQ}}$ is similar to H_{RQ} (an equation like (2.6) holds).

For the definition and properties of unitarily invariant norms, the reader is referred to [4, 22] for details. In this article, for convenience, any $\|\cdot\|_{\mathrm{ui}}$ we use is generic to matrix sizes in the sense that it applies to matrices of all sizes. Examples include the matrix spectral norm $\|\cdot\|_2$ and the Frobenius norm $\|\cdot\|_{\mathrm{F}}$. Two important properties of unitarily invariant norms are

$$\|X\|_2\le\|X\|_{\mathrm{ui}},\quad \|XYZ\|_{\mathrm{ui}}\le\|X\|_2\cdot\|Y\|_{\mathrm{ui}}\cdot\|Z\|_2 \tag{2.10}$$

for any matrices X, Y, and Z of compatible sizes.

3 Backward Errors and Optimal Residuals

Let $\mathcal{U} \subseteq \mathbb{R}^n$ and $\mathcal{V} \subseteq \mathbb{R}^n$ be two k-dimensional subspaces, and let $U \in \mathbb{R}^{n\times k}$ and $V \in \mathbb{R}^{n\times k}$ be the basis matrices for $\mathcal{U}$ and $\mathcal{V}$, respectively. As discussed in section 2, $\{\mathcal{U}, \mathcal{V}\}$ is a pair of deflating subspaces of $\{K, M\}$ if and only if the equations in (2.3) holds for some $K_{\rm R} \in \mathbb{R}^{k\times k}$ and $M_{\rm R} \in \mathbb{R}^{k\times k}$. In this case, $\{U, V, K_{\rm R}, M_{\rm R}\}$ is an eigen-quaternary.

But in practice, $\{\mathcal{U}, \mathcal{V}\}$ is likely a pair of approximate deflating subspaces in the sense that the *residuals*

$$\mathscr{R}_K(K_{\rm R}) := KU - VK_{\rm R}, \quad \mathscr{R}_M(M_{\rm R}) := MV - UM_{\rm R} \tag{3.1}$$

are tiny in norm for some $K_{\rm R} \in \mathbb{R}^{k\times k}$ and $M_{\rm R} \in \mathbb{R}^{k\times k}$. In this case, $\{U, V, K_{\rm R}, M_{\rm R}\}$ is an approximate eigen-quaternary. Set

$$H_{\rm R} = \begin{bmatrix} & K_{\rm R} \\ M_{\rm R} & \end{bmatrix}, \tag{3.2}$$

associated with such $K_{\rm R}$ and $M_{\rm R}$. Different from $\{\mathcal{U}, \mathcal{V}\}$ being exact, now $\operatorname{eig}(H_{\rm R}) \not\subset \operatorname{eig}(H)$ but hopefully some or all eigenvalues of $H_{\rm R}$ are good approximations to some eigenvalues of H. Naturally if

$$H_{\rm R}\hat{\boldsymbol{z}} := \begin{bmatrix} & K_{\rm R} \\ M_{\rm R} & \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{y}} \\ \hat{\boldsymbol{x}} \end{bmatrix} = \lambda \begin{bmatrix} \hat{\boldsymbol{y}} \\ \hat{\boldsymbol{x}} \end{bmatrix} =: \lambda\hat{\boldsymbol{z}},$$

we may take $(\lambda, \begin{bmatrix} V\hat{\boldsymbol{y}} \\ U\hat{\boldsymbol{x}} \end{bmatrix})$ as an approximate eigenpair of H [1, 2] in view of our discussions in the previous section.

In this section, we are interested in answering the following three questions:

1. Given an approximate eigen-quaternary, what are the smallest symmetric perturbations $\varDelta K$ and $\varDelta M$ (to K and M, respectively) in norm such that the given eigen-quaternary is an exact eigen-quaternary of $\{K + \varDelta K, M + \varDelta M\}$?

2. Given a pair of approximate deflating subspaces $\{\mathcal{U}, \mathcal{V}\}$, what are the smallest residuals $\mathscr{R}_K(K_{\rm R})$ and $\mathscr{R}_M(M_{\rm R})$ in norm optimizing among all possible $K_{\rm R}$ and $M_{\rm R}$?

3. It turns out that the so-called Rayleigh quotient pair $\{K_{\rm RQ}, M_{\rm RQ}\}$ is not the one that minimizes $\mathscr{R}_K(K_{\rm R})$ and $\mathscr{R}_M(M_{\rm R})$ in norm. But how far are $K_{\rm RQ}$ and $M_{\rm RQ}$ from their optimal counterparts?

3.1 Optimal Backward Errors

In this subsection, we shall investigate the first question raised at the beginning of the section.

Throughout this subsection, $\{U, V, K_{\rm R}, M_{\rm R}\}$ is assumed an approximate eigen-quaternary of $\{K, M\}$ with U, $V \in \mathbb{R}^{n\times k}$ satisfying

$$U^{\rm T}U = V^{\rm T}V = I_k, \quad \text{and} \quad \operatorname{rank}(U^{\rm T}V) = k, \tag{3.3}$$

$K_{\mathrm{R}}, M_{\mathrm{R}} \in \mathbb{R}^{k\times k}$. Define $\mathscr{R}_K(K_{\mathrm{R}})$ and $\mathscr{R}_M(M_{\mathrm{R}})$ by (3.1), H_{R} as in (3.2), and

$$S_K := (U^{\mathrm{T}}V)K_{\mathrm{R}}, \quad S_M := (V^{\mathrm{T}}U)M_{\mathrm{R}}. \tag{3.4}$$

Lemma 3.1. *Factorize $U^{\mathrm{T}}V$ as $U^{\mathrm{T}}V = W_1^{\mathrm{T}}W_2$, where $W_i \in \mathbb{R}^{k\times k}$ are nonsingular. Then*

$$H_{\mathrm{R}} = [W_2 \oplus W_1]^{-1} \begin{bmatrix} 0 & W_1^{-\mathrm{T}}S_K W_1^{-1} \\ W_2^{-\mathrm{T}}S_M W_2^{-1} & 0 \end{bmatrix} [W_2 \oplus W_1]. \tag{3.5}$$

In the case when $\{U, V, K_{\mathrm{R}}, M_{\mathrm{R}}\}$ is an exact eigen-quaternary of $\{K, M\}$,

$$S_K = U^{\mathrm{T}}KU, \quad S_M = V^{\mathrm{T}}MV. \tag{3.6}$$

Proof. The equation (3.5) can be verified straightforwardly after substituting in S_K and S_M as given by (3.4). When $\{U, V, K_{\mathrm{R}}, M_{\mathrm{R}}\}$ is exact, we can take K_{R} and M_{R} as in (2.7a). Now use (3.4) to see (3.6). □

Perturbations ΔK and ΔM (to K and M, respectively) such that the given eigen-quaternary $\{U, V, K_{\mathrm{R}}, M_{\mathrm{R}}\}$ is an exact eigen-quaternary of $\{K + \Delta K, M + \Delta M\}$ are the ones that satisfy

$$(K + \Delta K)U = VK_{\mathrm{R}}, \quad (M + \Delta M)V = UM_{\mathrm{R}}. \tag{3.7}$$

Since K and M are symmetric, we further restrict ΔK and ΔM to be symmetric, too. The first and foremost question is, naturally, if such perturbations ΔK and ΔM exist, and then if they do, what the smallest perturbations in norm are. For this purpose, we define

$$\mathbb{L} := \big\{(\Delta K, \Delta M) \,:\, \Delta K^{\mathrm{T}} = \Delta K \in \mathbb{R}^{n\times n}, \Delta M^{\mathrm{T}} = \Delta M \in \mathbb{R}^{n\times n} \text{ satisfying (3.7)}\big\}, \tag{3.8}$$

and investigate when $\mathbb{L} \neq \varnothing$.

Lemma 3.2 ([25, Lemma 1.4])**.** *Given $Z_1, Z_2 \in \mathbb{C}^{n\times k}$, define*

$$\mathbb{S} = \{S \in \mathbb{C}^{n\times n} : S^{\mathrm{H}} = S,\ SZ_1 = Z_2\}.$$

1. *$\mathbb{S} \neq \varnothing$ if and only if Z_1 and Z_2 satisfy*

$$Z_2 P_{Z_1^{\mathrm{H}}} = Z_2 \quad \text{and} \quad (P_{Z_1}Z_2Z_1^{\dagger})^{\mathrm{H}} = P_{Z_1}Z_2Z_1^{\dagger}.$$

2. *In the case of $\mathbb{S} \neq \varnothing$, any $S \in \mathbb{S}$ can be expressed by*

$$S = Z_2Z_1^{\dagger} + (Z_1^{\dagger})^{\mathrm{H}}Z_2^{\mathrm{H}} - (Z_1^{\dagger})^{\mathrm{H}}Z_2^{\mathrm{H}}P_{Z_1} + P_{Z_1}^{\perp}TP_{Z_1}^{\perp},$$

where $T \in \mathbb{C}^{n\times n}$ is Hermitian and arbitrary. Moreover,

$$S_{\mathrm{opt}} = Z_2Z_1^{\dagger} + (Z_1^{\dagger})^{\mathrm{H}}Z_2^{\mathrm{H}} - (Z_1^{\dagger})^{\mathrm{H}}Z_2^{\mathrm{H}}P_{Z_1^{\dagger}} \in \mathbb{S}$$

is the unique matrix such that

$$\|S_{\mathrm{opt}}\|_{\mathrm{F}} = \min_{S\in\mathbb{S}} \|S\|_{\mathrm{F}}.$$

Applying Lemma 3.2, we have

Lemma 3.3. *Given approximate eigen-quaternary* $\{U, V, K_{\rm R}, M_{\rm R}\}$ *satisfying* (3.3), $\mathbb{L} \neq \varnothing$ *if and only if* S_K *and* S_M *are symmetric.*

Proof. We first apply Lemma 3.2 with

$$S = \varDelta K, \quad Z_1 = U, \quad Z_2 = KU - VK_{\rm R}.$$

Notice $P_{Z_1^{\rm H}} = U^{\rm T}U = I$ and

$$P_{Z_1}Z_2Z_1^{\dagger} = UU^{\rm T}(VK_{\rm R} - KU)U^{\rm T} = U[\underbrace{(U^{\rm T}V)K_{\rm R}}_{S_K} - U^{\rm T}KU]U^{\rm T}$$

to conclude that $\varDelta K$ exists if and only if S_K is symmetric, as was to be shown. Next we apply Lemma 3.2 again but with $S = \varDelta M$, $Z_1 = V$, and $Z_2 = MV - UM_{\rm R}$. □

As we pointed out in section 2, for any given pair of deflating subspaces, the associated $\{K_{\rm R}, M_{\rm R}\}$ may be expressed in different ways, e.g., the ones in (2.7a) and (2.7b). Now, by Lemma 3.3, it becomes clear that (2.7a) is a good choice because it ensures that $\mathbb{L} \neq \varnothing$.

In the case of $\mathbb{L} \neq \varnothing$, we define the *optimal backward error* by

$$\zeta(U, V, K_{\rm R}, M_{\rm R}) := \min_{(\varDelta K, \varDelta M) \in \mathbb{L}} (\|\varDelta K\|_{\rm ui} + \|\varDelta M\|_{\rm ui}), \tag{3.9}$$

given a unitarily invariant norm $\|\cdot\|_{\rm ui}$. For any particular unitarily invariant norm, we will attach a suggestive subscript to ζ to indicate the norm used, e.g., $\zeta_2(U, V, K_{\rm R}, M_{\rm R})$ and $\zeta_{\rm F}(U, V, K_{\rm R}, M_{\rm R})$ defined under the spectral norm and the Frobenius norm, respectively.

Theorem 3.1. *Suppose* S_K *and* S_M *defined by* (3.4) *are symmetric. Then*

$$\begin{aligned} \zeta_{\rm F}(U, V, K_{\rm R}, M_{\rm R}) &= \sqrt{2\|\mathscr{R}_K(K_{\rm R})\|_{\rm F}^2 - \|U^{\rm T}\mathscr{R}_K(K_{\rm R})\|_{\rm F}^2} \\ &\quad + \sqrt{2\|\mathscr{R}_M(M_{\rm R})\|_{\rm F}^2 - \|V^{\rm T}\mathscr{R}_M(M_{\rm R})\|_{\rm F}^2}, \end{aligned} \tag{3.10}$$

$$\zeta_2(U, V, K_{\rm R}, M_{\rm R}) = \|\mathscr{R}_K(K_{\rm R})\|_2 + \|\mathscr{R}_M(M_{\rm R})\|_2, \tag{3.11}$$

and for a general unitarily invariant norm,

$$\|\mathscr{R}_K(K_{\rm R})\|_{\rm ui} + \|\mathscr{R}_M(M_{\rm R})\|_{\rm ui} \leq \zeta(U, V, K_{\rm R}, M_{\rm R}) \leq 2\Big[\|\mathscr{R}_K(K_{\rm R})\|_{\rm ui} + \|\mathscr{R}_M(M_{\rm R})\|_{\rm ui}\Big]. \tag{3.12}$$

Proof. Note that the minimization for $\zeta(U, V, K_{\rm R}, M_{\rm R})$ can be separated into the K-part and M-part. For the Frobenius norm, we can apply directly Lemma 3.2 with $Z_1 = U$ and $Z_2 = \mathscr{R}_K(K_{\rm R})$ to get the optimal $\varDelta K$ as

$$\varDelta K_{\rm opt}(K_{\rm R}) = \mathscr{R}_K(K_{\rm R})U^{\rm T} + [\mathscr{R}_K(K_{\rm R})]^{\rm T}U - U[\mathscr{R}_K(K_{\rm R})]^{\rm T}UU^{\rm T}, \tag{3.13}$$

whose Frobenius norm is $\sqrt{2\|\mathscr{R}_K(K_{\rm R})\|_{\rm F}^2 - \|U^{\rm T}\mathscr{R}_K(K_{\rm R})\|_{\rm F}^2}$, and similarly for the optimal $\varDelta M$ in the Frobenius norm.

Expand U to an orthogonal matrix $[U, U_\perp] \in \mathbb{R}^{n\times n}$ and write

$$\Delta K = [U, U_\perp] \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} [U, U_\perp]^{\mathrm{T}}.$$

Since $\Delta K U = -\mathscr{R}_K(K_{\mathrm{R}})$ by (3.7), we have $\begin{bmatrix} T_{11} \\ T_{21} \end{bmatrix} = -\begin{bmatrix} U^{\mathrm{T}} \\ U_\perp^{\mathrm{T}} \end{bmatrix} \mathscr{R}_K(K_{\mathrm{R}})$, and thus

$$\Delta K = -[U, U_\perp] \begin{bmatrix} U^{\mathrm{T}}\mathscr{R}_K(K_{\mathrm{R}}) & \mathscr{R}_K(K_{\mathrm{R}})^{\mathrm{T}} U_\perp \\ U_\perp^{\mathrm{T}}\mathscr{R}_K(K_{\mathrm{R}}) & T_{22} \end{bmatrix} [U, U_\perp]^{\mathrm{T}},$$

where T_{22} is symmetric and arbitrary. Therefore

$$\|\Delta K\|_{\mathrm{ui}} = \left\| \begin{bmatrix} U^{\mathrm{T}}\mathscr{R}_K(K_{\mathrm{R}}) & \mathscr{R}_K(K_{\mathrm{R}})^{\mathrm{T}} U_\perp \\ U_\perp^{\mathrm{T}}\mathscr{R}_K(K_{\mathrm{R}}) & T_{22} \end{bmatrix} \right\|_{\mathrm{ui}} \geq \left\| \begin{bmatrix} U^{\mathrm{T}}\mathscr{R}_K(K_{\mathrm{R}}) \\ U_\perp^{\mathrm{T}}\mathscr{R}_K(K_{\mathrm{R}}) \end{bmatrix} \right\|_{\mathrm{ui}} = \|\mathscr{R}_K(K_{\mathrm{R}})\|_{\mathrm{ui}}$$

for any T_{22}. Setting $T_{22} = 0$, we have

$$\|\Delta K\|_{\mathrm{ui}} \leq \left\| \begin{bmatrix} U^{\mathrm{T}}\mathscr{R}_K(K_{\mathrm{R}}) \\ U_\perp^{\mathrm{T}}\mathscr{R}_K(K_{\mathrm{R}}) \end{bmatrix} \right\|_{\mathrm{ui}} + \left\| \begin{bmatrix} \mathscr{R}_K(K_{\mathrm{R}})^{\mathrm{T}} U_\perp \\ 0 \end{bmatrix} \right\|_{\mathrm{ui}} \leq 2\|\mathscr{R}_K(K_{\mathrm{R}})\|_{\mathrm{ui}}.$$

Similar inequalities hold for the optimal ΔM. Together, they yield (3.12).

Finally for the spectral norm, by the dilation theorem of Kreĭn and Kahan (see, e.g., [9, 12] and [27, Theorem 1.2.3]), the optimal ΔK_{opt} in the sense that $\|\Delta K\|_2$ is smallest as T_{22} varies among all possible symmetric matrices is

$$\|\Delta K_{\mathrm{opt}}\|_2 = \left\| \begin{bmatrix} U^{\mathrm{T}}\mathscr{R}_K(K_{\mathrm{R}}) \\ U_\perp^{\mathrm{T}}\mathscr{R}_K(K_{\mathrm{R}}) \end{bmatrix} \right\|_2 = \|\mathscr{R}_K(K_{\mathrm{R}})\|_2,$$

and similarly for the optimal ΔM in the spectral norm. □

We remark that the optimal backward perturbation matrices ΔK and ΔM for the spectral norm and for the Frobenius norm may be different. In particular, the optimal $(\Delta K, \Delta M)$ for the Frobenius norm is unique and can be explicitly stated as by (3.13), while the optimal $(\Delta K, \Delta M)$ for the spectral norm, in general, is not unique and we do not have an explicit expression for it.

3.2 Optimal Residuals

Theorem 3.1 gives the minimal spectral norm and Frobenius norm for an approximate eigen-quaternary of $\{K, M\}$. In this subsection, we shall investigate the second question raised at the beginning of the section.

Given a pair of approximate deflating subspaces $\{\mathcal{U}, \mathcal{V}\}$, there are many $K_{\mathrm{R}} \in \mathbb{R}^{k\times k}$ and $M_{\mathrm{R}} \in \mathbb{R}^{k\times k}$, e.g., the ones in the form of (3.14) below, which, combined with the basis matrices U and V for $\mathcal{U}$ and $\mathcal{V}$, lead to approximate eigen-quaternary of $\{K, M\}$. Each approximate eigen-quaternary gives rise to an optimal backward error $\zeta(U, V, K_{\mathrm{R}}, M_{\mathrm{R}})$ as defined by (3.9). A natural question then is how small can $\zeta(U, V, K_{\mathrm{R}}, M_{\mathrm{R}})$ get by varying K_{R} and M_{R}.

Our investigation reveals a similar conclusion to that for the standard nonsymmetric eigenvalue problem in [10]: the Rayleigh quotient pair $\{K_{\rm RQ}, M_{\rm RQ}\}$ in (2.8) does not achieve the minimum in general, but is a reasonably good and computable choice.

Because of Lemma 3.3, we will consider these $K_{\rm R}$ and $M_{\rm R}$:

$$K_{\rm R} = (U^{\rm T}V)^{-1}S_K \quad \text{and} \quad M_{\rm R} = (V^{\rm T}U)^{-1}S_M, \tag{3.14}$$

where S_K, $S_M \in \mathbb{R}^{k\times k}$ are symmetric.

We begin by the case for $k = 1$. In this case, $K_{\rm R} = S_K/(\boldsymbol{u}^{\rm T}\boldsymbol{v})$ is a scalar, and by calculation, the optimal $K_{\rm R}$ in the spectral norm (using $\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2$=1) is

$$K_{\rm R} = \boldsymbol{u}^{\rm T}K\boldsymbol{v} \tag{3.15}$$

which is different from

$$K_{\rm RQ} = \boldsymbol{u}^{\rm T}K\boldsymbol{u}/(\boldsymbol{u}^{\rm T}\boldsymbol{v})$$

unless $\boldsymbol{u} = \pm\boldsymbol{v}$ or $\{\mathcal{R}(\boldsymbol{u}), \mathcal{R}(\boldsymbol{v})\}$ is already a pair of deflating subspaces. For the Frobenius norm, simple calculations yield the optimal $K_{\rm R}$ as

$$K_{\rm R} = \frac{2\boldsymbol{u}^{\rm T}K\boldsymbol{v} - \boldsymbol{u}^{\rm T}K\boldsymbol{u}(\boldsymbol{u}^{\rm T}\boldsymbol{v})}{2-(\boldsymbol{u}^{\rm T}\boldsymbol{v})^2} \tag{3.16}$$

which is not equal to $K_{\rm RQ}$, either. It is also noticed that the optimal $K_{\rm R}$ in (3.15) for the spectral norm differs from the one in (3.16) for the Frobenius norm.

In general for $k > 1$, it doesn't seem to be possible to derive closed formulas for the optimal $K_{\rm R}$ and $M_{\rm R}$ with respect to any unitarily invariant norm, even for the Frobenius norm which is often the easiest norm to handle. In Theorem 3.2 below, we present the determining equations for S_K and S_M that minimizes $\zeta_{\rm F}(U, V, K_{\rm R}, M_{\rm R})$.

Theorem 3.2. *For the Frobenius norm, there is a unique* $\{K_{\rm R}, M_{\rm R}\}$ *in the form of* (3.14) *that minimizes* $\zeta_{\rm F}(U, V, K_{\rm R}, M_{\rm R})$*, and* S_K *and* S_M *for the optimal* $\{K_{\rm R}, M_{\rm R}\}$ *satisfies*

$$\begin{aligned} &S_K(2Q^{\rm T}Q - I_k) + (2Q^{\rm T}Q - I_k)S_K - \operatorname{Diag}(S_K[2Q^{\rm T}Q - I_k]) \\ &\quad = -2U^{\rm T}KU + 2(Q^{\rm T}V^{\rm T}KU + U^{\rm T}KVQ) + \operatorname{Diag}(U^{\rm T}KU - 2Q^{\rm T}V^{\rm T}KU), \end{aligned} \tag{3.17a}$$

$$\begin{aligned} &S_M(2QQ^{\rm T} - I_k) + (2QQ^{\rm T} - I_k)S_M - \operatorname{Diag}(S_M[2QQ^{\rm T} - I_k]) \\ &\quad = -2V^{\rm T}MV + 2(QU^{\rm T}MV + V^{\rm T}MUQ^{\rm T}) + \operatorname{Diag}(V^{\rm T}MV - 2QU^{\rm T}MV), \end{aligned} \tag{3.17b}$$

where $Q := (U^{\rm T}V)^{-1}$ *and* $\operatorname{Diag}(Z)$ *denotes the diagonal matrix whose diagonal entries are those of* Z*.*

Proof. The equations in (3.17) are the first order optimality conditions for minimizing

$$2\|\mathscr{R}_K(K_{\rm R})\|_{\rm F}^2 - \|V^{\rm T}\mathscr{R}_K(K_{\rm R})\|_{\rm F}^2 \quad \text{and} \quad 2\|\mathscr{R}_M(M_{\rm R})\|_{\rm F}^2 - \|U^{\rm T}\mathscr{R}_M(M_{\rm R})\|_{\rm F}^2,$$

upon using

$$\|Z\|_{\rm F}^2 = \operatorname{trace}(Z^{\rm T}Z), \quad \text{and} \quad \frac{\partial \operatorname{trace}(ZS)}{\partial S} = Z + Z^{\rm T} - \operatorname{Diag}(Z) \quad \text{for } S = S^{\rm T}.$$

These equations are systems of linear equations. We claim that they have unique solutions and thus $\{K_{\rm R}, M_{\rm R}\}$ in the form of (3.14) must be unique. To see this, we show that the corresponding homogeneous systems have only the trivial solution, i.e., the solution 0. For (3.17a), denote by $X = S_K(2Q^{\rm T}Q - I_k)$ and note that

$$\mathrm{Diag}(X) = \mathrm{Diag}(\frac{X + X^{\rm T}}{2}).$$

Thus the corresponding homogeneous system is

$$X + X^{\rm T} = \mathrm{Diag}(\frac{X + X^{\rm T}}{2})$$

which is true if and only if $X + X^{\rm T} = 0$, i.e.,

$$(2Q^{\rm T}Q - I_k)S_K + S_K(2Q^{\rm T}Q - I_k) = 0.$$

This implies $S_K = 0$ since $2Q^{\rm T}Q - I_k$ is positive definite. This completes the proof. □

Even though (3.17a) and (3.17b) are both linear systems, it seems that there is no easy way to explicitly express S_K and S_M. In the special case $\mathcal{R}(U) = \mathcal{R}(V)$, the unique solutions are $S_K = U^{\rm T}KU$ and $S_M = V^{\rm T}MV$.

3.3 Near Optimality of the Rayleigh Quotient Pair

Previously, we introduced Rayleigh quotient pair $\{K_{\rm RQ}, M_{\rm RQ}\}$:

$$K_{\rm RQ} = (U^{\rm T}V)^{-1}U^{\rm T}KU \quad \text{and} \quad M_{\rm RQ} = (V^{\rm T}U)^{-1}V^{\rm T}MV. \tag{2.8}$$

It is in general not the optimal pair that minimizes $\zeta(U, V, K_{\rm R}, M_{\rm R})$ in the Frobenius norm and the spectral norm. So for a given pair of approximate deflating subspaces $\{\mathcal{U}, \mathcal{V}\}$, there are better pairs $\{K_{\rm R}, M_{\rm R}\}$, in the sense of giving smaller $\zeta(U, V, K_{\rm R}, M_{\rm R})$, than the Rayleigh quotient pair to extract partial spectral information for H from.

On the other hand, consider

$$H_{\rm RQ} = \begin{bmatrix} & K_{\rm RQ} \\ M_{\rm RQ} & \end{bmatrix}. \tag{2.9}$$

Factorize $U^{\rm T}V$ as $U^{\rm T}V = W_1^{\rm T}W_2$, where $W_i \in \mathbb{R}^{k\times k}$ are nonsingular. Recall the structure-preserving restriction

$$H_{\rm SR} = \begin{bmatrix} 0 & W_1^{-\rm T}U^{\rm T}KUW_1^{-1} \\ W_2^{-\rm T}V^{\rm T}MVW_2^{-1} & 0 \end{bmatrix} \tag{3.18}$$

introduced in [1, 2]. It can be verified that

$$H_{\rm RQ} = [W_2 \oplus W_1]^{-1}H_{\rm SR}[W_2 \oplus W_1],$$

and thus H_{RQ} and H_{SR} have the same eigenvalues. In [1, 2], it was the eigenvalues of H_{SR}, and thus of H_{RQ}, too, that were used to approximate part of the eigenvalues of H, given the pair of approximate deflating subspaces $\{\mathcal{U}, \mathcal{V}\}$. There, it was also proved that such eigenvalue approximation is optimal in the sense of the trace minimization principle obtained there. Therefore the Rayleigh quotient pair must be a reasonably good pair and cannot be too far from the optimal one $\{K_{\mathrm{R}}, M_{\mathrm{R}}\}$ in the form of (3.14) that minimizes $\zeta(U, V, K_{\mathrm{R}}, M_{\mathrm{R}})$. In this subsection, we will justify such a claim.

Recall our assumptions:

$$U^{\mathrm{T}}U = V^{\mathrm{T}}V = I_k, \quad \text{and} \quad \operatorname{rank}(U^{\mathrm{T}}V) = k. \tag{3.3}$$

Let K_{R} and M_{R} be in the form of (3.14), where S_K and S_M are symmetric. The angle between $\mathcal{U} = \mathcal{R}(U)$ and $\mathcal{V} = \mathcal{R}(V)$ is defined by

$$\theta_{\max}(\mathcal{U}, \mathcal{V}) := \arccos \sigma_{\min}(U^{\mathrm{T}}V),$$

where $\sigma_{\min}(U^{\mathrm{T}}V)$ is the smallest singular value of $U^{\mathrm{T}}V$.

Owing to the definition of $\zeta(U, V, K_{\mathrm{R}}, M_{\mathrm{R}})$ and Theorem 3.1, we will focus on minimizing the norms of $\mathscr{R}_K(K_{\mathrm{R}})$ and $\mathscr{R}_M(M_{\mathrm{R}})$, separately.

Lemma 3.4. *For any K_{R}, M_{R}, and unitarily invariant norm $\|\cdot\|_{\mathrm{ui}}$,*

$$\|K_{\mathrm{RQ}} - K_{\mathrm{R}}\|_{\mathrm{ui}} \le \alpha \cdot \|\mathscr{R}_K(K_{\mathrm{R}})\|_{\mathrm{ui}}, \tag{3.19a}$$

$$\|M_{\mathrm{RQ}} - M_{\mathrm{R}}\|_{\mathrm{ui}} \le \alpha \cdot \|\mathscr{R}_M(M_{\mathrm{R}})\|_{\mathrm{ui}}, \tag{3.19b}$$

where

$$\alpha = \frac{\sqrt{1 + \sin\theta_{\max}(\mathcal{U}, \mathcal{V})}}{\cos\theta_{\max}(\mathcal{U}, \mathcal{V})}. \tag{3.20}$$

Proof. We will prove (3.19a) only since (3.19b) can be proved in the same way. Let $V_\perp \in \mathbb{R}^{n\times(n-k)}$ that makes $[V, V_\perp]$ an orthogonal matrix and set $P = [U, V_\perp]$. Write

$$\begin{aligned} \mathscr{R}_K(K_{\mathrm{R}}) &= KU - VK_{\mathrm{R}} \\ &= (KU - VK_{\mathrm{RQ}}) + V(K_{\mathrm{RQ}} - K_{\mathrm{R}}) \\ &= \mathscr{R}_K(K_{\mathrm{RQ}}) + V(K_{\mathrm{RQ}} - K_{\mathrm{R}}). \end{aligned} \tag{3.21}$$

Now use $U^{\mathrm{T}}\mathscr{R}_K(K_{\mathrm{RQ}}) = 0$ and $V_\perp^{\mathrm{T}}V = 0$ to get

$$P^{\mathrm{T}}\mathscr{R}_K(K_{\mathrm{R}}) = \begin{bmatrix} U^{\mathrm{T}}V(K_{\mathrm{RQ}} - K_{\mathrm{R}}) \\ V_\perp^{\mathrm{T}}\mathscr{R}_K(K_{\mathrm{RQ}}) \end{bmatrix}. \tag{3.22}$$

Therefore

$$\begin{aligned} \|P^{\mathrm{T}}\mathscr{R}_K(K_{\mathrm{R}})\|_{\mathrm{ui}} &\ge \|(U^{\mathrm{T}}V)(K_{\mathrm{RQ}} - K_{\mathrm{R}})\|_{\mathrm{ui}} \\ &\ge \sigma_{\min}(U^{\mathrm{T}}V) \cdot \|K_{\mathrm{RQ}} - K_{\mathrm{R}}\|_{\mathrm{ui}}, \end{aligned} \tag{3.23}$$

$$\|P^{\mathrm{T}}\mathscr{R}_K(K_{\mathrm{R}})\|_{\mathrm{ui}} \le \|P\|_2 \|\mathscr{R}_K(K_{\mathrm{R}})\|_{\mathrm{ui}}$$

$$= \sqrt{1+\sin\theta_{\max}(\mathcal{U},\mathcal{V})}\,\|\mathscr{R}_K(K_{\rm R})\|_{\rm ui}. \tag{3.24}$$

In deriving (3.24), we have used

$$P^{\rm T}P = \begin{bmatrix} I_k & U^{\rm T}V_\perp \\ V_\perp^{\rm T}U & I_{n-k}\end{bmatrix} \quad \Rightarrow \quad \|P\|_2^2 = \|P^{\rm T}P\|_2 = 1 + \|U^{\rm T}V_\perp\|_2 = 1+\sin\theta_{\max}(\mathcal{U},\mathcal{V}).$$

Combine (3.23) and (3.24) to get (3.19a). ☐

Theorem 3.3. *For any unitarily invariant norm* $\|\cdot\|_{\rm ui}$*,*

$$\min\|K_{\rm RQ}-K_{\rm R}\|_{\rm ui} \le \alpha\cdot\min\|\mathscr{R}_K(K_{\rm R})\|_{\rm ui}, \tag{3.25a}$$
$$\min\|M_{\rm RQ}-M_{\rm R}\|_{\rm ui} \le \alpha\cdot\min\|\mathscr{R}_M(M_{\rm R})\|_{\rm ui}, \tag{3.25b}$$

and

$$\min\|\mathscr{R}_K(K_{\rm R})\|_{\rm ui} \le \|\mathscr{R}_K(K_{\rm RQ})\|_{\rm ui} \le (1+\alpha)\cdot\min\|\mathscr{R}_K(K_{\rm R})\|_{\rm ui}, \tag{3.26a}$$
$$\min\|\mathscr{R}_M(M_{\rm R})\|_{\rm ui} \le \|\mathscr{R}_M(M_{\rm RQ})\|_{\rm ui} \le (1+\alpha)\cdot\min\|\mathscr{R}_M(M_{\rm R})\|_{\rm ui}. \tag{3.26b}$$

*where the "*min*" in* (3.25a) *and* (3.26a) *are taken over all* $K_{\rm R}$ *in the form of* (3.14) *with symmetric* S_K*, and the ones in* (3.25b) *and* (3.26b) *are taken over all* $M_{\rm R}$ *in the form of* (3.14) *with symmetric* S_M*, and* α *is given by* (3.20)*.*

Proof. The inequalities in (3.25) are direct consequences of Lemma 3.4. In what follows, we will prove (3.26a) only since (3.26b) can be proved in the same way. The first inequality in (3.26a) is evident. For the second inequality there, we note

$$\mathscr{R}_K(K_{\rm RQ}) = \mathscr{R}_K(K_{\rm R}) - V(K_{\rm RQ}-K_{\rm R})$$

by (3.21) and thus

$$\begin{aligned}\|\mathscr{R}_K(K_{\rm RQ})\|_{\rm ui} &\le \|\mathscr{R}_K(K_{\rm R})\|_{\rm ui} + \|V(K_{\rm RQ}-K_{\rm R})\|_{\rm ui} \\ &\le \|\mathscr{R}_K(K_{\rm R})\|_{\rm ui} + \|K_{\rm RQ}-K_{\rm R}\|_{\rm ui} \\ &\le (1+\alpha)\|\mathscr{R}_K(K_{\rm R})\|_{\rm ui}\end{aligned} \tag{3.27}$$

for all $K_{\rm R}$ in the form of (3.14) with symmetric S_K. Minimizing the right-hand side of (3.27) over all S_K leads to the second inequality in (3.26a). ☐

4 Residual-based Error Bounds for Eigenvalues

As preparation, we first cite an eigenvalue perturbation result for a positive definite pencil in subsection 4.1 and apply it to LREP (1.1) in subsection 4.2, and then come to develop residual based error bounds in subsection 4.3. Results in both subsections 4.1 and 4.2 are of independent interests on their own from the rest of this article.

In what follows, $A \succ 0$ means that A is Hermitian and positive definite.

4.1 A Perturbation Bound for Positive Definite Pencil

Consider a Hermitian matrix pencil $A - \lambda B$, where A, $B \in \mathbb{C}^{n\times n}$ are Hermitian. It is called a *positive definite pencil* if there is a $\lambda_0 \in \mathbb{R}$ such that $A - \lambda_0 B \succ 0$ [11, 13, 16].

Suppose that $A - \lambda B$ is a positive definite pencil and B is nonsingular. Let n_+ and n_- be the numbers of positive and negative eigenvalues of B, respectively. Note $n_+ + n_- = n$. It is known [16] that $A - \lambda B$ has only real eigenvalues which we will divide into two groups $\{\lambda_i^-\}_{i=1}^{n_-}$ and $\{\lambda_i^+\}_{i=1}^{n_+}$ and which can be arranged in the order as

$$\lambda_{n_-}^- \le \cdots \le \lambda_1^- < \lambda_1^+ \le \cdots \le \lambda_{n_+}^+.$$

Moreover, $A - \lambda B$ is diagonalizable [6, 16]: there exists nonsingular $Z \in \mathbb{C}^{n\times n}$ such that

$$Z^{\mathrm{H}} A Z = \operatorname{diag}(\Lambda_+, -\Lambda_-), \quad Z^{\mathrm{H}} B Z = J := \operatorname{diag}(I_{n_+}, -I_{n_-}), \tag{4.1}$$

where $\Lambda_\pm = \operatorname{diag}(\lambda_1^\pm, \lambda_2^\pm, \dots, \lambda_{n_\pm}^\pm)$ and $\Lambda = \operatorname{diag}(\Lambda_+, \Lambda_-)$.

Lemma 4.1 ([17])**.** *Let $A - \lambda B$ be a positive definite pencil with nonsingular B and with the eigen-decomposition* (4.1)*. Suppose it is perturbed to another positive definite pencil $\widetilde{A} - \lambda\widetilde{B}$ with nonsingular $\widetilde{B}$, and adopt the same notations for this perturbed pencil as those for $A - \lambda B$ except with a* tilde *on each symbol. Then for any unitarily invariant norm $\|\cdot\|_{\mathrm{ui}}$,*

$$\|\widetilde{\Lambda} - \Lambda\|_{\mathrm{ui}} \le \|Z\|_2 \|\widetilde{Z}\|_2 \Big(\|\widetilde{A} - A\|_{\mathrm{ui}} + \xi \|\widetilde{B} - B\|_{\mathrm{ui}} \Big),$$

where $\xi = \max\{\|\Lambda\|_2, \|\widetilde{\Lambda}\|_2\}$.

The concept of positive definite pencil is closely related to that of the so-called *definite pencil* in the past literature [21, 23, 24]. The latter is more general, encompassing the former. In general, B may be singular, but Lemma 4.1 excludes the case. When B is singular, infinite eigenvalues occur. In order to be able to deal with both finite and infinite eigenvalues at the same time, in the literature number pairs (α, β) were used to represent eigenvalues α/β which is finite if $\beta \ne 0$ and infinite otherwise and the chordal distance was used to measure the difference between two eigenvalues, finite or not. Lemma 4.1 resembles various perturbation bounds in [5, 14, 15, 21, 23] for the definite pencils.

4.2 A Perturbation Bound for LREP

Consider LREP (1.1) with $K \succ 0$ and $M \succ 0$. It is equivalent to the generalized eigenvalue problem for the matrix pencil [1]

$$\boldsymbol{A} - \lambda \boldsymbol{B} \equiv \begin{bmatrix} M & \\ & K \end{bmatrix} - \lambda \begin{bmatrix} 0 & I_n \\ I_n & 0 \end{bmatrix}. \tag{4.2}$$

$\boldsymbol{A} - \lambda\boldsymbol{B}$ is a positive definite pencil because $\boldsymbol{A} - 0\cdot\boldsymbol{B} = \boldsymbol{A} \succ 0$. Recall Theorem 2.1. We find the eigen-decomposition for $\boldsymbol{A} - \lambda\boldsymbol{B}$:

$$Z^{\mathrm{T}} \boldsymbol{A} Z = \operatorname{diag}(\Lambda, \Lambda), \quad Z^{\mathrm{T}} \boldsymbol{B} Z = \operatorname{diag}(I_n, -I_n), \tag{4.3}$$

where

$$Z = \begin{bmatrix} \Psi\Lambda^{1/2} & \Psi\Lambda^{1/2} \\ \Phi\Lambda^{-1/2} & -\Phi\Lambda^{-1/2} \end{bmatrix}. \tag{4.4}$$

The next theorem bounds Z and its inverse from above and below.

Theorem 4.1. *For Z in* (4.4)*,*

$$2\gamma_1 \le \|Z\|_2^2 \le 2\gamma_2, \quad \frac{1}{2}\gamma_1 \le \|Z^{-1}\|_2^2 \le \frac{1}{2}\gamma_2, \tag{4.5}$$

where

$$\gamma_1 = \max\left\{\|M^{-1}\|_2\lambda_1, \frac{\|M\|_2}{\lambda_n}\right\}, \quad \gamma_2 = \left\{\|M^{-1}\|_2\lambda_n, \frac{\|M\|_2}{\lambda_1}\right\}. \tag{4.6}$$

They are also valid if all occurrences of M are replaced by K.

Proof. It can be verified that

$$ZZ^{\mathrm T} = 2\begin{bmatrix} \Psi\Lambda\Psi^{\mathrm T} & 0 \\ 0 & \Phi\Lambda^{-1}\Phi^{\mathrm T} \end{bmatrix}, \quad Z^{-\mathrm T}Z^{-1} = \frac{1}{2}\begin{bmatrix} \Phi\Lambda^{-1}\Phi^{\mathrm T} & 0 \\ 0 & \Psi\Lambda\Psi^{\mathrm T} \end{bmatrix}.$$

Therefore

$$\begin{aligned}
\|Z\|_2^2 &\le 2\max\left\{\|\Psi\|_2^2\lambda_n, \frac{\|\Phi\|_2^2}{\lambda_1}\right\} = 2\max\left\{\|M^{-1}\|_2\lambda_n, \frac{\|M\|_2}{\lambda_1}\right\}, \\
\|Z\|_2^2 &\ge 2\max\left\{\|\Psi\|_2^2\lambda_1, \frac{\|\Phi\|_2^2}{\lambda_n}\right\} = 2\max\left\{\|M^{-1}\|_2\lambda_1, \frac{\|M\|_2}{\lambda_n}\right\}, \\
\|Z^{-1}\|_2^2 &\le \frac{1}{2}\max\left\{\frac{\|\Phi\|_2^2}{\lambda_1}, \|\Psi\|_2^2\lambda_n\right\} = \frac{1}{2}\max\left\{\frac{\|M\|_2}{\lambda_1}, \|M^{-1}\|_2\lambda_n\right\}, \\
\|Z^{-1}\|_2^2 &\ge \frac{1}{2}\max\left\{\frac{\|\Phi\|_2^2}{\lambda_n}, \|\Psi\|_2^2\lambda_1\right\} = \frac{1}{2}\max\left\{\frac{\|M\|_2}{\lambda_n}, \|M^{-1}\|_2\lambda_1\right\},
\end{aligned}$$

where we have used $M = \Phi\Phi^{\mathrm T}$ and $M^{-1} = \Psi\Psi^{\mathrm T}$. Together, they yield (4.5). To see the last claim of this theorem, we let $\widehat{\Psi} = \Psi\Lambda$ and $\widehat{\Phi} = \Phi\Lambda^{-1}$. It can be verified that

$$K = \Psi\Lambda^2\Psi^{\mathrm T} = \widehat{\Psi}\widehat{\Psi}^{\mathrm T}, \quad M = \Phi\Phi^{\mathrm T} = \widehat{\Phi}\Lambda^2\widehat{\Phi}^{\mathrm T}, \quad Z = \begin{bmatrix} \widehat{\Psi}\Lambda^{-1/2} & \widehat{\Psi}\Lambda^{-1/2} \\ \widehat{\Phi}\Lambda^{1/2} & -\widehat{\Phi}\Lambda^{1/2} \end{bmatrix},$$

and $K^{-1} = \widehat{\Phi}\widehat{\Phi}^{\mathrm T}$. Following the same lines of argument as above, we see all inequalities in (4.5) are valid if all occurrences of M are replaced by K. □

A straightforward application of Lemma 4.1 leads to

Theorem 4.2. *For* LREP (1.1) *with $K \succ 0$ and $M \succ 0$ admitting the decompositions in* Theorem 2.1*, let Z be defined by* (4.4)*. Suppose that K and M are perturbed to $\widetilde{K} \succ 0$ and $\widetilde{M} \succ 0$, and accordingly H is perturbed to $\widetilde{H}$. Adopt the same notations for the perturbed*

LREP *for* $\widetilde{H}$ *as those for* H *except with a* tilde *on each symbol. Then for any unitarily invariant norm* $\|\cdot\|_{\rm ui}$,

$$\|\widetilde{\Lambda}\oplus\widetilde{\Lambda}-\Lambda\oplus\Lambda\|_{\rm ui}\le\|Z\|_2\|\widetilde{Z}\|_2\|\widetilde{M}\oplus\widetilde{K}-M\oplus K\|_{\rm ui}. \tag{4.7}$$

In particular,

$$\max_{1\le i\le n}|\widetilde{\lambda}_i-\lambda_i|\le\|Z\|_2\|\widetilde{Z}\|_2\max\{\|\widetilde{M}-M\|_2,\|\widetilde{K}-K\|_2\}, \tag{4.8a}$$

$$\sqrt{\sum_{i=1}^n|\widetilde{\lambda}_i-\lambda_i|^2}\le\frac{1}{\sqrt{2}}\|Z\|_2\|\widetilde{Z}\|_2\sqrt{\|\widetilde{M}-M\|_{\rm F}^2+\|\widetilde{K}-K\|_{\rm F}^2}. \tag{4.8b}$$

In the left-hand side of (4.7), the difference between $\widetilde{\Lambda}$ and Λ appears twice. This repetition is handily removed for the spectral and Frobenius norm in (4.8). In general, it is not so easy to remove the repetition without weakening the inequality a little bit. In the corollary below, we show one way of doing it.

Corollary 4.1. *Under the conditions of* Theorem 4.2*,*

$$\|\widetilde{\Lambda}-\Lambda\|_{\rm ui}\le\|Z\|_2\|\widetilde{Z}\|_2\Big[\|\widetilde{M}-M\|_{\rm ui}+\|\widetilde{K}-K\|_{\rm ui}\Big]. \tag{4.9}$$

Proof. It is suffices to show that (4.9) holds for all Ky Fan 2ℓ-norm $\|\cdot\|_{(\ell)}$ which is the sum of the ℓ largest singular values of its argument [4, 8, 22].

Let $\{i_1,i_2,\dots,i_n\}$ be a permutation of $\{1,2,\dots,n\}$ such that

$$|\widetilde{\lambda}_{i_1}-\lambda_{i_1}|\ge\cdots\ge|\widetilde{\lambda}_{i_n}-\lambda_{i_n}|.$$

For $1\le\ell\le n$, we have

$$\begin{aligned}\|\widetilde{\Lambda}\oplus\widetilde{\Lambda}-\Lambda\oplus\Lambda\|_{(2\ell)}&=2\sum_{j=1}^{\ell}|\widetilde{\lambda}_{i_j}-\lambda_{i_j}|\\&=2\|\widetilde{\Lambda}-\Lambda\|_{(\ell)},\\ \|\widetilde{M}\oplus\widetilde{K}-M\oplus K\|_{(2\ell)}&\le\|\widetilde{M}-M\|_{(2\ell)}+\|\widetilde{K}-K\|_{(2\ell)}\\&\le2\Big[\|\widetilde{M}-M\|_{(\ell)}+\|\widetilde{K}-K\|_{(\ell)}\Big].\end{aligned}$$

By Theorem 4.2, we have

$$\|\widetilde{\Lambda}-\Lambda\|_{(\ell)}\le\|Z\|_2\|\widetilde{Z}\|_2\Big[\|\widetilde{M}-M\|_{(\ell)}+\|\widetilde{K}-K\|_{(\ell)}\Big],$$

as expected. □

4.3 Residual Based Error Bounds for LREP

Consider an approximate eigen-quaternary $\{U, V, K_{\rm R}, M_{\rm R}\}$ of $\{K, M\}$, where $U, V \in \mathbb{R}^{n\times k}$ satisfying, as before,

$$U^{\rm T}U = V^{\rm T}V = I_k, \quad \text{and} \quad \operatorname{rank}(U^{\rm T}V) = k, \tag{3.3}$$

and $K_{\rm R}, M_{\rm R} \in \mathbb{R}^{k\times k}$, and define $\mathscr{R}_K(K_{\rm R})$ and $\mathscr{R}_M(M_{\rm R})$ by (3.1) and $H_{\rm R}$ by (3.2). In subsection 3.1, we showed that $\{U, V, K_{\rm R}, M_{\rm R}\}$ of $\{K, M\}$ is an exact eigen-quaternary of

$$\{\widetilde{K}, \widetilde{M}\} := \{K + \Delta K, M + \Delta M\} \tag{4.10}$$

with bounds in norm on ΔK and ΔM in terms of the residuals $\mathscr{R}_K(K_{\rm R})$ and $\mathscr{R}_M(M_{\rm R})$. If the two residuals are sufficiently small, then $\widetilde{K} \succ 0$ and $\widetilde{M} \succ 0$ and the eigenvalue problem for the corresponding $\widetilde{H}$ is again an LREP, making all results in subsection 4.2 applicable.

Lemma 4.2. *Suppose $\|\mathscr{R}_K(K_{\rm R})\|_2 < \sigma_{\min}(K)$ and $\|\mathscr{R}_M(M_{\rm R})\|_2 < \sigma_{\min}(M)$. Then $H_{\rm R}$ is similar to an* LREP *of $2k \times 2k$. Consequently, all eigenvalues of $H_{\rm R}$ are real and they come in $\{\pm\lambda\}$ pairs.*

Proof. By Theorem 3.1, the approximate exact eigen-quaternary $\{U, V, K_{\rm R}, M_{\rm R}\}$ of $\{K, M\}$ is an exact eigen-quaternary of $\{\widetilde{K}, \widetilde{M}\}$ as in (4.10) with

$$\|\Delta K\|_2 = \|\mathscr{R}_K(K_{\rm R})\|_2 < \sigma_{\min}(K), \quad \|\Delta M\|_2 = \|\mathscr{R}_M(M_{\rm R})\|_2 < \sigma_{\min}(M).$$

Now apply Lemma 3.1 to conclude that $H_{\rm R}$ is similar to

$$\begin{bmatrix} 0 & W_1^{-\rm T}(U^{\rm T}\widetilde{K}U)W_1^{-1} \\ W_2^{-\rm T}(V^{\rm T}\widetilde{M}V)W_2^{-1} & 0 \end{bmatrix}$$

whose eigenvalue problem is an LREP, where W_i are as defined in Lemma 3.1. □

In what follows, whenever $H_{\rm R}$ is similar to an LREP of $2k \times 2k$, we will denote its eigenvalues by

$$-\mu_k \le \cdots \le -\mu_1 < \mu_1 \le \cdots \le \mu_k. \tag{4.11}$$

Let $Z \in \mathbb{R}^{2n\times 2n}$ be the one that diagonalizes $\boldsymbol{A} - \lambda\boldsymbol{B}$ defined in (4.2) and (4.3) and similarly $\widetilde{Z}$ diagonalizes $\widetilde{\boldsymbol{A}} - \lambda\widetilde{\boldsymbol{B}}$ which is similarly defined in terms of $\widetilde{K}$ and $\widetilde{M}$.

The appearance of $\widetilde{Z}$ is the unsatisfactory part of the results below since ΔK and ΔM are usually unknown. But we argue that it does not necessarily invalidate the usefulness of these results. Because for sufficiently small $\mathscr{R}_K(K_{\rm R})$ and $\mathscr{R}_M(M_{\rm R})$ in norm, it is reasonable to expect $\|\widetilde{Z}\|_2 \approx \|Z\|_2$ and the latter can be bounded as in Theorem 4.1.

Theorem 4.3. *If $\|\mathscr{R}_K(K_{\rm R})\|_2 < \sigma_{\min}(K)$ and $\|\mathscr{R}_M(M_{\rm R})\|_2 < \sigma_{\min}(M)$, then there are k positive eigenvalues of H:*

$$\lambda_{i_1} \le \cdots \le \lambda_{i_k}$$

such that

$$\max_{1\le j\le k} |\lambda_{i_j} - \mu_j| \le \|Z\|_2 \|\widetilde{Z}\|_2 \max\{\|\mathscr{R}_K(K_{\rm R})\|_2, \|\mathscr{R}_M(M_{\rm R})\|_2\}. \tag{4.12}$$

Proof. The conditions of the theorem ensure that $\{U, V, K_{\rm R}, M_{\rm R}\}$ is an exact eigen-quaternary of $\{\widetilde{K}, \widetilde{M}\}$ in (4.10) with $\|\mathit{\Delta} K\|_2 = \|\mathscr{R}_K(K_{\rm R})\|_2$ and $\|\mathit{\Delta} M\|_2 = \|\mathscr{R}_M(M_{\rm R})\|_2$. Thus μ_j for $1 \le j \le k$ are among the positive eigenvalues of $\widetilde{H}$. Let μ_j be the i_jth positive eigenvalue of $\widetilde{H}$. The inequality (4.12) is now a consequence of (4.8a). □

In a similar way, we can prove

Theorem 4.4. *If $\|\mathscr{R}_K(K_{\rm R})\|_{\rm F} < \sigma_{\min}(K)$ and $\|\mathscr{R}_M(M_{\rm R})\|_{\rm F} < \sigma_{\min}(M)$, then there are k eigenvalues of H:*

$$\lambda_{i_1} \le \cdots \le \lambda_{i_k}$$

such that

$$\sqrt{\sum_{1\le j\le k} |\lambda_{i_j} - \mu_j|^2} \le \|Z\|_2 \|\widetilde{Z}\|_2 \Big[2\|\mathscr{R}_K(K_{\rm R})\|_{\rm F}^2 - \|U^{\rm T}\mathscr{R}_K(K_{\rm R})\|_{\rm F}^2 + 2\|\mathscr{R}_M(M_{\rm R})\|_{\rm F}^2 - \|V^{\rm T}\mathscr{R}_M(M_{\rm R})\|_{\rm F}^2 \Big]^{1/2}. \tag{4.13}$$

Proof. The conditions on $\|\mathscr{R}_K(K_{\rm R})\|_{\rm F}$ and $\|\mathscr{R}_M(M_{\rm R})\|_{\rm F}$ ensure that $\widetilde{H}$ defined with the optimal $\mathit{\Delta} K$ and $\mathit{\Delta} M$ in the Frobenius norm is an LREP because, by the proof of Theorem 3.1,

$$\|\mathit{\Delta} K\|_2 \le \|\mathit{\Delta} K\|_{\rm F} = \sqrt{2\|\mathscr{R}_K(K_{\rm R})\|_{\rm F}^2 - \|U^{\rm T}\mathscr{R}_K(K_{\rm R})\|_{\rm F}^2} \le \|\mathscr{R}_K(K_{\rm R})\|_{\rm F} < \sigma_{\min}(K)$$

and similarly $\mathit{\Delta} M\|_2 < \sigma_{\min}(M)$. The inequality (4.13) is now a consequence of (4.8b). □

The conditions on $\mathscr{R}_K(K_{\rm R})$ and $\mathscr{R}_M(M_{\rm R})$ in Theorem 4.4 seem to be stronger than necessary at first sight. It would be more natural to have the same conditions as stated in Theorem 4.3. The thing is that we don't know if $\|\mathit{\Delta} K\|_2 \le \|\mathscr{R}_K(K_{\rm R})\|_2$ for the optimal $\mathit{\Delta} K$ in the Frobenius norm while we do know $\|\mathit{\Delta} K\|_2 = \|\mathscr{R}_K(K_{\rm R})\|_2$ for the optimal $\mathit{\Delta} K$ in the spectral norm. This same reasoning explains the seemingly stronger than necessary conditions in Theorem 4.5 below for any unitarily invariant norm $\|\cdot\|_{\rm ui}$.

Theorem 4.5. *If $2\|\mathscr{R}_K(K_{\rm R})\|_{\rm ui} < \sigma_{\min}(K)$ and $2\|\mathscr{R}_M(M_{\rm R})\|_{\rm ui} < \sigma_{\min}(M)$, then there are k eigenvalues of H:*

$$\lambda_{i_1} \le \cdots \le \lambda_{i_k}$$

such that

$$\|\widetilde{\mathit{\Omega}} - \mathit{\Omega}\|_{\rm ui} \le 2\|Z\|_2\|\widetilde{Z}\|_2 \Big[\|\mathscr{R}_K(K_{\rm R})\|_{\rm ui} + \|\mathscr{R}_M(M_{\rm R})\|_{\rm ui}\Big], \tag{4.14}$$

where $\mathit{\Omega} = \operatorname{diag}\big(\lambda_{i_1}, \ldots, \lambda_{i_k}\big)$ and $\widetilde{\mathit{\Omega}} = \operatorname{diag}\big(\mu_1, \ldots, \mu_k\big)$.

Proof. The conditions on $\|\mathscr{R}_K(K_{\rm R})\|_{\rm ui}$ and $\|\mathscr{R}_M(M_{\rm R})\|_{\rm ui}$ ensure that $\widetilde{H}$ defined with the optimal $\mathit{\Delta} K$ and $\mathit{\Delta} M$ in the unitary invariant norm is an LREP because, by the proof of Theorem 3.1,

$$\|\mathit{\Delta} K\|_2 \le \|\mathit{\Delta} K\|_{\rm ui} \le 2\|\mathscr{R}_K(K_{\rm R})\|_{\rm ui} < \sigma_{\min}(K)$$

and similarly $\|\Delta M\|_2 < \sigma_{\min}(M)$. By Corollary 4.1, we have

$$\begin{aligned}\|\widetilde{\Omega} - \Omega\|_{\rm ui} &\le \|\widetilde{\Lambda} - \Lambda\|_{\rm ui} \\ &\le \|Z\|_2\|\widetilde{Z}\|_2\Big[\|\Delta K\|_{\rm ui} + \|\Delta M\|_{\rm ui}\Big] \\ &\le 2\|Z\|_2\|\widetilde{Z}\|_2\Big[\|\mathscr{R}_K(K_{\rm R})\|_{\rm ui} + \|\mathscr{R}_M(M_{\rm R})\|_{\rm ui}\Big],\end{aligned}$$

as expected. □

Remark 4.1. For a given pair of approximate deflating subspaces $\{\mathcal{U}, \mathcal{V}\}$, as in [1, 2], likely we will use the associated Rayleigh quotient pair $\{K_{\rm RQ}, M_{\rm RQ}\}$ to make up an approximate eigen-quaternary $\{U, V, K_{\rm RQ}, M_{\rm RQ}\}$ of $\{K, M\}$. Theorems 4.3 and 4.4 can be applied with $K_{\rm R} = K_{\rm RQ}$ and $M_{\rm R} = M_{\rm RQ}$ to arrive at corresponding error bounds on eigenvalue approximations by the eigenvalues of $H_{\rm RQ}$ in (2.9).

5 Compare with the Standard Eigenvalue Problems

Analogous questions to what we have been investigating so far had been thoroughly studied for the standard eigenvalue problems. Our results here resemble those in the literature. In what follows, we give a brief review on the related results.

Consider the eigenvalue problem for $C \in \mathbb{C}^{n\times n}$. Suppose $X \in \mathbb{C}^{n\times k}$ whose columns span an approximate invariant subspace, i.e., $AX \approx XD$ for some $D \in \mathbb{C}^{k\times k}$ in the sense that

$$\mathscr{R}(D) = AX - XD$$

is relatively small in norm. Let

$$\mathbb{E}_1 := \{E \in \mathbb{C}^{n\times n} : (C + E)X = XD\}.$$

Each E in $\mathbb{E}_1$ makes $\mathcal{R}(X)$ an invariant subspace of $C + E$ associated with its partial spectrum $\operatorname{eig}(D)$. Define the optimal backward error by

$$\eta(X, D) := \min_{E\in\mathbb{E}_1} \|E\|_{\rm ui}\,.$$

It is shown [27, Theorem 2.4.2] that

$$\eta(X, D) = \|\mathscr{R}(D)\|_{\rm ui}\,.$$

In the other word, for any given approximate eigen-matrix pair (D, X), the minimal norm $\|E\|_{\rm ui}$ among all possible backward perturbations is given by $\|\mathscr{R}(D)\|_{\rm ui}$. The next question is to minimize $\eta(X, D)$ as D varies. If also $X^{\rm H}X = I_k$, we have (see, e.g., [27, Theorem 2.4.1], [25, Theorem 2.1], [22, Theorem IV.1.15])

$$\min_{D\in\mathbb{C}^{k\times k}} \|\mathscr{R}(D)\|_{\rm ui} = \|CX - X(X^{\rm H}CX)\|_{\rm ui},$$

i.e., the Rayleigh quotient matrix $D = X^{\rm H}CX$ achieves the minimum of $\|\mathscr{R}(D)\|_{\rm ui}$ over $D \in \mathbb{C}^{k\times k}$, and it is unique for $\|\cdot\|_{\rm ui} = \|\cdot\|_{\rm F}$.

For Hermitian $C \in \mathbb{C}^{n\times n}$, it is often desirable to enforce that $C+E$ remains Hermitian too. In this case, define [26]

$$\eta(X, D) := \min_{E\in\mathbb{E}_2} \|E\|_{\rm ui}\,,$$

where $\mathbb{E}_2 := \big\{E = E^{\rm H} \in \mathbb{C}^{n\times n} : (C+E)X = XD\big\}$. Suppose $X^{\rm H}X = I_k$. It is shown that (a special case of the main theorem in [10, p.478])

$$\begin{aligned}\eta_2(X, D) &:= \min_{E\in\mathbb{E}_2} \|E\|_2 = \|\mathscr{R}(D)\|_2,\\ \eta_{\rm F}(X, D) &:= \min_{E\in\mathbb{E}_2} \|E\|_{\rm F} = \sqrt{2\|\mathscr{R}(D)\|_{\rm F}^2 - \|X^{\rm H}\mathscr{R}(D)\|_{\rm F}}\,,\end{aligned}$$

but no expression for $\eta(X, D)$ for a general $\|\cdot\|_{\rm ui}$ is known. Moreover, among all Hermitian $D \in \mathbb{C}^{k\times k}$, the Rayleigh quotient $D = X^{\rm H}CX$ achieves the minimums for $\eta(X, D)$ for all $\|\cdot\|_{\rm ui}$ [25, Theorem 2.2], i.e.,

$$X^{\rm H}CX = \operatorname*{argmin}_{D} \eta(X, D),$$

including $\eta_2(X, D)$ and $\eta_{\rm F}(X, D)$. Let the eigenvalues of C and those of $D = X^{\rm H}CX$ be

$$\lambda_1 \le \cdots \le \lambda_n, \quad \mu_1 \le \cdots \le \mu_k,$$

respectively. As a consequence of well-known perturbation results for Hermitian matrices [22], there exist $i_1 < i_2 < \cdots < i_k$ such that

$$\begin{aligned}\max_{1\le j\le k} |\lambda_{i_j} - \mu_j| &\le \|\mathscr{R}(D)\|_2,\\ \sqrt{\sum_{1\le j\le k} |\lambda_{i_j} - \mu_j|^2} &\le \sqrt{2\|\mathscr{R}(D)\|_{\rm F}^2 - \|X^{\rm H}\mathscr{R}(D)\|_{\rm F}}\,.\end{aligned}$$

Similar inequalities in a unitarily invariant norm can be derived, too [25].

Finally, for (nonnormal) $C \in \mathbb{C}^{n\times n}$ with the availability of both left and right approximate invariant subspaces, Kahan, Parlett, and Jiang [10] analyzed the backward perturbation and the residuals for a given quaternary $(X_{\rm L}, X_{\rm R}, D_{\rm L}, D_{\rm R})$, where $(D_{\rm R}, X_{\rm R})$ and $(D_{\rm L}, X_{\rm L})$ are approximate right and left eigen-matrix pairs of C, respectively, and $X_{\rm L}, X_{\rm R} \in \mathbb{C}^{n\times k}$ have orthonormal columns. Let

$$\begin{aligned}\mathbb{E}_3 &:= \big\{E \in \mathbb{C}^{n\times n} : (C+E)X_{\rm R} = X_{\rm R}D_{\rm R} \quad\text{and}\quad X_{\rm L}^{\rm H}(C+E) = D_{\rm L}X_{\rm L}^{\rm H}\big\},\\ \mathscr{R}_{\rm R}(D_{\rm R}) &= X_{\rm R}D_{\rm R} - CX_{\rm R}, \qquad \mathscr{R}_{\rm L}(D_{\rm L}) = D_{\rm L}X_{\rm L}^{\rm H} - X_{\rm L}^{\rm H}C.\end{aligned}$$

It is shown that [10] $\mathbb{E}_3 \ne \varnothing$ if and only if $D_{\rm R} = (X_{\rm L}^{\rm H}X_{\rm R})^{-1}D_{\rm L}(X_{\rm L}^{\rm H}X_{\rm R})$ and

$$\eta_2(X_{\rm L}, X_{\rm R}, D_{\rm L}, D_{\rm R}) := \min_{E\in\mathbb{E}_3} \|E\|_2 = \max\{\|\mathscr{R}_{\rm L}(D_{\rm L})\|_2, \|\mathscr{R}_{\rm R}(D_{\rm R})\|_2\},$$

$$\eta_{\rm F}(X_{\rm L},X_{\rm R},D_{\rm L},D_{\rm R}) := \min_{E\in\mathbb{E}_3}\|E\|_{\rm F} = \sqrt{\|\mathscr{R}_{\rm L}(D_{\rm L})\|_{\rm F}^2+\|\mathscr{R}_{\rm R}(D_{\rm R})\|_{\rm F}^2-\|X_{\rm L}^{\rm H}\mathscr{R}_{\rm R}(D_{\rm R})\|_{\rm F}^2}.$$

In this case, the Rayleigh quotient matrices

$$D_{\rm R;RQ}=(X_{\rm L}^{\rm H}X_{\rm R})^{-1}X_{\rm L}^{\rm H}CX_{\rm R},\quad D_{\rm L;RQ}=X_{\rm L}^{\rm H}CX_{\rm R}(X_{\rm L}^{\rm H}X_{\rm R})^{-1}$$

in general do not achieve the minimum of either η_2 or $\eta_{\rm F}$ any more over all possible $D_{\rm R}$ and $D_{\rm L}$.

6 Concluding remarks

In approximations for the standard eigenvalue problem, much attention was drawn to investigate the approximation accuracy by a given approximate invariant subspace in the past. Numerous results some of which are reviewed in section 5 have been obtained and can be found in, e.g., [20, 22, 25, 27] and references therein. They are particularly important for today's large scale eigenvalue problems because often it is an approximate invariant subspace that gets computed first and then the interested eigenvalues/eigenvectors are then extracted from projecting the problems by approximate invariant subspaces into much smaller eigenvalue problems.

While the linear response eigenvalue problem (1.1) is a standard eigenvalue problem, it has its own block and symmetry structures that are not exploited in the existing theory. Keeping these special structures in mind, in this paper, we have developed a backward perturbation analysis and error bounds for the approximation accuracy of eigenvalues generated by a pair of approximate deflating subspaces or eigen-quaternary. Our results are specific for LREP and cannot be derived from the existing ones such as those in section 5, and they are useful for convergence analysis and designing stopping criteria for iterative methods for LREP.

We have assumed so far that K and M in LREP (1.1) under investigation are real and symmetric, beside assumptions on their definiteness. We remark that all results are valid for complex Hermitian K and M, with the same assumptions on their definiteness, after minor changes: replacing all $\mathbb{R}$ by $\mathbb{C}$ and all superscripts $(\cdot)^{\rm T}$ by complex conjugate transposes $(\cdot)^{\rm H}$.

References

[1] Zhaojun Bai and Ren-Cang Li. Minimization principle for linear response eigenvalue problem, I: Theory. *SIAM J. Matrix Anal. Appl.*, 33(4):1075–1100, 2012.

[2] Zhaojun Bai and Ren-Cang Li. Minimization principles for the linear response eigenvalue problem II: Computation. *SIAM J. Matrix Anal. Appl.*, 34(2):392–416, 2013.

[3] P. Benner, V. Mehrmann, and H. Xu. Perturbation analysis for the eigenvalue problem of a formal product of matrices. *BIT*, 42(1):1–43, 2002.

[4] R. Bhatia. *Matrix Analysis.* Graduate Texts in Mathematics, vol. 169. Springer, New York, 1996.

[5] R. Bhatia and Ren-Cang Li. On perturbations of matrix pencils with real spectra. II. *Math. Comp.*, 65(214):637–645, 1996.

[6] D. C. Dzeng and W. W. Lin. Homotopy continuation method for the numerical solutions of generalised symmetric eigenvalue problems. *J. Austral. Math. Soc. Ser. B*, 32:437–456, 4 1991.

[7] R. Granat, B. Kågström, and D. Kressner. Computing periodic deflating subspaces associated with a specified set of eigenvalues. *BIT*, 43(1):1–18, 2003.

[8] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, Cambridge, 1985.

[9] W. Kahan. Inclusion theorems for clusters of eigenvalues of Hermitian matrices. Technical report, Computer Scinence Department, University of Toronoto, 1967.

[10] W. Kahan, B. N. Parlett, and E. Jiang. Residual bounds on approximate eigensystems of nonnormal matrices. *SIAM J. Numer. Anal.*, 19:470–484, 1982.

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

[12] M. G. Kreǐn. The theory of self-adjoint extensions of semi-bounded Hermitian transformations and its applications. *Mat. Sb.*, 20:431–495, 1947.

[13] P. Lancaster and Q. Ye. Variational properties and Rayleigh quotient algorithms for symmetric matrix pencils. *Oper. Theory: Adv. Appl.*, 40:247–278, 1989.

[14] Ren-Cang Li. On perturbations of matrix pencils with real spectra. *Math. Comp.*, 62:231–265, 1994.

[15] Ren-Cang Li. On perturbations of matrix pencils with real spectra, a revisit. *Math. Comp.*, 72:715–728, 2003.

[16] Xin Liang, Ren-Cang Li, and Zhaojun Bai. Trace minimization principles for positive semi-definite pencils. *Linear Algebra Appl.*, 438:3085–3106, 2013.

[17] Xin Liang and Ren-Cang Li. The hyperbolic quadratic eigenvalue problem. Technical Report 2014-01, Department of Mathematics, University of Texas at Arlington, January 2014. Available at `http://www.uta.edu/math/preprint/`.

[18] Wen-Wei Lin, P. van Dooren, and Q.-F. Xu. Equivalent characterizations of periodical invariant subspaces. NCTS Preprints Series 1998-8, National Center for Theoretical Sciences, Math. Division, National Tsing Hua University, Hsinchu, Taiwan, 1998.

[19] Wen-Wei Lin and Ji guang Sun. Perturbation analysis for the eigenproblem of periodic matrix pairs. *Linear Algebra Appl.*, 337(13):157–187, 2001.

[20] B. N. Parlett. *The Symmetric Eigenvalue Problem.* SIAM, Philadelphia, 1998.

[21] G. W. Stewart. Perturbation bounds for the definite generalized eigenvalue problem. *Linear Algebra Appl.*, 23:69–86, 1979.

[22] G. W. Stewart and Ji-Guang Sun. *Matrix Perturbation Theory.* Academic Press, Boston, 1990.

[23] Ji-Guang Sun. A note on Stewart's theorem for definite matrix pairs. *Linear Algebra Appl.*, 48:331–339, 1982.

[24] Ji-Guang Sun. Perturbation bounds for eigenspaces of a definite matrix pair. *Numer. Math.*, 41:321–343, 1983.

[25] Ji-Guang Sun. Backward perturbation analysis of certain characteristic subspaces. *Numer. Math.*, 65:357–382, 1993.

[26] Ji-Guang Sun. A note on backward perturbations for the Hermitian eigenvalue problem. *BIT*, 35:385–393, 1995.

[27] Ji-Guang Sun. Stability and accuracy: Perturbation analysis of algebraic eigenproblems. Technical Report UMINF 1998-07, ISSN-0348-0542, Faculty of Science and Technology, Department of Computing Science, Umeå universitet, Augest 1998. http://www8.cs.umu.se/research/reports/showl.cgi?year=1998.

[28] Zhongming Teng and Ren-Cang Li. Convergence analysis of Lanczos-type methods for the linear response eigenvalue problem. *J. Comput. Appl. Math.*, 247:17–33, 2013.

[29] Zhongming Teng, Yunkai Zhou, and Ren-Cang Li. A block chebyshev-davidson method for linear response eigenvalue problems. Technical Report 2013-11, Department of Mathematics, University of Texas at Arlington, September 2013. Available at `http://www.uta.edu/math/preprint/`, submited.

[30] D. J. Thouless. Vibrational states of nuclei in the random phase approximation. *Nuclear Physics*, 22(1):78–95, 1961.

[31] D. J. Thouless. *The quantum mechanics of many-body systems.* Academic, 1972.

[32] E. V. Tsiper. Variational procedure and generalized Lanczos recursion for small-amplitude classical oscillations. *JETP Letters*, 70(11):751–755, 1999.

[33] E. V. Tsiper. A classical mechanics technique for quantum linear response. *J. Phys. B: At. Mol. Opt. Phys.*, 34(12):L401–L407, 2001.

[34] L.-H. Zhang, J.-G. Xue, and R.-C. Li. Rayleigh-Ritz approximation for the linear response eigenvalue problem. Technical Report 2013-12, Department of Mathematics, University of Texas at Arlington, November 2013. Available at `http://www.uta.edu/math/preprint/`, submited.