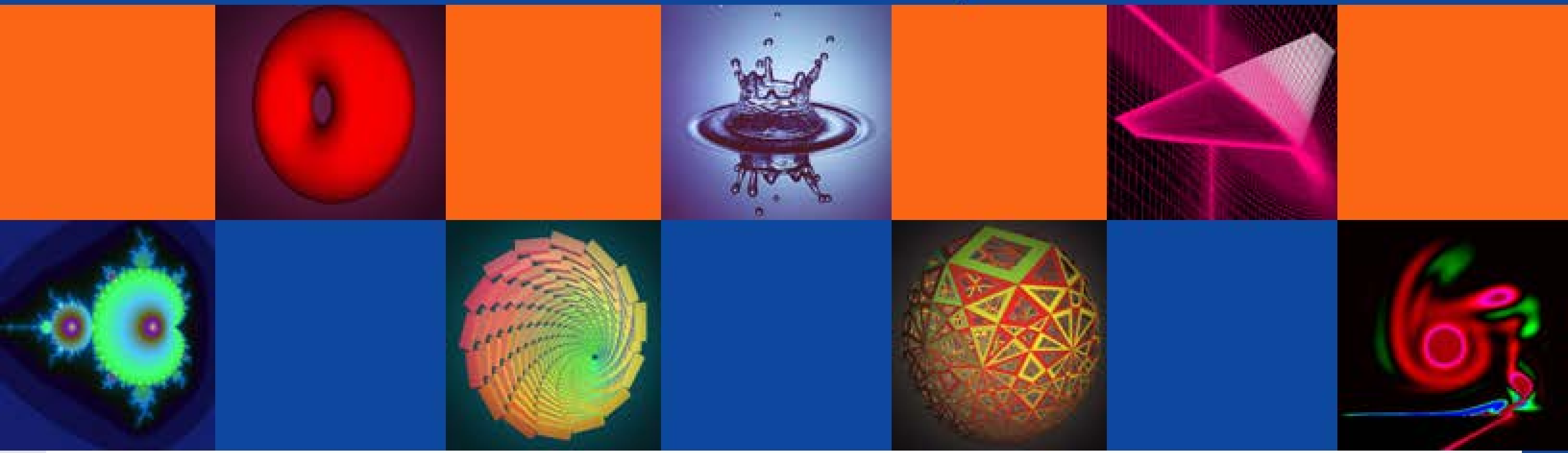
The University of Texas ARLINGTON

MATHEMATICS
Preprint Series



The Hyperbolic Quadratic Eigenvalue Problem

Xin Liang
Ren-Cang Li

Technical Report 2014-01

http://www.uta.edu/math/preprint/

The Hyperbolic Quadratic Eigenvalue Problem

Xin Liang* Ren-Cang Li†

January 8, 2014

Abstract

The hyperbolic quadratic eigenvalue problem (HQEP) was shown to admit the Courant-Fischer type min-max principles in 1955 by Duffin and Cauchy type interlacing inequalities in 2010 by Veselić. It can be regarded as the closest analogue (among all kinds of quadratic eigenvalue problems) to the standard Hermitian eigenvalue problem (among all kinds of standard eigenvalue problems). In this paper, we conduct a systematic study on HQEP both theoretically and numerically. In the theoretic front, we generalize Wiedlandt-Lidskii type min-max principles and, as a special case, Ky-Fan type trace min/max principles and establish Weyl type and Mirsky type perturbation results when an HQEP is perturbed to another HQEP. In the numerical front, we justify the natural generalization of the Rayleigh-Ritz procedure with the existing and our new optimization principles and, as consequences of these principles, we extend various current optimization approaches – steepest descent/ascent and nonlinear conjugate gradient type methods for the Hermitian eigenvalue problem – to calculate few extreme quadratic eigenvalues (of both pos- and neg-type). A detailed convergent analysis is given on the steepest descent/ascent methods. The analysis reveals the intrinsic quantities that control convergence rates and consequently yields ways of constructing effective preconditioners. Numerical examples are presented to demonstrate the proposed theory and algorithms.

Key words. Hyperbolic quadratic eigenvalue problem, Rayleigh quotient, min-max principle, Cauchy interlacing inequality, eigenvalue perturbation, extended steepest descent/ascent method, locally optimal extended conjugate gradient method, preconditioning

AMS subject classifications. 15A18, 15A42, 65F08, 65F30, 65G99

---

*School of Mathematical Sciences, Peking University, Beijing, 100871, P. R. China. E-mail: `liangxinslm@pku.edu.cn`. Supported in part by China Scholarship Council and National Natural Science Foundation of China NSFC-61075119. This work is primarily done while this author was a visiting student, from August 2011 to September 2013, at Department of Mathematics, University of Texas at Arlington, Arlington, TX 76019.

†Department of Mathematics, University of Texas at Arlington, P.O. Box 19408, Arlington, TX 76019. E-mail: `rcli@uta.edu`. Supported in part by NSF grants DMS-1115834 and DMS-1317330, and a Research Gift Grant from Intel Corporation.

Contents

1 Introduction

It was argued in [26] that the hyperbolic quadratic eigenvalue problem (HQEP) is the closest analogue of the standard Hermitian eigenvalue problem when it comes to the quadratic eigenvalue problem (QEP)

$$(\lambda^2 A + \lambda B + C)x = 0. \tag{1.1}$$

In many ways, both problems share common properties: the eigenvalues are all real, and for HQEP there is a version of the min-max principles [12, 1955] that is very much like the Courant-Fischer min-max principles.

One source of QEPs (1.1) is dynamical systems with friction, where A, C are associated with the kinetic-energy and potential-energy quadratic form, respectively, and B is associated with the Rayleigh dissipation function [16, 65]. When A, B, and C are Hermitian, and A and B are positive definite and C positive semidefinite, we say the dynamical system is *overdamped* if

$$(x^{\mathrm{H}}Bx)^2 - 4(x^{\mathrm{H}}Ax)(x^{\mathrm{H}}Cx) > 0 \quad \text{for any nonzero vector } x.$$

Overdamped dynamical systems are common in elevator and car braking systems[1]. A HQEP is slightly more general than an overdamped QEP in that B and C are no longer required positive definite or positive semidefinite, respectively. However, a a suitable shift in λ can turn a HQEP into an overdamped QEP [20].

If (1.1) is satisfied for a scalar λ and nonzero vector x, we call λ a *quadratic eigenvalue*, x an associated *quadratic eigenvector*, and (λ, x) a *quadratic eigenpair*.

In this paper, we will launch a systematic study of the HQEP both in theory and numerical computations that will further reinforce the belief that this class of QEP is the closest analogue to the standard Hermitian eigenvalue problem. In the theoretical front, we will

- review existing results of Courant-Fischer type min-max principles, Cauchy interlacing inequalities;
- establish Wielandt-Lidskii type min-max principles for the sums of selected quadratic eigenvalues and, as corollaries, trace min/max type principles;
- establish perturbation results in the spectral and Frobenius norm, as well as general unitarily invariant norms on how the quadratic eigenvalues will change if A, B, C are perturbed.

In the numerical front, we will

- justify a naturally extended Rayleigh-Ritz type procedure, with the existing and newly established min-max principles, why the procedure will produce the best approximations to quadratic eigenvalues/eigenvectors;
- propose extended steepest descent/ascent and CG type methods for computing extreme quadratic eigenpairs;

[1] W. Kahan, private cmmunications, November 2013.

- establish convergence results, including the rate of convergence for the extended steepest descent/ascent methods, which shed light on preconditioning in what constitutes a good preconditioner and how to construct one.

In a separate paper, we will extend most of the development in this paper to the hyperbolic polynomial eigenvalue problem.

The rest of this paper is organized as follows. In section 2, we collect some properties for hyperbolic quadratic matrix polynomials and establish a few more about an HQEP. Wielandt-Lidskii type min-max principles, among others, are given in section 3. Eigenperturbation analysis for HQEP is done in section 4. In section 5, we justify the use of the Rayleigh-Ritz procedure for extracting interested quadratic eigenvalues and their associated quadratic eigenvectors within a given subspace. The steepest descent/ascent method and its extended variation are studied in section 6, where a detailed convergence analysis is performed. Section 7 investigates the preconditioning techniques to speed up the extended steepest descent/ascent method and explain how an effective preconditioner should be constructed from two different perspectives. Section 8 introduces the block variations of the methods in the previous two sections. Various conjugate gradient methods – the plain, locally optimal, and extended subspace search versions combined with suitable preconditoners and blocking – are described in detail in section 9. Two numerical examples are presented in section 10 to demonstrate the effectiveness of the locally optimal block preconditioned conjugate gradient method in the previous section. Finally in section 11, we present our concluding remarks. In appendix section A, we review the Jordan canonical form of a positive semidefinite matrix pencil and establish a perturbation theory for a positive definite matrix pencil for use in section 4.

Notation. Throughout this paper, $\mathbb{C}^{n\times m}$ is the set of all $n\times m$ complex matrices, $\mathbb{C}^n=\mathbb{C}^{n\times 1}$, and $\mathbb{C}=\mathbb{C}^1$. $\mathbb{R}$ is the set of all real numbers. I_n (or simply I if its dimension is clear from the context) is the $n\times n$ identity matrix, and e_j is its jth column. $X^{\rm H}$ is the conjugate transpose of a vector or matrix. For $X\in\mathbb{C}^{n\times m}$, $\sigma_{\min}(X)$ is the smallest singular value of X (X has $\min\{m,n\}$ singular values), $\|X\|_2$ and $\|X\|_{\rm F}$ and $\|X\|_{\rm ui}$ are the spectral, Frobenius, and a general unitarily invariant norm of X, and $\kappa_2(X)=\|X\|_2\|X^{-1}\|_2$ is the condition number of X.

$A\succ 0$ ($A\succeq 0$) means that A is Hermitian positive (semi-)definite, and $A\prec 0$ ($A\preceq 0$) if $-A\succ 0$ ($-A\succeq 0$). $A^{1/2}\succeq 0$ is the unique square root of $A\succeq 0$.

The integer triplet $(i_-(H), i_0(H), i_+(H))$ denotes the inertia of an Hermitian matrix H, meaning that H has $i_-(H)$ negative, $i_0(H)$ zero, and $i_+(H)$ positive eigenvalues, respectively, and $\lambda_{\min}(H)$ and $\lambda_{\max}(H)$ are its smallest and largest eigenvalue.

Generic notation $\mathrm{eig}(\,\cdot\,)$ is the set of all eigenvalues, counting algebraic multiplicities, of a matrix or a matrix pencil, depending on its argument(s): $\mathrm{eig}(A)$ is for A, and $\mathrm{eig}(A,B)$ is for $A-\lambda B$. We use $\mathrm{polyeig}(A_0,A_1,\cdots,A_k)$ as MATLAB's function `polyeig` for the set of all polynomial eigenvalues of $\lambda^kA_k+\cdots+\lambda A_1+A_0$. Note $\mathrm{polyeig}(A_0,A_1)$ is not the same of $\mathrm{eig}(A_0,A_1)$.

2 Hyperbolic quadratic matrix polynomial

Given $A, B, C \in \mathbb{C}^{n\times n}$, define

$$\boldsymbol{Q}(\lambda) := \lambda^2 A + \lambda B + C, \tag{2.1}$$

a quadratic matrix polynomial of order n.

Definition 2.1. $\boldsymbol{Q}(\lambda)$ is said *Hermitian* if A, B, and C are all Hermitian, *hyperbolic* if it is Hermitian, $A \succ 0$, and

$$(x^{\rm H}Bx)^2 - 4(x^{\rm H}Ax)(x^{\rm H}Cx) > 0, \quad \text{for all } 0 \neq x \in \mathbb{C}^n, \tag{2.2}$$

overdamped if it is hyperbolic as well as $B \succ 0, C \succeq 0$. For a hyperbolic $\boldsymbol{Q}(\lambda)$, define

$$\varsigma(x) := \left[(x^{\rm H}Bx)^2 - 4(x^{\rm H}Ax)(x^{\rm H}Cx)\right]^{1/2}, \quad \varsigma_0(x) := \frac{\varsigma(x)}{x^{\rm H}x}. \tag{2.3}$$

The quadratic eigenvalue problem (QEP) for $\boldsymbol{Q}(\,\cdot\,)$ is to find $\lambda \in \mathbb{C}$ and $0 \neq x \in \mathbb{C}^n$ such that

$$\boldsymbol{Q}(\lambda)x = 0.$$

When this equation is satisfied, λ is called a *quadratic eigenvalue* and x the associated *quadratic eigenvector*. Evidently all quadratic eigenvalues of $\boldsymbol{Q}(\,\cdot\,)$ is the roots of $\det \boldsymbol{Q}(\lambda) = 0$ which has $2n$ (complex) roots, counting multiplicities.

The next theorem summarizes some of the relevant theoretical results on hyperbolic quadratic polynomials. They can be found in Guo and Lancaster [20] which is an excellent gateway to references of origins for these results. Item 3(c) can be found in [64, (0.7)].

Theorem 2.1. *Let $\boldsymbol{Q}(\lambda) = \lambda^2 A + \lambda B + C$ as in (2.1) be Hermitian with $A \succ 0$.*

1. *$\boldsymbol{Q}(\lambda)$ is hyperbolic if and only if there exists $\lambda_0 \in \mathbb{R}$ such that $\boldsymbol{Q}(\lambda_0) \prec 0$.*

2. *If $\boldsymbol{Q}(\lambda)$ is hyperbolic, then its quadratic eigenvalues are all real.*

3. *Suppose $\boldsymbol{Q}(\lambda)$ is hyperbolic. Denote its quadratic eigenvalues by $\lambda_i^{\pm}$ and arrange them in the order of*

$$\lambda_1^- \le \cdots \le \lambda_n^- < \lambda_1^+ \le \cdots \le \lambda_n^+. \tag{2.4}$$

   Then

   (a) $\boldsymbol{Q}(\lambda) \prec 0$ for all $\lambda \in (\lambda_n^-, \lambda_1^+)$;
   (b) $\boldsymbol{Q}(\lambda) \succ 0$ for all $\lambda \in (-\infty, \lambda_1^-) \cup (\lambda_n^+, +\infty)$;
   (c) the inertia of $\boldsymbol{Q}(\lambda)$ is $(n-k, 0, k)$ for $\lambda \in (\lambda_k^+, \lambda_{k+1}^+)$ or $\lambda \in (\lambda_{n-k}^-, \lambda_{n+1-k}^-)$ for $k = 1, \cdots, n$, concluding that $\boldsymbol{Q}(\lambda)$ is indefinite for $\lambda \in (\lambda_1^{\rm typ}, \lambda_n^{\rm typ})$;
   (d) $\boldsymbol{Q}(\lambda)$ is overdamped if and only if $\lambda_n^+ \le 0$.

An immediate consequence of Theorem 2.1 is a test to determine whether $\boldsymbol{Q}(\lambda)$ is hyperbolic or not [20]: check if its quadratic eigenvalues are all real and, in the case they are all real, check if $\boldsymbol{Q}(\lambda_0) \prec 0$, where $\lambda_0 = (\lambda_n^- + \lambda_1^+)/2$.

A common technique of solving QEP (1.1), or more generally the polynomial eigenvalue problem, is *linearization* that converts a polynomial eigenvalue problem to an equivalent generalized (linear) eigenvalue problem of a matrix pencil [16, 25, 42].

Under the condition that A is nonsingular, QEP (1.1) is equivalent to the generalized eigenvalue problem of the following matrix pencil

$$\mathscr{L}_{\boldsymbol{Q}}(\lambda) := \begin{bmatrix} -C & 0 \\ 0 & A \end{bmatrix} - \lambda \begin{bmatrix} B & A \\ A & 0 \end{bmatrix} = \mathscr{A} - \lambda \mathscr{B}, \tag{2.5}$$

or

$$\mathscr{K}_{\boldsymbol{Q}}(\lambda) := \begin{bmatrix} 0 & -C \\ -C & -B \end{bmatrix} - \lambda \begin{bmatrix} -C & 0 \\ 0 & A \end{bmatrix} = \mathscr{A} - \lambda \mathscr{B} \tag{2.6}$$

in the sense that $\operatorname{polyeig}(C, B, A) = \operatorname{eig}(\mathscr{A}, \mathscr{B})$ and associated eigenvectors of one can be recovered from those for the other. More can be said if $\boldsymbol{Q}(\lambda) = \lambda^2 A + \lambda B + C$ is hyperbolic. Relevant results are summarized in the following lemma, where item 5 is essentially in [4] (see also [9], [26, Theorem 3.6], and [63, Theorem 5A]).

Theorem 2.2. *Let $\boldsymbol{Q}(\lambda) = \lambda^2 A + \lambda B + C$ as in (2.1) and let $\mathscr{L}_{\boldsymbol{Q}}(\lambda)$ be as in (2.5). Suppose A is nonsingular.*

1. $\operatorname{polyeig}(C, B, A) = \operatorname{eig}(\mathscr{A}, \mathscr{B})$.
2. *If $A \succ 0$ and B is Hermitian, then the inertia of $\mathscr{B}$ is $(n, 0, n)$.*
3. *If (μ, x) is an eigenpair of $\boldsymbol{Q}(\lambda)$, then $(\mu, \begin{bmatrix} x \\ \mu x \end{bmatrix})$ is an eigenpair of $\mathscr{L}_{\boldsymbol{Q}}(\lambda)$.*
4. *If $(\mu, \begin{bmatrix} x \\ y \end{bmatrix})$ is an eigenpair of $\mathscr{L}_{\boldsymbol{Q}}(\lambda)$, then (μ, x) is an eigenpair of $\boldsymbol{Q}(\lambda)$ and $y = \mu x$.*
5. *Suppose $\boldsymbol{Q}(\lambda)$ is Hermitian. $\boldsymbol{Q}(\lambda)$ is hyperbolic if and only if $\mathscr{L}_{\boldsymbol{Q}}(\lambda)$ is a positive definite pencil.*
6. *Suppose $\boldsymbol{Q}(\lambda)$ is hyperbolic, and adopt the notation in* item 3 *of* Theorem 2.1. *Then $\mathscr{L}_{\boldsymbol{Q}}(\lambda) \succ 0$ for all $\lambda \in (\lambda_n^-, \lambda_1^+)$.*

Proof. Since for any $\lambda \in \mathbb{C}$,

$$\begin{bmatrix} I & 0 \\ -\lambda I & I \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} -\boldsymbol{Q}(\lambda) & 0 \\ 0 & A \end{bmatrix} \begin{bmatrix} I & 0 \\ -\lambda I & I \end{bmatrix} = \begin{bmatrix} -C - \lambda B & -\lambda A \\ -\lambda A & A \end{bmatrix} = \mathscr{L}_{\boldsymbol{Q}}(\lambda). \tag{2.7}$$

Thus $(-1)^n \det \boldsymbol{Q}(\lambda) \cdot \det A \equiv \det \mathscr{L}_{\boldsymbol{Q}}(\lambda)$ and item 1 follows. For item 2, $A \succ 0$ guarantees that there is a nonsingular matrix $X \in \mathbb{C}^{n \times n}$ such that

$$X^{\mathrm{H}} A X = I_n, \quad X^{\mathrm{H}} B X = \operatorname{diag}(\omega_1, \ldots, \omega_n) =: \mathit{\Omega},$$

where $\omega_i \in \mathbb{R}$. We have

$$\begin{bmatrix} X & \\ & X \end{bmatrix}^{\mathrm{H}} \mathscr{B} \begin{bmatrix} X & \\ & X \end{bmatrix} = \begin{bmatrix} \Omega & I_n \\ I_n & 0 \end{bmatrix} \tag{2.8}$$

whose eigenvalues are the union of all the eigenvalues of

$$\begin{bmatrix} \omega_i & 1 \\ 1 & 0 \end{bmatrix} \quad \text{for } i = 1, 2, \ldots, n.$$

But the two eigenvalues of each one of these 2×2 matrices are

$$\frac{\omega_i - \sqrt{\omega_i^2 + 4}}{2} < 0, \quad \frac{\omega_i + \sqrt{\omega_i^2 + 4}}{2} > 0.$$

Therefore the last matrix in (2.8) has n positive and n negative eigenvalues, as expected. Items 3 and 4 can be verified in a straightforward way by using (2.7). Also by using (2.7), we see that $\mathrm{diag}(-\boldsymbol{Q}(\lambda), A)$ and $\mathscr{L}_{\boldsymbol{Q}}(\lambda)$ are congruent for all $\lambda \in \mathbb{R}$, and hence items 5 and 6 follow from items 1 and 3(a) of Theorem 2.1, respectively. $\square$

One consequence of Theorem 2.2 is that any hyperbolic $\boldsymbol{Q}(\lambda) = \lambda^2 A + \lambda B + C$ gives rise to a positive definite matrix pencil $\mathscr{L}_{\boldsymbol{Q}}(\lambda)$ as defined by (2.5) with $\mathscr{B}$ having inertia $(n, 0, n)$. There is a converse to the statement, too.

Theorem 2.3. *Let $L(\lambda) = \mathscr{A} - \lambda\mathscr{B}$ be a positive definite Hermitian pair of order $2n$. If the inertia of $\mathscr{B}$ is $(n, 0, n)$, then there exists a hyperbolic $\boldsymbol{Q}(\lambda) = \lambda^2 A + \lambda B + C$ and a nonsingular matrix $U \in \mathbb{C}^{2n\times 2n}$ such that the following statements are true.*

1. *If (μ, x) is a quadratic eigenpair of $\boldsymbol{Q}(\lambda)$, then $(\mu, U \begin{bmatrix} x \\ \mu x \end{bmatrix})$ is an eigenpair of $L(\lambda)$.*
2. *If $(\mu, \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix})$ is an eigenpair of $L(\lambda)$ and we define $\begin{bmatrix} x \\ y \end{bmatrix} = U^{-1} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix}$, where $x \in \mathbb{C}^n$, then (μ, x) is a quadratic eigenpair of $\boldsymbol{Q}(\lambda)$ and $y = \mu x$.*

Proof. Since $L(\lambda)$ is positive definite and the inertia of $\mathscr{B}$ is $(n, 0, n)$, by Theorem A.1 there exists a nonsingular matrix W such that $W^{\mathrm{H}}\mathscr{A}W = \mathrm{diag}(\Lambda_+, -\Lambda_-)$ and $W^{\mathrm{H}}\mathscr{B}W = \mathrm{diag}(I, -I)$, where $\Lambda_+ = \mathrm{diag}(\lambda_1^+, \cdots, \lambda_n^+), \Lambda_- = \mathrm{diag}(\lambda_1^-, \cdots, \lambda_n^-)$ and $\lambda_i^{\pm} \in \mathbb{R}$ and $\lambda_i^+ > \lambda_j^-$ for all i and j. Set

$$A = I, \quad B = -(\Lambda_+ + \Lambda_-), \quad C = \Lambda_+\Lambda_-,$$
$$S = \begin{bmatrix} \Lambda_- & -I \\ \Lambda_+ & -I \end{bmatrix} \begin{bmatrix} (\Lambda_+ - \Lambda_-)^{-1/2} & 0 \\ 0 & (\Lambda_+ - \Lambda_-)^{-1/2} \end{bmatrix},$$

and $\boldsymbol{Q}(\lambda) = \lambda^2 A + \lambda B + C$. It can be verified that corresponding to this $\boldsymbol{Q}(\lambda)$, $\mathscr{L}_{\boldsymbol{Q}}(\lambda)$ of (2.5) satisfies $\mathscr{L}_{\boldsymbol{Q}}(\lambda) = S^{\mathrm{H}}W^{\mathrm{H}}L(\lambda)WS$. Since $L(\lambda)$ is positive definite, there is a $\lambda_0 \in \mathbb{R}$ such that $L(\lambda_0) \succ 0$ which implies $\mathscr{L}_{\boldsymbol{Q}}(\lambda_0) \succ 0$ and thus $\boldsymbol{Q}(\lambda_0) \prec 0$ by (2.7). Consequently, this $\boldsymbol{Q}(\lambda)$ is hyperbolic by item 1 of Theorem 2.1. Finally take $U = WS$ for items 1 and 2. $\square$

Theorem 2.4. *Let $\boldsymbol{Q}(\lambda) = \lambda^2 A + \lambda B + C$ be hyperbolic. Then for any $X \in \mathbb{C}^{n\times m}$ satisfying $X^{\mathrm{H}}AX = I_m$,*

$$(X^{\mathrm{H}}BX)^2 - 4(X^{\mathrm{H}}CX) \succ 0. \tag{2.9}$$

Proof. For any $y \in \mathbb{C}^m$ with $\|y\|_2 = 1$, write $x = Xy$. We have

$$\begin{aligned}
&y^{\mathrm{H}}\left[(X^{\mathrm{H}}BX)^2 - 4(X^{\mathrm{H}}CX)\right]y \\
&\quad= (X^{\mathrm{H}}BXy)^{\mathrm{H}}(X^{\mathrm{H}}BXy) - 4(Xy)^{\mathrm{H}}C(Xy) \\
&\quad= \|y\|_2^2 \cdot \|X^{\mathrm{H}}BXy\|_2^2 - 4(Xy)^{\mathrm{H}}C(Xy)\cdot y^{\mathrm{H}}(X^{\mathrm{H}}AX)y && (2.10)\\
&\quad\ge \left[y^{\mathrm{H}}(X^{\mathrm{H}}BXy)\right]^2 - 4(Xy)^{\mathrm{H}}C(Xy)\cdot(Xy)^{\mathrm{H}}A(Xy) && (2.11)\\
&\quad= (x^{\mathrm{H}}Bx)^2 - 4x^{\mathrm{H}}Cx\cdot x^{\mathrm{H}}Ax \\
&\quad> 0, && (2.12)
\end{aligned}$$

where we have used $\|y\|_2 = 1$ and $X^{\mathrm{H}}AX = I_m$ for (2.10), and used the Cauchy-Bunyakovsky-Schwarz inequality for (2.11). Therefore $(X^{\mathrm{H}}BX)^2 - 4(X^{\mathrm{H}}CX) \succ 0$ by (2.12). $\square$

Theorem 2.5. *Let $\boldsymbol{Q}(\lambda) = \lambda^2 A + \lambda B + C$ be a hyperbolic quadratic matrix polynomial of order n, and denote by $\lambda_i^{\pm}$ its quadratic eigenvalues which are arranged as in* (2.4)*. Set*

$$\varLambda_+ = \operatorname{diag}(\lambda_1^+, \cdots, \lambda_n^+), \quad \varLambda_- = \operatorname{diag}(\lambda_1^-, \cdots, \lambda_n^-). \tag{2.13}$$

Then there exists nonsingular $Z \in \mathbb{C}^{2n\times 2n}$ of the form

$$Z = \begin{bmatrix} U_+ & U_- \\ U_+\varLambda_+ & U_-\varLambda_- \end{bmatrix}, \tag{2.14}$$

where U_+, $U_- \in \mathbb{C}^{n\times n}$ are nonsingular and

$$\varUpsilon := U_+^{-1}U_- \tag{2.15}$$

is unitary, such that

$$Z^{\mathrm{H}}\mathscr{A}Z = Z^{\mathrm{H}}\begin{bmatrix} -C & \\ & A \end{bmatrix}Z = \begin{bmatrix} \varLambda_+ & \\ & -\varLambda_- \end{bmatrix}, \tag{2.16a}$$

$$Z^{\mathrm{H}}\mathscr{B}Z = Z^{\mathrm{H}}\begin{bmatrix} B & A \\ A & \end{bmatrix}Z = \begin{bmatrix} I_n & \\ & -I_n \end{bmatrix}. \tag{2.16b}$$

Write

$$U_+ = [u_1^+, u_2^+, \ldots, u_n^+], \quad U_- = [u_1^-, u_2^-, \ldots, u_n^-].$$

As a consequence of (2.14) *and* (2.16)*, we have the following statements.*

1. *$\boldsymbol{Q}(\lambda_i^+)u_i^+ = 0$, $\boldsymbol{Q}(\lambda_i^-)u_i^- = 0$ for $i = 1, 2, \cdots, n$. Thus there are n linearly independent quadratic eigenvectors associated with all λ_i^+, and the same can be said about quadratic eigenvectors associated with all λ_i^-.*

2. *$\varsigma(u_i^{\pm}) = 1$ for $i = 1, 2, \ldots, n$.*

3. $\boldsymbol{Q}(\lambda)$ *admits*

$$\boldsymbol{Q}(\lambda) = U_-^{-\mathrm{H}}(\lambda I - \mathit{\Lambda}_-)U_-^{\mathrm{H}}AU_+(\lambda I - \mathit{\Lambda}_+)U_+^{-1}, \tag{2.17a}$$

$$\boldsymbol{Q}(\lambda) = U_+^{-\mathrm{H}}(\lambda I - \mathit{\Lambda}_+)U_+^{\mathrm{H}}AU_-(\lambda I - \mathit{\Lambda}_-)U_-^{-1}. \tag{2.17b}$$

4. $U_-^{\mathrm{H}}AU_+ = (\mathit{\Lambda}_+\mathit{\Upsilon} - \mathit{\Upsilon}\mathit{\Lambda}_-)^{-1}$. *As a result,* A, B, C *and* $\boldsymbol{Q}(\lambda)$ *can be expressed in terms of* $\mathit{\Lambda}_\pm$ *and any two of* U_+, U_-, *and* $\mathit{\Upsilon}$, *assuming* (2.15). *In particular,*

$$A = U_+^{-\mathrm{H}}\mathit{\Theta}U_+^{-1}, \tag{2.18a}$$

$$B = U_+^{-\mathrm{H}}(I - \mathit{\Theta}\mathit{\Lambda}_+ - \mathit{\Lambda}_+\mathit{\Theta})U_+^{-1}, \tag{2.18b}$$

$$C = U_+^{-\mathrm{H}}(\mathit{\Lambda}_+\mathit{\Theta}\mathit{\Lambda}_+ - \mathit{\Lambda}_+)U_+^{-1}, \tag{2.18c}$$

$$\boldsymbol{Q}(\lambda) = U_+^{-\mathrm{H}}\Big[(\lambda I - \mathit{\Lambda}_+)\mathit{\Theta}(\lambda I - \mathit{\Lambda}_+) + (\lambda I - \mathit{\Lambda}_+)\Big]U_+^{-1}, \tag{2.18d}$$

where

$$\mathit{\Theta} = (\mathit{\Lambda}_+ - \mathit{\Upsilon}\mathit{\Lambda}_-\mathit{\Upsilon}^{\mathrm{H}})^{-1}. \tag{2.18e}$$

5. *We have*

$$\|U_+\|_2 = \|U_-\|_2 \le \frac{\|A^{-1/2}\|_2}{\sqrt{\lambda_1^+ - \lambda_n^-}}, \tag{2.19a}$$

$$\|U_+^{-1}\|_2 = \|U_-^{-1}\|_2 \le \|A^{1/2}\|_2\sqrt{\lambda_n^+ - \lambda_1^-}, \tag{2.19b}$$

$$\kappa(U_+) = \kappa(U_-) \le \sqrt{\kappa(A)}\sqrt{\frac{\lambda_n^+ - \lambda_1^-}{\lambda_1^+ - \lambda_n^-}}, \tag{2.19c}$$

and

$$\|Z\|_2 \le \mathit{\Xi}\|U_\pm\|_2, \quad \|Z^{-1}\|_2 \le \frac{\mathit{\Xi}}{\lambda_1^+ - \lambda_n^-}\|U_\pm^{-1}\|_2, \tag{2.20}$$

where $\xi_\pm = \max\{|\lambda_1^\pm|, |\lambda_n^\pm|\}$ *and*

$$\mathit{\Xi} = \frac{2 + \xi_+^2 + \xi_-^2 + \sqrt{[(\xi_+ - 1)^2 + (\xi_- + 1)^2][(\xi_+ + 1)^2 + (\xi_- - 1)^2]}}{2}.$$

The following converse to item 4 *is also true: given diagonal matrices* $\mathit{\Lambda}_\pm$ *as in* (2.13) *and two of* U_+, U_-, *and* $\mathit{\Upsilon}$, *where* $\mathit{\Upsilon} \in \mathbb{C}^{n\times n}$ *as in* (2.15) *is unitary and* $U_+, U_- \in \mathbb{C}^{n\times n}$ *are nonsingular, if* $\lambda_i^\pm$ *can be arranged as in* (2.4), *then the quadratic matrix polynomial constructed by* (2.18) *is hyperbolic.*

Proof. Since $\boldsymbol{Q}(\lambda)$ is hyperbolic, $\mathscr{L}_{\boldsymbol{Q}}(\lambda)$ in (2.5) is a positive definite pencil. By Theorem A.1, there exists a nonsingular $Z \in \mathbb{C}^{2n\times 2n}$ to give (2.16). We have to show that Z must take the form (2.14).

Since each column of Z is an eigenvector of the pencil $\mathscr{L}_{\boldsymbol{Q}}(\lambda)$, by Theorem 2.2, we conclude that the ith column of Z can be expressed as $\begin{bmatrix} u_i^+ \\ \lambda_i^+ u_i^+ \end{bmatrix}$ for $1 \le i \le n$ and $\begin{bmatrix} u_j^- \\ \lambda_j^- u_j^- \end{bmatrix}$

for $1 \le j = i - n \le n$, where u_i^+, u_j^- are the corresponding quadratic eigenvectors of $\boldsymbol{Q}(\lambda)$ associated with λ_i^+ and λ_j^-, respectively. Hence Z takes the form (2.14).

Blockwise, the equations in (2.16) yield

$$\begin{aligned} U_+^{\mathrm{H}} C U_+ - \Lambda_+ U_+^{\mathrm{H}} A U_+ \Lambda_+ &= -\Lambda_+, && \text{(2.21a)}\\ U_-^{\mathrm{H}} C U_- - \Lambda_- U_-^{\mathrm{H}} A U_- \Lambda_- &= \ \ \Lambda_-, && \text{(2.21b)}\\ U_+^{\mathrm{H}} C U_- - \Lambda_+ U_+^{\mathrm{H}} A U_- \Lambda_- &= \ \ 0, && \text{(2.21c)}\\ U_+^{\mathrm{H}} B U_+ + U_+^{\mathrm{H}} A U_+ \Lambda_+ + \Lambda_+ U_+^{\mathrm{H}} A U_+ &= \ \ I, && \text{(2.21d)}\\ U_-^{\mathrm{H}} B U_- + U_-^{\mathrm{H}} A U_- \Lambda_- + \Lambda_- U_-^{\mathrm{H}} A U_- &= -I, && \text{(2.21e)}\\ U_+^{\mathrm{H}} B U_- + U_+^{\mathrm{H}} A U_- \Lambda_- + \Lambda_+ U_+^{\mathrm{H}} A U_- &= \ \ 0. && \text{(2.21f)} \end{aligned}$$

We claim that U_+ is nonsingular. Consider $U_+ x = 0$ for some $x \in \mathbb{C}^n$. We will prove that $x = 0$ and thus U_+ is nonsingular. By (2.21d),

$$x^{\mathrm{H}} x = x^{\mathrm{H}} I x = x^{\mathrm{H}} (U_+^{\mathrm{H}} B U_+ + U_+^{\mathrm{H}} A U_+ \Lambda_+ + \Lambda_+ U_+^{\mathrm{H}} A U_+) x = 0$$

which implies $x = 0$, as was to be shown. Similarly, U_- is nonsingular.

Next, we define

$$\widehat{\Lambda}_+ := U_+ \Lambda_+ U_+^{-1}, \quad \widehat{\Lambda}_- := U_- \Lambda_- U_-^{-1}. \tag{2.22}$$

We deduce from (2.21c) and (2.21f) the expressions for C and B in (2.23a) below, and then use $C = C^{\mathrm{H}}$ and $B = B^{\mathrm{H}}$ to get (2.23b).

$$\begin{aligned} C &= \widehat{\Lambda}_-^{\mathrm{H}} A \widehat{\Lambda}_+, \quad B = -A\widehat{\Lambda}_+ - \widehat{\Lambda}_-^{\mathrm{H}} A, && \text{(2.23a)}\\ C &= \widehat{\Lambda}_+^{\mathrm{H}} A \widehat{\Lambda}_-, \quad B = -A\widehat{\Lambda}_- - \widehat{\Lambda}_+^{\mathrm{H}} A. && \text{(2.23b)} \end{aligned}$$

Using the second equation in (2.23a), we deduce from (2.21d) and (2.21e) that

$$\begin{aligned} U_+^{-\mathrm{H}} U_+^{-1} &= \ \ B + A\widehat{\Lambda}_+ + \widehat{\Lambda}_+^{\mathrm{H}} A = (\widehat{\Lambda}_+ - \widehat{\Lambda}_-)^{\mathrm{H}} A,\\ U_-^{-\mathrm{H}} U_-^{-1} &= -B - A\widehat{\Lambda}_- - \widehat{\Lambda}_-^{\mathrm{H}} A = A(\widehat{\Lambda}_+ - \widehat{\Lambda}_-). \end{aligned}$$

So $U_+^{-\mathrm{H}} U_+^{-1} = (U_-^{-\mathrm{H}} U_-^{-1})^{\mathrm{H}} = U_-^{-\mathrm{H}} U_-^{-1}$. Thus,

$$(U_+^{-1} U_-)^{\mathrm{H}} U_+^{-1} U_- = U_-^{\mathrm{H}} U_+^{-\mathrm{H}} U_+^{-1} U_- = I,$$

which infers $\Upsilon := U_+^{-1} U_-$ is unitary.

Item 1 is straightforward. We now prove item 2 for u_i^+ and the case for u_i^- can be handled in exactly the same way. Write $a_i = (u_i^+)^{\mathrm{H}} A u_i^+$, $b_i = (u_i^+)^{\mathrm{H}} B u_i^+$, and $c_i = (u_i^+)^{\mathrm{H}} C u_i^+$. By (2.21a) and (2.21d), we have

$$c_i - (\lambda_i^+)^2 a_i = -\lambda_i^+, \quad b_i + 2a_i \lambda_i^+ = 1$$

solving which for c_i and b_i to get

$$b_i^2 - 4a_i c_i = (1 - 2a_i\lambda_i^+)^2 - 4a_i[-\lambda_i^+ + (\lambda_i^+)^2 a_i] = 1.$$

For item 3, we have, by (2.23),

$$\boldsymbol{Q}(\lambda) = (\lambda I - \widehat{\Lambda}_-^{\mathrm{H}}) A (\lambda I - \widehat{\Lambda}_+), \quad \boldsymbol{Q}(\lambda) = (\lambda I - \widehat{\Lambda}_+^{\mathrm{H}}) A (\lambda I - \widehat{\Lambda}_-)$$

which, together with (2.22), yield (2.17). For item 4, write $\Lambda_{-;\Upsilon} = \Upsilon\Lambda_-\Upsilon^{\mathrm{H}}$, then $\Lambda_+ - \Lambda_{-;\Upsilon} \succ 0$ because for $x \neq 0$,

$$x^{\mathrm{H}}(\Lambda_+ - \Lambda_{-;\Upsilon})x \geq \lambda_1^+ x^{\mathrm{H}}x - \lambda_n^- x^{\mathrm{H}}\Upsilon^{\mathrm{H}}\Upsilon x = (\lambda_1^+ - \lambda_n^-)x^{\mathrm{H}}x > 0$$

which also implies

$$0 \prec (\Lambda_+ - \Lambda_{-;\Upsilon})^{-1} \preceq (\lambda_1^+ - \lambda_n^-)^{-1} I. \tag{2.24}$$

Substitute $U_- = U_+\Upsilon$ into (2.21c) to get $U_+^{\mathrm{H}}CU_+ - \Lambda_+ U_+^{\mathrm{H}}AU_+\Lambda_{-;\Upsilon} = 0$ and thus by (2.21a), we have

$$\begin{aligned}
0 &= U_+^{\mathrm{H}}CU_+ - \Lambda_+U_+^{\mathrm{H}}AU_+\Lambda_+ + \Lambda_+ \\
&= \Lambda_+U_+^{\mathrm{H}}AU_+\Lambda_{-;\Upsilon} - \Lambda_+U_+^{\mathrm{H}}AU_+\Lambda_+ + \Lambda_+ \\
&= \Lambda_+\Big[I - U_+^{\mathrm{H}}AU_+(\Lambda_+ - \Lambda_{-;\Upsilon})\Big].
\end{aligned} \tag{2.25}$$

Substitute $U_+ = U_-\Upsilon^{\mathrm{H}}$ into (2.21c) to get $U_-^{\mathrm{H}}CU_- - \Lambda_{+;\Upsilon}U_-^{\mathrm{H}}AU_-\Lambda_- = 0$, where $\Lambda_{+;\Upsilon} = \Upsilon^{\mathrm{H}}\Lambda_+\Upsilon$. Thus by (2.21b), we have

$$\begin{aligned}
0 &= U_-^{\mathrm{H}}CU_- - \Lambda_-U_-^{\mathrm{H}}AU_-\Lambda_- - \Lambda_- \\
&= \Lambda_{+;\Upsilon}U_-^{\mathrm{H}}AU_-\Lambda_- - \Lambda_-U_-^{\mathrm{H}}AU_-\Lambda_- - \Lambda_- \\
&= -\Big[I - (\Lambda_{+;\Upsilon} - \Lambda_-)U_-^{\mathrm{H}}AU_-\Big]\Lambda_-.
\end{aligned} \tag{2.26}$$

We note that at least one of Λ_+ and Λ_- is nonsingular. If Λ_+ is nonsingular, then (2.25) implies

$$U_+^{\mathrm{H}}AU_+(\Lambda_+ - \Lambda_{-;\Upsilon}) = I \quad \Rightarrow \quad U_+^{\mathrm{H}}AU_+ = (\Lambda_+ - \Lambda_{-;\Upsilon})^{-1}. \tag{2.27}$$

If Λ_- is nonsingular, then (2.26) implies $(\Lambda_{+;\Upsilon} - \Lambda_-)U_-^{\mathrm{H}}AU_- = I$ which, upon using $U_- = U_+\Upsilon$, also implies the second equation in (2.27). Then $U_-^{\mathrm{H}}AU_+ = (\Lambda_+\Upsilon - \Upsilon\Lambda_-)^{-1}$. So, $U_+^{\mathrm{H}}AU_+ = \Theta$, $U_+^{\mathrm{H}}BU_+ = -\Theta\Lambda_+ - \Lambda_{-;\Upsilon}\Theta$, and $U_+^{\mathrm{H}}CU_+ = \Lambda_{-;\Upsilon}\Theta\Lambda_+$. Noticing

$$\Lambda_{-;\Upsilon}\Theta = -(\Lambda_+ - \Lambda_{-;\Upsilon})\Theta + \Lambda_+\Theta = -I + \Lambda_+\Theta,$$

we have (2.18).

For item 5, the equalities in (2.19) is a consequence of $U_- = U_+\Upsilon$ and that Υ is unitary. We now prove (2.19) for U_+. Use $(A^{1/2}U_+)^{\mathrm{H}}(A^{1/2}U_+) = \Theta$ to get

$$\|U_+\|_2 \leq \|A^{-1/2}\|_2\|A^{1/2}U_+\|_2 = \|A^{-1/2}\|_2\sqrt{\|\Theta\|_2} \leq \frac{\|A^{-1/2}\|_2}{\sqrt{\lambda_1^+ - \lambda_n^-}},$$

and use $(U_+^{-1}A^{-1/2})(U_+^{-1}A^{-1/2})^{\mathrm{H}} = \Theta^{-1}$ to get

$$\|U_+^{-1}\|_2 \leq \|U_+^{-1}A^{-1/2}\|_2\|A^{1/2}\|_2 = \sqrt{\|\Theta^{-1}\|_2}\|A^{1/2}\|_2 \leq \|A^{1/2}\|_2\sqrt{\lambda_n^+ - \lambda_1^-}.$$

They give (2.19a) and (2.19b) for U_+. Combine (2.19a) and (2.19b) to get (2.19c). For the first inequality in (2.20), we have

$$\|Z\|_2 \leq \left\|\begin{bmatrix} \|U_+\|_2 & \|U_-\|_2 \\ \|U_+\|_2\xi_+ & \|U_-\|_2\xi_- \end{bmatrix}\right\|_2 = \|U_+\|_2\left\|\begin{bmatrix} 1 & 1 \\ \xi_+ & \xi_- \end{bmatrix}\right\|_2 = \|U_+\|_2\Xi.$$

For the second inequality, we notice by using $U_- = U_+ \varUpsilon$

$$Z = \begin{bmatrix} U_+ & 0 \\ 0 & U_+ \end{bmatrix} \begin{bmatrix} I & \varUpsilon \\ \varLambda_+ & \varUpsilon \varLambda_- \end{bmatrix} = \begin{bmatrix} U_+ & 0 \\ 0 & U_+ \end{bmatrix} \begin{bmatrix} I & 0 \\ \varLambda_+ & I \end{bmatrix} \begin{bmatrix} I & \varUpsilon \\ 0 & S \end{bmatrix},$$

where $S = \varUpsilon \varLambda_- - \varLambda_+ \varUpsilon = -\varTheta^{-1}\varUpsilon$. This expression, after some calculations, leads to

$$\begin{aligned} Z^{-1} &= \begin{bmatrix} I & -\varUpsilon S^{-1} \\ 0 & S^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\varLambda_+ & I \end{bmatrix} \begin{bmatrix} U_+^{-1} & 0 \\ 0 & U_+^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \varUpsilon S^{-1} \varUpsilon \varLambda_- \varUpsilon^{\mathrm{H}} & \varUpsilon S^{-1} \\ -S^{-1}\varLambda_+ & S^{-1} \end{bmatrix} \begin{bmatrix} U_+^{-1} & 0 \\ 0 & U_+^{-1} \end{bmatrix}, \end{aligned}$$

and thus

$$\|Z^{-1}\|_2 \le \|S^{-1}\|_2 \left\| \begin{bmatrix} \xi_- & 1 \\ \xi_+ & 1 \end{bmatrix} \right\|_2 \|U_+^{-1}\|_2 = \|U_+^{-1}\|_2 \|\varTheta\|_2 \varXi$$

which implies the second inequality in (2.20).

We now prove the converse of item 4. First $\varTheta$ is Hermitian and $\varTheta \succ 0$ by (2.24). Obviously A, B, C in (2.18) is Hermitian and $A \succ 0$. Choose $\lambda_0 = (\lambda_1^+ + \lambda_n^-)/2$, then $\varTheta^{-1} \succ \varLambda_+ - \lambda_0 I \succ 0$ and $\varTheta \prec (\varLambda_+ - \lambda_0 I)^{-1}$. Thus,

$$U_+^{\mathrm{H}} \boldsymbol{Q}(\lambda_0) U_+ = (\varLambda_+ - \lambda_0 I)\varTheta(\varLambda_+ - \lambda_0 I) - (\varLambda_+ - \lambda_0 I) \prec 0$$

which says $\boldsymbol{Q}(\lambda_0) \prec 0$. By item 1 of Theorem 2.1, $\boldsymbol{Q}(\lambda)$ is hyperbolic. □

Remark 2.1.

1. Each of the decompositions in (2.17) doesn't reflect the symmetry property in $\boldsymbol{Q}(\lambda)$ somewhat. However, using the fact that $\varUpsilon = U_+^{-1} U_-$ is unitary, we can turn them into

$$\boldsymbol{Q}(\lambda) = U_+^{-\mathrm{H}} (\lambda I - \varUpsilon \varLambda_- \varUpsilon^{\mathrm{H}})(\varLambda_+ - \varUpsilon \varLambda_- \varUpsilon^{\mathrm{H}})^{-1} (\lambda I - \varLambda_+) U_+^{-1}, \tag{2.28a}$$

$$\boldsymbol{Q}(\lambda) = U_-^{-\mathrm{H}} (\lambda I - \varUpsilon^{\mathrm{H}} \varLambda_+ \varUpsilon)(\varUpsilon \varLambda_+ \varUpsilon^{\mathrm{H}} - \varLambda_-)^{-1} (\lambda I - \varLambda_-) U_-^{-1}. \tag{2.28b}$$

These equations are essentially the decomposition in [43, Theorem 31.24] but with more detail.

2. [22, Lemma 6.1] and Problem `gen_hyper2` of [5] provide a different set of formulas for B and C:

$$B = U_+^{-\mathrm{H}} \big[-\varTheta(\varLambda_+^2 - \varUpsilon \varLambda_-^2 \varUpsilon^{\mathrm{H}})\varTheta \big] U_+^{-1}, \tag{2.29a}$$

$$\begin{aligned} C = U_+^{-\mathrm{H}} \big[&-\varTheta(\varLambda_+^3 - \varUpsilon \varLambda_-^3 \varUpsilon^{\mathrm{H}})\varTheta \\ &+ \varTheta(\varLambda_+^2 - \varUpsilon \varLambda_-^2 \varUpsilon^{\mathrm{H}})\varTheta(\varLambda_+^2 - \varUpsilon \varLambda_-^2 \varUpsilon^{\mathrm{H}})\varTheta \big] U_+^{-1}. \end{aligned} \tag{2.29b}$$

[31, Corollary 6] provides yet another formula for C:

$$C = U_+^{-\mathrm{H}} \big[-(\varLambda_+^{-1} - \varUpsilon \varLambda_-^{-1} \varUpsilon^{\mathrm{H}})^{-1} \big] U_+^{-1}. \tag{2.30}$$

Although both (2.29) and (2.30) look more complicated than ours for B and C in (2.18b) and (2.18c), they are actually the same in theory. In fact, we have

$$\varTheta(\varLambda_+^2 - \varUpsilon \varLambda_-^2 \varUpsilon^{\mathrm{H}})\varTheta = \varTheta(\varLambda_+^2 - [\varLambda_+ - \varTheta^{-1}]^2)\varTheta$$

$$= \varLambda_+\varTheta + \varTheta\varLambda_+ - I \tag{2.31}$$

which says (2.29a) is the same as (2.18b).

$$\begin{aligned}
\varLambda_+^{-1} - \varUpsilon\varLambda_-^{-1}\varUpsilon^{\mathrm{H}} &= \varLambda_+^{-1} - [\varLambda_+ - \varTheta^{-1}]^{-1} \qquad \text{(use (2.18e))} \\
&= \varLambda_+^{-1}(-\varTheta^{-1})[\varLambda_+ - \varTheta^{-1}]^{-1} \quad (\text{use } X^{-1} - Y^{-1} = X^{-1}[Y - X]Y^{-1}) \\
&= -(\varLambda_+\varTheta\varLambda_+ - \varLambda_+)^{-1}.
\end{aligned}$$

So (2.30) is the same as (2.18c). Finally

$$\begin{aligned}
\varTheta(\varLambda_+^3 - \varUpsilon\varLambda_-^3\varUpsilon^{\mathrm{H}})\varTheta &= \varTheta(\varLambda_+^3 - [\varLambda_+ - \varTheta^{-1}]^3)\varTheta \\
&= \varTheta^{-1} + \varTheta\varLambda_+^2 + \varLambda_+^2\varTheta + \varTheta\varLambda_+\varTheta^{-1}\varLambda_+\varTheta - \varTheta\varLambda_+\varTheta^{-1} \\
&\quad - \varTheta^{-1}\varLambda_+\varTheta - \varLambda_+.
\end{aligned}$$

Therefore use also (2.31) to get

$$\begin{aligned}
&-\varTheta(\varLambda_+^3 - \varUpsilon\varLambda_-^3\varUpsilon^{\mathrm{H}})\varTheta + \varTheta(\varLambda_+^2 - \varUpsilon\varLambda_-^2\varUpsilon^{\mathrm{H}})\varTheta(\varLambda_+^2 - \varUpsilon\varLambda_-^2\varUpsilon^{\mathrm{H}})\varTheta \\
&\quad = -(\varTheta^{-1} + \varTheta\varLambda_+^2 + \varLambda_+^2\varTheta + \varTheta\varLambda_+\varTheta^{-1}\varLambda_+\varTheta - \varTheta\varLambda_+\varTheta^{-1} - \varTheta^{-1}\varLambda_+\varTheta - \varLambda_+) \\
&\qquad + (\varTheta\varLambda_+ + \varLambda_+\varTheta - I)\varTheta^{-1}(\varTheta\varLambda_+ + \varLambda_+\varTheta - I) \\
&\quad = -(\varTheta^{-1} + \varTheta\varLambda_+^2 + \varLambda_+^2\varTheta + \varTheta\varLambda_+\varTheta^{-1}\varLambda_+\varTheta - \varTheta\varLambda_+\varTheta^{-1} - \varTheta^{-1}\varLambda_+\varTheta - \varLambda_+) \\
&\qquad + \varTheta^{-1} - \varTheta\varLambda_+\varTheta^{-1} - \varLambda_+ - \varLambda_+ + \varTheta\varLambda_+^2 + \varLambda_+\varTheta\varLambda_+ \\
&\qquad - \varTheta^{-1}\varLambda_+\varTheta + \varTheta\varLambda_+\varTheta^{-1}\varLambda_+\varTheta + \varLambda_+^2\varTheta \\
&\quad = -\varLambda_+ + \varLambda_+\varTheta\varLambda_+
\end{aligned}$$

which proves that (2.29b) is the same as (2.18c).

3. $\widehat{\varLambda}_\pm$ in (2.22) are two solutions of the matrix equation

$$AX^2 + BX + C = 0. \tag{2.32}$$

In fact,

$$A(U_+\varLambda_+U_+^{-1})^2 + B(U_+\varLambda_+U_+^{-1}) + C = (AU_+\varLambda_+^2 + BU_+\varLambda_+ + CU_+)U_+^{-1} = 0,$$

and similarly for $A(U_-\varLambda_-U_-^{-1})^2 + B(U_-\varLambda_-U_-^{-1}) + C = 0$. On the other hand, the ability of solving (2.32) factorizes $\boldsymbol{Q}(\lambda)$ into the product of two linear matrix polynomials, based on which Guo and Lancaster [20] proposed their solvent approach for solving HQEP (1.1) of modest sizes.

3 Variational principles

Throughout this section, $\boldsymbol{Q}(\lambda) = \lambda^2 A + \lambda B + C \in \mathbb{C}^{n\times n}$ will be always assumed a hyperbolic quadratic matrix polynomial and the notations in Theorem 2.5 will be kept. The scalar λ_0 is as in item 1 of Theorem 2.1 such that $\boldsymbol{Q}(\lambda_0) \prec 0$.

Consider the following equation in λ

$$f(\lambda, x) := x^{\mathrm{H}}\boldsymbol{Q}(\lambda)x = \lambda^2(x^{\mathrm{H}}Ax) + \lambda(x^{\mathrm{H}}Bx) + (x^{\mathrm{H}}Cx) = 0, \tag{3.1}$$

given $x \neq 0$. Since $\boldsymbol{Q}(\lambda)$ is hyperbolic, this equation always has two distinct real roots (as functions of x)

$$\rho_{\pm}(x) = \frac{-x^{\mathrm{H}}Bx \pm \left[(x^{\mathrm{H}}Bx)^2 - 4(x^{\mathrm{H}}Ax)(x^{\mathrm{H}}Cx)\right]^{1/2}}{2(x^{\mathrm{H}}Ax)}. \tag{3.2}$$

We shall call $\rho_+(x)$ the *pos-type Rayleigh quotient* of $\boldsymbol{Q}(\lambda)$ at x, and $\rho_-(x)$ the *neg-type Rayleigh quotient* of $\boldsymbol{Q}(\lambda)$ at x. It is easy to verify that for any $x \neq 0$, $\rho_{\pm}(x) \in \mathbb{R}$, and $\rho_{\pm}(\alpha x) = \rho_{\pm}(x)$ for any $\alpha \in \mathbb{C}$. By the elementary knowledge of scalar quadratic polynomials, we have

$$\rho_+(x) + \rho_-(x) = -\frac{x^{\mathrm{H}}Bx}{x^{\mathrm{H}}Ax}, \quad \rho_+(x)\cdot\rho_-(x) = \frac{x^{\mathrm{H}}Cx}{x^{\mathrm{H}}Ax}. \tag{3.3}$$

Both will be used later in this paper.

Theorem 3.1. *We have*

$$\rho_+(x) \in [\lambda_1^+, \lambda_n^+], \quad \rho_-(x) \in [\lambda_1^-, \lambda_n^-], \tag{3.4}$$

$$\varsigma(x) := \left[(x^{\mathrm{H}}Bx)^2 - 4(x^{\mathrm{H}}Ax)(x^{\mathrm{H}}Cx)\right]^{1/2} = \pm[2\rho_{\pm}(x)\, x^{\mathrm{H}}Ax + x^{\mathrm{H}}Bx], \tag{3.5}$$

$$\varsigma_0(x) = \frac{\varsigma(x)}{x^{\mathrm{H}}x} \in [(\lambda_1^+ - \lambda_n^-)\lambda_{\min}(A),\, (\lambda_n^+ - \lambda_1^-)\lambda_{\max}(A)]. \tag{3.6}$$

Consequently, $\lambda_i^+ = \rho_+(u_i^+)$ for the quadratic eigenpair (λ_i^+, u_i^+) and $\rho_-(u_j^-) = \lambda_j^-$ for the quadratic eigenpair (λ_j^-, u_j^-).

Proof. By item 3 of Theorem 2.1, for any fixed nonzero x, $f(\lambda, x) < 0$ for $\lambda \in (\lambda_n^-, \lambda_1^+)$ and $f(\lambda, x) > 0$ for $\lambda \in (-\infty, \lambda_1^-) \cup (\lambda_n^+, +\infty)$. Thus, the larger root of the scalar quadratic equation $f(\lambda, x) = 0$ in λ must lie in $[\lambda_1^+, \lambda_n^+]$ and the smaller one in $[\lambda_1^-, \lambda_n^-]$. That is (3.4). For (3.5), we have

$$\begin{aligned} 2\rho_{\pm}(x)\, x^{\mathrm{H}}Ax + x^{\mathrm{H}}Bx &= \left[-x^{\mathrm{H}}Bx \pm \sqrt{(x^{\mathrm{H}}Bx)^2 - 4(x^{\mathrm{H}}Ax)(x^{\mathrm{H}}Cx)}\ \right] + x^{\mathrm{H}}Bx \\ &= \pm\varsigma(x). \end{aligned}$$

Lastly, the inclusion (3.6) is a result of $\varsigma(x) = [\rho_+(x) - \rho_-(x)]\, x^{\mathrm{H}}Ax$. $\square$

3.1 Courant-Fischer type min-max principles

Theorem 3.2 below is a restatement of [43, Theorem 32.10, Theorem 32.11 and Remark 32.13]. However, it is essentially due to Duffin [12, 1955] whose proof, although for overdamped $\boldsymbol{Q}$, works for the general hyperbolic case. Closely related ones for more general nonlinear eigenvalue problems (other than quadratic eigenvalue problems) can be found in [49, 50, 66, 67]. They can be considered as a generalization of the Courant-Fischer min-max principles (see [47, p.206], [56, p.201]).

Theorem 3.2 ([12])**.** *We have for* $1 \le i \le n$

$$\lambda_i^+ = \max_{\substack{\mathcal{X}\subseteq\mathbb{C}^n \\ \operatorname{codim}\mathcal{X}=i-1}} \min_{\substack{x\in\mathcal{X} \\ x\neq 0}} \rho_+(x), \tag{3.7a}$$

$$\lambda_i^+ = \min_{\substack{\mathcal{X}\subseteq\mathbb{C}^n \\ \dim\mathcal{X}=i}} \max_{\substack{x\in\mathcal{X} \\ x\neq 0}} \rho_+(x), \tag{3.7b}$$

$$\lambda_i^- = \max_{\substack{\mathcal{X}\subseteq\mathbb{C}^n \\ \operatorname{codim}\mathcal{X}=i-1}} \min_{\substack{x\in\mathcal{X} \\ x\neq 0}} \rho_-(x), \tag{3.7c}$$

$$\lambda_i^- = \min_{\substack{\mathcal{X}\subseteq\mathbb{C}^n \\ \dim\mathcal{X}=i}} \max_{\substack{x\in\mathcal{X} \\ x\neq 0}} \rho_-(x). \tag{3.7d}$$

In particular,

$$\lambda_1^+ = \min_{x\neq 0} \rho_+(x), \quad \lambda_n^+ = \max_{x\neq 0} \rho_+(x), \tag{3.8a}$$

$$\lambda_1^- = \min_{x\neq 0} \rho_-(x), \quad \lambda_n^- = \max_{x\neq 0} \rho_-(x). \tag{3.8b}$$

3.2 Wielandt-Lidskii type min-max principles

Theorems 3.3 and 3.4 which can be considered as generalizations of Amir-Moéz type min-max principles [1] and Theorem 3.5 which can be considered as generalizations of the Wielandt-Lidskii min-max principles ([39, 69] and also [6, p.67], [56, p.199]) and Ky-Fan trace min/max principles [15] are new. For the ease of stating them, let $\lambda_\pm \in \mathbb{R}$ such that

$$\lambda_- \le \lambda_1^- \le \lambda_n^- \le \lambda_0 \le \lambda_1^+ \le \lambda_n^+ \le \lambda_+.$$

Such $\lambda_\pm$ exist, e.g., $\lambda_- = \lambda_1^-$ or $-\infty$ and $\lambda_+ = \lambda_n^+$ or ∞. Set intervals

$$\mathscr{I}_+ = \begin{cases} [\lambda_0, \lambda_+], & \text{if } \lambda_+ < \infty, \\ [\lambda_0, \infty), & \text{otherwise,} \end{cases} \qquad \mathscr{I}_- = \begin{cases} [\lambda_-, \lambda_0], & \text{if } \lambda_- > -\infty, \\ (-\infty, \lambda_0], & \text{otherwise.} \end{cases} \tag{3.9}$$

The following lemma is also essentially due to Duffin [12] whose proof, although for overdamped Q, again works for the general hyperbolic case.

Lemma 3.1. *We have*

$$\lambda_i^+ \ge \rho_+(x) \textit{ for any } x \in \operatorname{span}\{u_1^+, u_2^+, \dots, u_i^+\}, \tag{3.10a}$$

$$\lambda_i^+ \le \rho_+(x) \textit{ for any } x \in \operatorname{span}\{u_i^+, u_{i+1}^+, \dots, u_n^+\}. \tag{3.10b}$$

To generalize Amir-Moéz/Wielandt-Lidskii min-max principles, we introduce the following notations. For $X \in \mathbb{C}^{n\times k}$ with $\operatorname{rank}(X) = k$, $X^{\mathrm{H}}\boldsymbol{Q}(\lambda)X$ is a $k \times k$ hyperbolic quadratic matrix polynomial. Hence its quadratic eigenvalues are real. Denote them by $\lambda_{i,X}^{\pm}$ arranged as

$$\lambda_{1,X}^{-} \le \cdots \le \lambda_{k,X}^{-} \le \lambda_{1,X}^{+} \le \cdots \le \lambda_{k,X}^{+}. \tag{3.11}$$

Theorem 3.3. *Let $1 \le i_1 < \cdots < i_k \le n$. For any*

$$\varPhi \,:\, \underbrace{\mathscr{I}_{+} \times \cdots \times \mathscr{I}_{+}}_{k} \to \mathbb{R}$$

that is non-decreasing in each of its arguments, we have[2]

$$\min_{\substack{\mathcal{X}_1\subset\cdots\subset\mathcal{X}_k\\ \dim\mathcal{X}_j=i_j}} \;\sup_{\substack{x_j\in\mathcal{X}_j,\,j=1,\dots,k\\ X=[x_1,\dots,x_k]\\ \operatorname{rank}(X)=k}} \varPhi(\lambda_{1,X}^{+},\cdots,\lambda_{k,X}^{+}) = \varPhi(\lambda_{i_1}^{+},\cdots,\lambda_{i_k}^{+}), \tag{3.12a}$$

$$\max_{\substack{\mathcal{X}_1\supset\cdots\supset\mathcal{X}_k\\ \operatorname{codim}\mathcal{X}_j=i_j-1}} \;\inf_{\substack{x_j\in\mathcal{X}_j,\,j=1,\dots,k\\ X=[x_1,\dots,x_k]\\ \operatorname{rank}(X)=k}} \varPhi(\lambda_{1,X}^{+},\cdots,\lambda_{k,X}^{+}) = \varPhi(\lambda_{i_1}^{+},\cdots,\lambda_{i_k}^{+}). \tag{3.12b}$$

*If also $\varPhi$ is continuous, then "*sup*" in* (3.12a) *and "*inf*" in* (3.12b) *can be replaced by "*max*" and "*min*", respectively. In particular, setting $i_j = j$ in* (3.12a) *and setting $i_j = j + n - k$ in* (3.12b)*, respectively, give*

$$\min_{\operatorname{rank}(X)=k} \varPhi(\lambda_{1,X}^{+},\cdots,\lambda_{k,X}^{+}) = \varPhi(\lambda_{1}^{+},\cdots,\lambda_{k}^{+}), \tag{3.13a}$$

$$\max_{\operatorname{rank}(X)=k} \varPhi(\lambda_{1,X}^{+},\cdots,\lambda_{k,X}^{+}) = \varPhi(\lambda_{n-k+1}^{+},\cdots,\lambda_{n}^{+}). \tag{3.13b}$$

Proof. For convenience, we define, for a matrix $W = [w_1, \dots, w_p]$,

$$\mathcal{S}_{j,W} := \operatorname{span}\{w_1,\cdots,w_j\},\; \mathcal{T}_{j,W} := \operatorname{span}\{w_j,\cdots,w_p\} \quad \text{for } j = 1,\cdots,p.$$

In particular $\mathcal{S}_W = \mathcal{S}_{p,W}$, $\mathcal{T}_W = \mathcal{T}_{1,W}$, and thus $\mathcal{S}_W = \mathcal{T}_W$.

First we prove (3.12b). Recall the quadratic eigenvectors u_j^{+} introduced in Theorem 2.5. Choose

$$\widehat{\mathcal{X}}_j = \operatorname{span}\{u_{i_j}^{+},\cdots,u_n^{+}\} \quad \text{for } j = 1,2,\dots,k. \tag{3.14}$$

Then $\widehat{\mathcal{X}}_1 \supset \cdots \supset \widehat{\mathcal{X}}_k$ and $\operatorname{codim}\widehat{\mathcal{X}}_j = i_j - 1$. By Lemma 3.1, $\rho_+(x) \ge \lambda_{i_j}^{+}$ for any $x \in \widehat{\mathcal{X}}_j$. Therefore

$$\min_{\substack{x\in\widehat{\mathcal{X}}_j\\ x\ne0}} \rho_+(x) = \lambda_{i_j}^{+}.$$

For any $X = [x_1, \dots, x_k]$ with $x_j \in \widehat{\mathcal{X}}_j$ for $j = 1,\cdots,k$ such that $\operatorname{rank}(X) = k$, consider $X^{\mathrm{H}}\boldsymbol{Q}(\lambda)X$ which is a $k \times k$ hyperbolic quadratic matrix polynomial. For $j = 1,\cdots,k$, noticing $\mathcal{T}_{j,X} \subset \widehat{\mathcal{X}}_j$, we have by Theorem 3.2

$$\lambda_{j,X}^{+} = \max_{\substack{\mathcal{X}\subset\mathcal{T}_X\\ \dim\mathcal{X}=k-j+1}} \min_{\substack{x\in\mathcal{X}\\ x\ne0}} \rho_+(x) \ge \min_{\substack{x\in\mathcal{T}_{j,X}\\ x\ne0}} \rho_+(x) \ge \min_{\substack{x\in\widehat{\mathcal{X}}_j\\ x\ne0}} \rho_+(x) = \lambda_{i_j}^{+}.$$

[2]In (3.12a), it is not clear if the "sup" is attainable for any given $\mathcal{X}_j$ satisfying the given assumptions, except for continuous $\varPhi$. The same comment applies to the "inf" in (3.12b).

Since $\Phi(\cdot)$ is non-decreasing in each of its arguments,

$$\Phi(\lambda^+_{1,X}, \cdots, \lambda^+_{k,X}) \geq \Phi(\lambda^+_{i_1}, \cdots, \lambda^+_{i_k})$$

which gives

$$\min_{\substack{x_j \in \widehat{\mathcal{X}}_j,\, j=1,\dots,k \\ X=[x_1,\dots,x_k] \\ \operatorname{rank}(X)=k}} \Phi(\lambda^+_{1,X}, \cdots, \lambda^+_{k,X}) \geq \Phi(\lambda^+_{i_1}, \cdots, \lambda^+_{i_k})$$

because $x_j \in \widehat{\mathcal{X}}_j$ for $1 \leq i \leq k$ are arbitrary, subject to $\operatorname{rank}(X) = k$. Therefore

$$\sup_{\substack{\mathcal{X}_1 \supset \cdots \supset \mathcal{X}_k \\ \operatorname{codim} \mathcal{X}_j = i_j - 1}} \inf_{\substack{x_j \in \mathcal{X}_j,\, j=1,\dots,k \\ X=[x_1,\dots,x_k] \\ \operatorname{rank}(X)=k}} \Phi(\lambda^+_{1,X}, \cdots, \lambda^+_{k,X}) \geq \Phi(\lambda^+_{i_1}, \cdots, \lambda^+_{i_k}). \tag{3.15}$$

On the other hand, let $\mathcal{X}_j$ for $j = 1, \cdots, k$ be any subspaces that satisfy the assumptions: $\mathcal{X}_1 \supset \cdots \supset \mathcal{X}_k$ and $\operatorname{codim} \mathcal{X}_j = i_j - 1$. Define $\mathcal{Y}_j = \operatorname{span}\{u^+_1, \cdots, u^+_{i_j}\}$. Then $\mathcal{Y}_1 \subset \cdots \subset \mathcal{Y}_k$ and $\dim \mathcal{Y}_j = i_j$. By [1, Corollary 2.2] (see also [37, Lemma 3.2]), there exists two A-orthonormal sets $\{x_1, \dots, x_k\}$ and $\{y_1, \cdots, y_k\}$ with $x_j \in \mathcal{X}_j$ for $j = 1, \dots, k$ and $y_j \in \mathcal{Y}_j$ for $1 \leq j \leq k$ such that

$$\mathcal{T}_X := \operatorname{span}\{x_1, \cdots, x_k\} = \operatorname{span}\{y_1, \cdots, y_k\} =: \mathcal{S}_Y.$$

where $X = [x_1, \dots, x_k]$ and $Y = [y_1, \cdots, y_k]$ satisfy $X^{\mathrm{H}}AX = Y^{\mathrm{H}}AY = I_k$. $Y^{\mathrm{H}}\boldsymbol{Q}(\lambda)Y$ is a hyperbolic quadratic matrix polynomial whose pos-type quadratic eigenvalues are $\lambda^+_{1,Y} \leq \cdots \leq \lambda^+_{k,Y}$. Since $\mathcal{S}_Y = \mathcal{T}_X$, $\lambda^+_{j,Y} = \lambda^+_{j,X}$ for $j = 1, \cdots, k$. By Lemma 3.1, $\rho_+(y) \leq \lambda^+_{i_j}$ for any $y \in \mathcal{Y}_j$. Therefore

$$\max_{\substack{y \in \mathcal{Y}_j \\ y \neq 0}} \rho_+(y) = \lambda^+_{i_j}.$$

By Theorem 3.2 and noticing $\mathcal{S}_{j,Y} \subset \mathcal{Y}_j$, we have, for $j = 1, \cdots, k$,

$$\lambda^+_{j,X} = \lambda^+_{j,Y} = \min_{\substack{\mathcal{Y} \subset \mathcal{S}_Y \\ \dim \mathcal{Y} = j}} \max_{\substack{y \in \mathcal{Y} \\ y \neq 0}} \rho_+(y) \leq \max_{\substack{y \in \mathcal{S}_{j,Y} \\ y \neq 0}} \rho_+(y) \leq \max_{\substack{y \in \mathcal{Y}_j \\ y \neq 0}} \rho_+(y) = \lambda^+_{i_j}.$$

Since $\Phi(\cdot)$ is non-decreasing in each of its arguments,

$$\Phi(\lambda^+_{1,X}, \cdots, \lambda^+_{k,X}) \leq \Phi(\lambda^+_{i_1}, \cdots, \lambda^+_{i_k}),$$

which gives

$$\inf_{\substack{x_j \in \mathcal{X}_j,\, j=1,\dots,k \\ X=[x_1,\dots,x_k] \\ \operatorname{rank}(X)=k}} \Phi(\lambda^+_{1,X}, \cdots, \lambda^+_{k,X}) \leq \Phi(\lambda^+_{i_1}, \cdots, \lambda^+_{i_k}).$$

Since $\mathcal{X}_j$ are arbitrary, we conclude

$$\sup_{\substack{\mathcal{X}_1 \supset \cdots \supset \mathcal{X}_k \\ \operatorname{codim} \mathcal{X}_j = i_j - 1}} \inf_{\substack{x_j \in \mathcal{X}_j,\, j=1,\dots,k \\ X=[x_1,\dots,x_k] \\ \operatorname{rank}(X)=k}} \Phi(\lambda^+_{1,X}, \cdots, \lambda^+_{k,X}) \leq \Phi(\lambda^+_{i_1}, \cdots, \lambda^+_{i_k}). \tag{3.16}$$

Combine (3.15) and (3.16) to get

$$\sup_{\substack{\mathcal{X}_1\supset\cdots\supset\mathcal{X}_k\\ \operatorname{codim}\mathcal{X}_j=i_j-1}}\ \inf_{\substack{x_j\in\mathcal{X}_j,\,j=1,\ldots,k\\ X=[x_1,\ldots,x_k]\\ \operatorname{rank}(X)=k}} \Phi(\lambda^+_{1,X},\cdots,\lambda^+_{k,X}) = \Phi(\lambda^+_{i_1},\cdots,\lambda^+_{i_k}). \tag{3.12b$'$}$$

But the "sup" here is achievable by the selection in (3.14). Thus we have (3.12b).

Now we claim the "inf" can be replaced by "min" for a continuous Φ. Let $\mathcal{X}_j$ for $j = 1,\cdots,k$ be given and satisfy the assumptions: $\mathcal{X}_1 \supset \cdots \supset \mathcal{X}_k$ and $\operatorname{codim}\mathcal{X}_j = i_j - 1$. There exist a sequence $X^{(i)} \in \mathbb{C}^{n\times k}$ with $\operatorname{rank}(X^{(i)}) = k$ and its jth column in $\mathcal{X}_j$ such that

$$\lim_{i\to\infty} \Phi(\lambda^+_{1,X^{(i)}},\cdots,\lambda^+_{k,X^{(i)}}) = \inf_{\substack{x_j\in\mathcal{X}_j,\,j=1,\ldots,k\\ X=[x_1,\ldots,x_k]\\ \operatorname{rank}(X)=k}} \Phi(\lambda^+_{1,X},\cdots,\lambda^+_{k,X}). \tag{3.17}$$

Without loss of generality, we may assume $X^{(i)}$ has A-orthonormal columns, i.e.,

$$(X^{(i)})^{\mathrm{H}} A X^{(i)} = I_k;$$

otherwise we can perform the Gram-Schimdt A-orthogonalization on the columns of $X^{(i)}$ from the last column backwards, and the new $X^{(i)}$ has the same property as the old $X^{(i)}$: $\operatorname{rank}(X^{(i)}) = k$ and its jth column in $\mathcal{X}_j$, and also $\lambda^\pm_{j,X^{(i)}}$ remain the same. Since $\{X^{(i)}\}$ is a bounded set in $\mathbb{C}^{n\times k}$, it has a convergent subsequence. Through renaming, we may assume that $\{X^{(i)}\}$ itself is convergent, and let $Y \in \mathbb{C}^{n\times k}$ be the limit. It is not hard to see that $Y^{\mathrm{H}}AY = I_k$ which implies $\operatorname{rank}(Y) = k$ and that the jth column of Y is in $\mathcal{X}_j$. Since $(X^{(i)})^{\mathrm{H}}\boldsymbol{Q}(\lambda)X^{(i)}$ goes to $Y^{\mathrm{H}}\boldsymbol{Q}(\lambda)Y$, by the continuity of quadratic eigenvalues with respect to the coefficient matrices we conclude

$$\lim_{i\to\infty} \lambda^\pm_{j,X^{(i)}} = \lambda^\pm_{j,Y} \quad \text{for } 1 \le j \le k.$$

Therefore the left-hand side of (3.17) is equal to $\Phi(\lambda^+_{1,Y},\cdots,\lambda^+_{k,Y})$, and thus the "inf" in (3.17) is attainable.

For (3.12a), a proof similar to what we did above for (3.12b) works: choosing $\widehat{\mathcal{X}}_j = \operatorname{span}\{u^+_1,\cdots,u^+_{i_j}\}$ will lead to that the left-hand side is no bigger than its right-hand side, and choosing $\mathcal{Y}_j = \operatorname{span}\{u^+_{i_j},\cdots,u^+_n\}$ will give the opposite. $\square$

Similarly to Theorem 3.3, we have

Theorem 3.4. *Let $1 \le i_1 < \cdots < i_k \le n$. For any*

$$\Psi\ :\ \underbrace{\mathscr{I}_-\times\cdots\times\mathscr{I}_-}_{k} \to \mathbb{R}$$

that is non-decreasing in each of its arguments, we have[3]

$$\min_{\substack{\mathcal{X}_1\subset\cdots\subset\mathcal{X}_k\\ \dim\mathcal{X}_j=i_j}}\ \sup_{\substack{x_j\in\mathcal{X}_j,\,j=1,\ldots,k\\ X=[x_1,\ldots,x_k]\\ \operatorname{rank}(X)=k}} \Psi(\lambda^-_{1,X},\cdots,\lambda^-_{k,X}) = \Psi(\lambda^-_{i_1},\cdots,\lambda^-_{i_k}), \tag{3.18a}$$

[3]In (3.18a), it is not clear if the "sup" is attainable for any given $\mathcal{X}_j$ satisfying the given assumptions. The same comment applies to the "inf" in (3.18b).

$$\max_{\substack{\mathfrak{X}_1\supset\cdots\supset\mathfrak{X}_k\\ \operatorname{codim}\mathfrak{X}_j=i_j-1}} \inf_{\substack{x_j\in\mathfrak{X}_j,\, j=1,\dots,k\\ X=[x_1,\dots,x_k]\\ \operatorname{rank}(X)=k}} \Psi(\lambda^-_{1,X},\cdots,\lambda^-_{k,X}) = \Psi(\lambda^-_{i_1},\cdots,\lambda^-_{i_k}). \tag{3.18b}$$

*If also Ψ is continuous, then "*sup*" in* (3.18a) *and "*inf*" in* (3.18b) *can be replaced by "*max*" and "*min*", respectively. In particular, setting $i_j = j$ in* (3.18a) *and setting $i_j = j + n - k$ in* (3.18b), *respectively, give*

$$\min_{\operatorname{rank}(X)=k} \Psi(\lambda^-_{1,X},\cdots,\lambda^-_{k,X}) = \Psi(\lambda^-_1,\cdots,\lambda^-_k), \tag{3.19a}$$

$$\max_{\operatorname{rank}(X)=k} \Psi(\lambda^-_{1,X},\cdots,\lambda^-_{k,X}) = \Psi(\lambda^-_{n-k+1},\cdots,\lambda^-_n). \tag{3.19b}$$

Proof. Consider the hyperbolic quadratic matrix polynomial $\widehat{\boldsymbol{Q}}(\lambda) = \lambda^2 A + \lambda(-B) + C$ whose quadratic eigenvalues are

$$\hat\lambda^-_1 \le \cdots \le \hat\lambda^-_n < \hat\lambda^+_1 \le \cdots \le \hat\lambda^+_n,$$

where $\hat\lambda^-_i = -\lambda^+_{n-i+1}$ and $\hat\lambda^+_j = -\lambda^-_{n-j+1}$. Apply (3.12b) to $\widehat{\boldsymbol{Q}}(\lambda)$ with

$$\Phi(\xi_1,\dots,\xi_k) := -\Psi(-\xi_k,\dots,-\xi_1)$$

to get (3.18a), and apply (3.12a) to $\widehat{\boldsymbol{Q}}(\lambda)$ with the same Φ to get (3.18b). $\square$

Specializing Theorems 3.3 and 3.4 to the case where Φ and Ψ are the sum of its arguments gives us Wielandt-Lidskii type min-max principles as summarized in the following theorem and Ky-Fan type trace min/max principles.

Theorem 3.5. *Let $1 \le i_1 < \cdots < i_k \le n$ and* $\mathrm{typ} \in \{+,-\}$. *Then*

$$\min_{\substack{\mathfrak{X}_1\subset\cdots\subset\mathfrak{X}_k\\ \dim\mathfrak{X}_j=i_j}} \max_{\substack{x_j\in\mathfrak{X}_j\\ X=[x_1,\dots,x_k]\\ \operatorname{rank}(X)=k}} \sum_{j=1}^k \lambda^{\mathrm{typ}}_{j,X} = \sum_{j=1}^k \lambda^{\mathrm{typ}}_{i_j}, \tag{3.20a}$$

$$\max_{\substack{\mathfrak{X}_1\supset\cdots\supset\mathfrak{X}_k\\ \operatorname{codim}\mathfrak{X}_j=i_j-1}} \min_{\substack{x_j\in\mathfrak{X}_j\\ X=[x_1,\dots,x_k]\\ \operatorname{rank}(X)=k}} \sum_{j=1}^k \lambda^{\mathrm{typ}}_{j,X} = \sum_{j=1}^k \lambda^{\mathrm{typ}}_{i_j}. \tag{3.20b}$$

In particular, setting $i_j = j$ in (3.20a) *and setting $i_j = j + n - k$ in* (3.20b) *give*

$$\min_{\operatorname{rank}(X)=k} \sum_{j=1}^k \lambda^{\mathrm{typ}}_{j,X} = \sum_{j=1}^k \lambda^{\mathrm{typ}}_{j}, \quad \max_{\operatorname{rank}(X)=k} \sum_{j=1}^k \lambda^{\mathrm{typ}}_{j,X} = \sum_{j=1}^k \lambda^{\mathrm{typ}}_{n-k+j}. \tag{3.21}$$

3.3 Cauchy type interlacing inequalities

The Cauchy type interlacing inequalities in (3.22) were recently obtained by Veselić [64]. Here we present a simple proof, using our generalizations of Amir-Moéz type min-max principles in Theorems 3.3 and 3.4.

Theorem 3.6 (Cauchy-type interlacing inequalities [64]). *Suppose* $X \in \mathbb{C}^{n\times k}$ *with* $\operatorname{rank}(X) = k$*. Denote the quadratic eigenvalues of* $X^{\mathrm{H}}\boldsymbol{Q}(\lambda)X$ *by*

$$\mu_1^- \le \cdots \le \mu_k^- < \mu_1^+ \le \cdots \le \mu_k^+.$$

Then

$$\lambda_i^+ \le \mu_i^+ \le \lambda_{i+n-k}^+, \quad i = 1, \cdots, k, \tag{3.22a}$$

$$\lambda_j^- \le \mu_j^- \le \lambda_{j+n-k}^-, \quad j = 1, \cdots, k. \tag{3.22b}$$

Proof. Let

$$\Phi(\alpha_1, \cdots, \alpha_k) = \text{the } i\text{th largest } \alpha_j.$$

Then this Φ satisfies the condition of Theorem 3.3. Making use of (3.13a) and (3.13b) gives $\mu_i^+ \ge \lambda_i^+$ and $\mu_i^+ \le \lambda_{i+n-k}^+$, respectively. That is (3.22a). Similarly, we get (3.22b) by Theorem 3.4. □

Remark 3.1. The Cauchy type interlacing inequalities in Theorem 3.6 are sharper than those possibly derivable by linearization. Actually, through linearization and by item 1 of [38, Theorem 1.1] (which is, in fact, [30, Theorem 2.1]), we can only obtain

$$\lambda_i^+ \le \mu_i^+ \le \lambda_{i+2n-2k}^+, \quad i = 1, \cdots, k,$$

$$\lambda_{j-(n-k)}^- \le \mu_j^- \le \lambda_{j+n-k}^-, \quad j = 1, \cdots, k,$$

where we set $\lambda_i^+ = +\infty$ for $i > n$ and $\lambda_j^- = -\infty$ for $j < 1$.

4 Perturbation analysis

4.1 Setting the stage

Throughout this section, we suppose that Hermitian matrices A, B, and C are perturbed to Hermitian matrices $\widetilde{A}$, $\widetilde{B}$, and $\widetilde{C}$ and set

$$\varDelta A = \widetilde{A} - A, \quad \varDelta B = \widetilde{B} - B, \quad \varDelta C = \widetilde{C} - C. \tag{4.1}$$

This notational convention of placing a "$\widetilde{\ }$" over a symbol for the corresponding perturbed quantity and a "$\varDelta$" before a symbol for the change in the quantity will be generalized to all quantities that depend on A, B, and C. For example, $\boldsymbol{Q}(\lambda) = \lambda^2 A + \lambda B + C$ is perturbed to $\widetilde{\boldsymbol{Q}}(\lambda) = \lambda^2 \widetilde{A} + \lambda \widetilde{B} + \widetilde{C}$, as a result, and

$$\varDelta\rho_{\pm}(x) = \frac{-(x^{\rm H}\widetilde{B}x) \pm \left[(x^{\rm H}\widetilde{B}x)^2 - 4(x^{\rm H}\widetilde{A}x)(x^{\rm H}\widetilde{C}x)\right]^{1/2}}{2(x^{\rm H}\widetilde{A}x)} - \frac{-(x^{\rm H}Bx) \pm \left[(x^{\rm H}Bx)^2 - 4(x^{\rm H}Ax)(x^{\rm H}Cx)\right]^{1/2}}{2(x^{\rm H}Ax)}.$$

Besides $A \succ 0$, the other key condition for $\boldsymbol{Q}(\lambda) = \lambda^2 A + \lambda B + C$ to be hyperbolic is

$$[\varsigma(x)]^2 = (x^{\rm H}Bx)^2 - 4(x^{\rm H}Ax)(x^{\rm H}Cx) > 0, \quad \text{for all } 0 \neq x \in \mathbb{C}^n. \tag{2.2}$$

We first establish a condition under which (2.2) is weakly[4] satisfied for all convex combination $(1-t)\,\boldsymbol{Q}(\lambda) + t\,\widetilde{\boldsymbol{Q}}(\lambda)$. To this end, we define

$$\phi(x) := (x^{\rm H}\varDelta Bx)^2 - 4(x^{\rm H}\varDelta Ax)(x^{\rm H}\varDelta Cx), \tag{4.2}$$

$$\psi(x) := (x^{\rm H}Bx)(x^{\rm H}\varDelta Bx) - 2(x^{\rm H}Ax)(x^{\rm H}\varDelta Cx) - 2(x^{\rm H}Cx)(x^{\rm H}\varDelta Ax), \tag{4.3}$$

and define $\tilde{\phi}(x)$ and $\tilde{\psi}(x)$ in the same way, except by swapping the positions of A, B, C with those of $\widetilde{A}$, $\widetilde{B}$, and $\widetilde{C}$. It can be verified that

$$\tilde{\phi}(x) = \phi(x), \quad \tilde{\psi}(x) = -\psi(x) - \phi(x).$$

Also define

$$\begin{aligned} g(t) :&= (x^{\rm H}[B + t\varDelta B]x)^2 - 4(x^{\rm H}[A + t\varDelta A]x)(x^{\rm H}[C + t\varDelta C]x) \\ &= \varsigma(x)^2 + 2\psi(x)t + \phi(x)t^2. \end{aligned}$$

So $g(0) = \varsigma(x)$ and $g(1) = \tilde{\varsigma}(x)$. Correspondingly,

$$\begin{aligned} \tilde{g}(t) :&= (x^{\rm H}[\widetilde{B} - t\varDelta B]x)^2 - 4(x^{\rm H}[\widetilde{A} - t\varDelta A]x)(x^{\rm H}[\widetilde{C} - t\varDelta C]x) \\ &= \tilde{\varsigma}(x)^2 + 2\tilde{\psi}(x)t + \phi(x)t^2. \end{aligned}$$

Note that $g(t) = \tilde{g}(1-t)$.

By definition, if $A \succ 0$, then $\boldsymbol{Q}(\lambda)$ is hyperbolic if and only if $g(0) > 0$ for any nonzero $x \in \mathbb{C}^n$, and if $\widetilde{A} \succ 0$, then $\widetilde{\boldsymbol{Q}}(\lambda)$ is hyperbolic if and only if $g(1) > 0$ for any nonzero $x \in \mathbb{C}^n$.

[4]By *weakly*, we mean the strict positivity in (2.2) is given in to nonnegativity.

Lemma 4.1. *Suppose* $\min\{g(0), g(1)\} \ge 0$*. Then* $g(t) \ge 0$ *for all* $0 \le t \le 1$ *and nonzero* $x \in \mathbb{C}^n$ *if and only if*

$$\min\{\phi(x), -\psi(x), -\tilde{\psi}(x), \psi(x)^2 - \phi(x)\varsigma(x)^2\} \le 0 \ \text{ for all } x \ne 0. \tag{4.4}$$

Proof. The condition (4.4) is equivalent to that for any x, at least one of

$$\phi(x) \le 0, \quad \psi(x) \ge 0, \quad \tilde{\psi}(x) = -\psi(x) - \phi(x) \ge 0, \quad \psi(x)^2 - \phi(x)\varsigma(x)^2 \le 0$$

holds. Note that $g(0) \ge 0$ and $g(1) \ge 0$ by assumption. We first prove that (4.4) implies $g(t) \ge 0$ for all $0 \le t \le 1$ and for any nonzero $x \in \mathbb{C}^n$. To this end, we let $0 \le t \le 1$ and $0 \ne x \in \mathbb{C}^n$.

1. If $\phi(x) \le 0$, then $g(t)$ is concave and thus $g(t) \ge (1-t)g(0) + tg(1) \ge 0$;
2. If $\psi(x) \ge 0$, then
$$\begin{aligned} g(t) &= \varsigma(x)^2 + 2\psi(x)t + \phi(x)t^2 \\ &\ge \varsigma(x)^2 + 2\psi(x)t^2 + \phi(x)t^2 \\ &= (1-t^2)g(0) + t^2 g(1) \\ &\ge 0; \end{aligned}$$
3. If $\tilde{\psi}(x) \ge 0$, then similarly $\tilde{g}(t) \ge (1-t^2)\tilde{g}(0) + t^2\tilde{g}(1) \ge 0$;
4. Consider the case $\psi(x)^2 - \phi(x)\varsigma(x)^2 \le 0$. Suppose[5] $\phi(x) > 0$. Then $g(t)$ is a nontrivial quadratic function and has at most one zero in $\mathbb{R}$. Going through the cases either there is no zero or the zero is in $(0, 1)$ or the zero is outside of $(0, 1)$, we can see $g(t) \ge 0$ for all $0 \le t \le 1$.

Next for the necessity of (4.4), suppose there were an $x \ne 0$ satisfying $\phi(x) > 0$, $\psi(x) < 0$, $-\tilde{\psi}(x) = \psi(x) + \phi(x) > 0$, and $\psi(x)^2 - \phi(x)\varsigma(x)^2 > 0$. Then

$$\min_t g(t) = -\frac{\psi(x)^2 - \phi(x)\varsigma(x)^2}{\phi(x)} < 0$$

and $\min_t g(t)$ is attained at $t_{\min} = -\frac{\psi(x)}{\phi(x)} \in (0, 1)$, contradicting the assumption that $g(t) \ge 0$ for $0 \le t \le 1$. $\square$

Given a shift $\lambda_0 \in \mathbb{R}$, define

$$\begin{aligned} \boldsymbol{Q}_{\lambda_0}(\lambda) := \boldsymbol{Q}(\lambda + \lambda_0) &= \lambda^2 A + \lambda(2\lambda_0 A + B) + \boldsymbol{Q}(\lambda_0) &(4.5)\\ &= \lambda^2 A + \lambda B_{\lambda_0} + C_{\lambda_0}, &(4.6) \end{aligned}$$

where

$$B_{\lambda_0} = 2\lambda_0 A + B, \quad C_{\lambda_0} = \boldsymbol{Q}(\lambda_0). \tag{4.7}$$

It can be verified that (μ, x) *is a quadratic eigenpair of* $\boldsymbol{Q}_{\lambda_0}(\lambda)$ *if and only if* $(\mu + \lambda_0, x)$ *is a quadratic eigenpair of* $\boldsymbol{Q}(\lambda)$.

[5] The case $\phi(x) \le 0$ has already been dealt with.

Lemma 4.2. *Suppose that $\boldsymbol{Q}(\lambda)$ is hyperbolic, and adopt the notations introduced in* Theorem 2.5.

1. *If $\lambda_0 \in (\lambda_n^-, \lambda_1^+)$, then* $\operatorname{diag}(-C_{\lambda_0}, A) = \operatorname{diag}(-\boldsymbol{Q}(\lambda_0), A) \succ 0$.
2. *If $\lambda_0 \in [\lambda_n^+, +\infty)$, then $\boldsymbol{Q}_{\lambda_0}(\lambda)$ is overdamped, i.e. $B_{\lambda_0} \succ 0$ and $C_{\lambda_0} \succeq 0$. Moreover,*
$$-(\lambda_n^- + \lambda_n^+ - 2\lambda_0)A \preceq B_{\lambda_0} \preceq -(\lambda_1^- + \lambda_1^+ - 2\lambda_0)A, \tag{4.8}$$
$$(\lambda_n^- - \lambda_0)(\lambda_n^+ - \lambda_0)A \preceq C_{\lambda_0} \preceq (\lambda_1^- - \lambda_0)(\lambda_1^+ - \lambda_0)A. \tag{4.9}$$
3. *If $\|A^{-1/2}\Delta A A^{-1/2}\|_2 < 1$, then $\widetilde{A} \succ 0$.*

Proof. Item 1 is a consequence of Theorem 2.1 and (4.7). For (4.8) of item 2, we have for any $x \neq 0$
$$\begin{aligned} x^{\mathrm{H}} B_{\lambda_0} x &= 2\lambda_0 x^{\mathrm{H}} A x + x^{\mathrm{H}} B x \\ &= x^{\mathrm{H}} A x \left(2\lambda_0 + \frac{x^{\mathrm{H}} B x}{x^{\mathrm{H}} A x} \right) \\ &= x^{\mathrm{H}} A x \big(2\lambda_0 - [\rho_+(x) + \rho_-(x)] \big) \end{aligned}$$
which, together with (3.4), yields (4.8). For (4.9), we have for any $x \neq 0$
$$x^{\mathrm{H}} C_{\lambda_0} x = x^{\mathrm{H}} \boldsymbol{Q}(\lambda_0) x = x^{\mathrm{H}} A x [\lambda_0 - \rho_+(x)][\lambda_0 - \rho_-(x)]$$
which, together with (3.4), yields (4.9). For item 3, we notice the smallest eigenvalue of $A^{-1/2}\widetilde{A}A^{-1/2}$ satisfies
$$\lambda_{\min}(A^{-1/2}\widetilde{A}A^{-1/2}) = 1 + \lambda_{\min}(A^{-1/2}\Delta A A^{-1/2}) \geq 1 - \|A^{-1/2}\Delta A A^{-1/2}\|_2 > 0$$
if $\|A^{-1/2}\Delta A A^{-1/2}\|_2 < 1$. □

4.2 Asymptotical analysis

It is a common technique to perform an asymptotical analysis in numerical analysis for at least three reasons:

1. it is mathematically sound, provided it is known that the interested quantities are continuous with respect to what is being perturbed;
2. it is relatively easy because it is a first order analysis, and
3. it is powerful in revealing the intrinsic sensitivity of the interested quantities.

Let (μ, x) is a simple quadratic eigenpair of HQEP (1.1) for $\boldsymbol{Q}(\lambda)$. Since HQEP (1.1) is equivalent to the eigenvalue problem for the regular matrix pencil $\mathscr{L}_{\boldsymbol{Q}}(\lambda)$ in (2.5) and since the eigenvalues of a regular matrix pencil and the eigenvectors associated with simple eigenvalues are continuous with respect to the entries of the involved matrices [56], $\widetilde{\boldsymbol{Q}}(\lambda)$ has a simple quadratic eigenpair $(\tilde{\mu}, \tilde{x}) = (\mu + \Delta\mu, x + \Delta x)$ such that $\Delta\mu \to 0$ and $\Delta x \to 0$ as ΔA, ΔB, $\Delta C \to 0$. Now suppose that $\|\Delta A\|$, $\|\Delta B\|$, and $\|\Delta C\|$ are sufficiently tiny, and so are $\Delta\mu$ and $\|\Delta x\|$. Ignoring terms of order 2 or higher and noticing $\boldsymbol{Q}(\mu)x = 0$, we have from $\widetilde{\boldsymbol{Q}}(\mu + \Delta\mu)\,(x + \Delta x) = 0$

$$\Delta\mu\big[2\mu A + B\big]x + \big[\mu^2 \Delta A + \mu \Delta B + \Delta C\big]x + \big[\mu^2 A + \mu B + C\big]\Delta x \approx 0, \tag{4.10}$$

where the "$\approx$" means the equation is true after ignoring terms of order 2 or higher. Pre-multiply (4.10) by $x^{\rm H}$ and use $x^{\rm H}\boldsymbol{Q}(\mu) = 0$ to get

$$\Delta\mu \approx -\frac{x^{\rm H}\big[\mu^2 \Delta A + \mu \Delta B + \Delta C\big]x}{x^{\rm H}\big[2\mu A + B\big]x} \tag{4.11}$$

$$= -\frac{x^{\rm H}\big[\mu^2 \Delta A + \mu \Delta B + \Delta C\big]x}{\varsigma(x)} \tag{4.12}$$

$$= -\frac{\mu^2}{\pm\varsigma(x)} \cdot x^{\rm H}\Delta A x - \frac{\mu}{\pm\varsigma(x)} \cdot x^{\rm H}\Delta B x - \frac{1}{\pm\varsigma(x)} \cdot x^{\rm H}\Delta C x. \tag{4.13}$$

where the equality in (4.12) is due to (3.5). There is a clear interpretation of (4.13): the change $\Delta\mu$ is proportional to ΔA, ΔB, ΔC with multiplying factors $|\mu^2/\varsigma(x)|$, $|\mu/\varsigma(x)|$, and $1/|\varsigma(x)|$, respectively. Our following strict bounds reflect this interpretation.

The expression (4.11) is not new and its derivation follows a rather standard technique (see, e.g., [62]). What is new here is the use of (3.5) to relate its denominator $x^{\rm H}\big[2\mu A{+}B\big]x$ to $\varsigma(x)$, a quantity that determines the hyperbolicity of $\boldsymbol{Q}$.

4.3 Perturbation bounds in the spectral norm

Throughout the rest of this section, we assume $\boldsymbol{Q}(\lambda)$ and $\widetilde{\boldsymbol{Q}}(\lambda)$ are hyperbolic and

$$\|A^{-1/2}\Delta A A^{-1/2}\|_2 < 1 \tag{4.14}$$

which guarantees $\widetilde{A} \succ 0$. We will adopt the notations introduced in Theorem 2.5. Our goal is to bound the norms of

$$\Delta\varLambda_+ = \operatorname{diag}(\tilde\lambda_1^+ - \lambda_1^+, \ldots, \tilde\lambda_n^+ - \lambda_n^+), \quad \Delta\varLambda_- = \operatorname{diag}(\tilde\lambda_1^- - \lambda_1^-, \ldots, \tilde\lambda_n^- - \lambda_n^-).$$

Bounds on norms of the change to $\varLambda = \operatorname{diag}(\varLambda_-, \varLambda_+)$ are easily derivable through

$$\|\Delta\varLambda\|_2 = \max_{\pm}\|\Delta\varLambda_\pm\|_2, \quad \|\Delta\varLambda\|_{\rm F} = \sqrt{\|\Delta\varLambda_+\|_{\rm F}^2 + \|\Delta\varLambda_-\|_{\rm F}^2},$$
$$\|\Delta\varLambda\|_{\rm ui} \le 2\max_{\pm}\|\Delta\varLambda_\pm\|_{\rm ui}.$$

In this subsection, we will focus on the spectral norm, and leave the case for the Frobenius norms and more generally unitarily invariant norms to the next subsection. Our main results of this subsection are summarized in Theorem 4.1.

Theorem 4.1. *Let* $\mathrm{typ} \in \{+,-\}$, *and*

$$\epsilon_a = \|A^{-1/2}\Delta A A^{-1/2}\|_2, \quad \epsilon_b = \frac{\|\Delta B\|_2}{\|B\|_2}, \quad \epsilon_c = \frac{\|\Delta C\|_2}{\|C\|_2}, \tag{4.15}$$

$$\lambda_{\max}^{\rm typ} = \max\{|\lambda_1^{\rm typ}|, |\lambda_n^{\rm typ}|\}, \quad \tilde\lambda_{\max}^{\rm typ} = \max\{|\tilde\lambda_1^{\rm typ}|, |\tilde\lambda_n^{\rm typ}|\}, \tag{4.16}$$

$$\chi_\varsigma = \min_{x\ne 0}\{\varsigma_0(x), \tilde\varsigma_0(x)\}, \quad \chi_{\lambda^{\rm typ}} = \max\{\lambda_{\max}^{\rm typ}, \tilde\lambda_{\max}^{\rm typ}\}. \tag{4.17}$$

1. *If* $\Delta A = \Delta B = 0$ *and*
$$\epsilon_c < \frac{\chi_\varsigma^2}{4\|A\|_2\|C\|_2}, \tag{4.18}$$
then
$$\|\Delta\Lambda_{\rm typ}\|_2 \le \frac{1}{\chi_\varsigma}\,\|\Delta C\|_2. \tag{4.19}$$

2. *If* $\Delta B = \Delta C = 0$ *and*
$$\epsilon_a < \min\left\{1, \frac{\chi_\varsigma^2}{4\|A\|_2\|C\|_2}\right\}, \tag{4.20}$$
then
$$\|\Delta\Lambda_{\rm typ}\|_2 \le \frac{\chi_{\lambda^{\rm typ}}^2}{(1-\epsilon_a)\chi_\varsigma}\,\|\Delta A\|_2. \tag{4.21}$$

3. *If* $\Delta A = \Delta C = 0$ *and*
$$\epsilon_b < \frac{\chi_\varsigma^2}{\|B\|_2(\|B\|_2 + 2\sqrt{\|A\|_2\|C\|_2})}, \tag{4.22}$$
then
$$\|\Delta\Lambda_{\rm typ}\|_2 \le \frac{\chi_{\lambda^{\rm typ}}}{\chi_\varsigma}\,\|\Delta B\|_2 + \frac{\|C\|_2}{\chi_\varsigma^3}\,\|\Delta B\|_2^2. \tag{4.23}$$

4. *If* $\Delta A = \Delta C = 0$ *and*
$$\|\Delta B\|_2 < \frac{\chi_\varsigma^2}{\|2\lambda_0 A + B\|_2 + 2\sqrt{\|A\|_2\|\boldsymbol{Q}(\lambda_0)\|_2}}, \tag{4.24}$$
where $\lambda_0 \in (-\infty, \min\{\lambda_1^-, \tilde\lambda_1^-\}] \cup [\max\{\lambda_n^+, \tilde\lambda_n^+\}, +\infty)$, *then*
$$\|\Delta\Lambda_{\rm typ}\|_2 \le \frac{\chi_{\lambda^{\rm typ}} + |\lambda_0|}{\chi_\varsigma}\,\|\Delta B\|_2. \tag{4.25}$$

5. *In general, without assuming two of* ΔA, ΔB, *and* ΔC *are zeros, if*
$$\epsilon_a < \gamma\min\left\{1, \frac{\chi_\varsigma^2}{4\|A\|_2\|C\|_2}\right\}, \tag{4.26a}$$
$$\epsilon_b < \gamma\frac{\chi_\varsigma^2}{\|B\|_2(\|B\|_2 + 2\sqrt{\|A\|_2\|C\|_2})}, \tag{4.26b}$$
$$\epsilon_c < \gamma\frac{\chi_\varsigma^2}{4\|A\|_2\|C\|_2}, \tag{4.26c}$$
where
$$\gamma = \frac{\chi_\varsigma^2}{\|B\|_2^2 + \chi_\varsigma^2 + \sqrt{(\|B\|_2^2+\chi_\varsigma^2)(\|B\|_2^2+2\chi_\varsigma^2)}} < \sqrt{2}-1, \tag{4.27}$$
then
$$\begin{aligned}\|\Delta\Lambda_{\rm typ}\|_2 \le{}& \frac{4}{(1-\epsilon_a)\chi_\varsigma^3}\|C\|_2\big[\|A\|_2\|C\|_2(\epsilon_a+\epsilon_c)^2 + \|B\|_2^2(\epsilon_b+\epsilon_a)(\epsilon_b+\epsilon_c)\big]\\ &+ \frac{1}{(1-\epsilon_a)\chi_\varsigma}\left[(\chi_{\lambda^{\rm typ}})^2\|\Delta A\|_2 + \chi_{\lambda^{\rm typ}}\|\Delta B\|_2 + \|\Delta C\|_2\right].\end{aligned} \tag{4.28}$$

All bounds by this theorem are strict. They are consistent with the asymptotic expression (4.13) rather well after dropping terms of order 2 or higher in ϵ_a, ϵ_b, and ϵ_c. For example, (4.28) yields

$$\|\Delta\Lambda_{\rm typ}\|_2 \lesssim \frac{1}{\chi_\varsigma}\left[(\chi_{\lambda^{\rm typ}})^2\|\Delta A\|_2 + \chi_{\lambda^{\rm typ}}\|\Delta B\|_2 + \|\Delta C\|_2\right]. \tag{4.29}$$

The rest of this subsection is devoted for the proof of Theorem 4.1. Later in the next subsection, we will extend (4.19) to a general unitarily invariant norm.

Each of many expressions below is in its compact form for two. For example, (4.30) includes two displayed equations: one for $\Delta\rho_+$ and one for $\Delta\rho_+$ with all "$\pm$" selected as either "$+$" or "$-$", accordingly.

Lemma 4.3. *If* (4.4) *and* (4.14) *hold, then there exists* $0 \le \xi \le 1$ *such that*

$$\Delta\rho_\pm(x) = \delta^\pm(x,\xi) := \pm\left[\delta_3(x,\xi) - \frac{x^{\rm H}Ax}{x^{\rm H}\widetilde{A}x}\,\delta_2^\pm(x)\right] \tag{4.30}$$

for any $x \neq 0$, *where*

$$\delta_2^\pm(x) = \frac{\rho_\pm(x)^2(x^{\rm H}\Delta Ax) + \rho_\pm(x)(x^{\rm H}\Delta Bx) + x^{\rm H}\Delta Cx}{\varsigma(x)}, \tag{4.31a}$$

$$\delta_3(x,\xi) = \frac{\varsigma(x)^2\phi(x) - \psi(x)^2}{4(x^{\rm H}\widetilde{A}x)\left[\varsigma(x)^2 + 2\psi(x)\xi + \phi(x)\xi^2\right]^{3/2}}, \tag{4.31b}$$

$\phi(x)$ *and* $\psi(x)$ *are defined in* (4.2) *and* (4.3). *In addition,*

$$\frac{1}{1+\|A^{-1/2}\Delta AA^{-1/2}\|_2} \le \frac{x^{\rm H}Ax}{x^{\rm H}\widetilde{A}x} \le \frac{1}{1-\|A^{-1/2}\Delta AA^{-1/2}\|_2}, \tag{4.32}$$

$$|\delta_2^\pm(x)| \le \frac{\max\{|\lambda_1^\pm|^2, |\lambda_n^\pm|^2\}\|\Delta A\|_2 + \max\{|\lambda_1^\pm|, |\lambda_n^\pm|\}\|\Delta B\|_2 + \|\Delta C\|_2}{\min\limits_{x\neq 0}\varsigma_0(x)}. \tag{4.33}$$

Proof. According to how the difference operator Δ is defined at the beginning of subsection 4.1, we have

$$\pm\Delta\rho_\pm(x) = \frac{\Delta\varsigma(x) \mp x^{\rm H}\Delta Bx}{2(x^{\rm H}Ax)} + \frac{\tilde{\varsigma}(x) \mp x^{\rm H}\widetilde{B}x}{2}\Delta\left(\frac{1}{x^{\rm H}Ax}\right) =: \epsilon_1 + \epsilon_2. \tag{4.34}$$

The rest of this proof is to calculate ϵ_1 and ϵ_2. By Lemma 4.1,

$$f(t;x) := \left[\varsigma(x)^2 + 2\psi(x)t + \phi(x)t^2\right]^{1/2} \tag{4.35}$$

is well-defined and differentiable for $0 \le t \le 1$. By the Taylor expansion, there exists $0 \le \xi \le 1$ such that

$$\begin{aligned}\tilde{\varsigma}(x) = f(1;x) &= f(0;x) + f'(0;x) + \frac{1}{2}f''(\xi;x)\\ &= \varsigma(x) + \frac{\psi(x)}{\varsigma(x)} + \frac{\varsigma(x)^2\phi(x) - \psi(x)^2}{2[f(\xi;x)]^3}.\end{aligned} \tag{4.36}$$

This ξ depends on x. Now we are ready to calculate ϵ_1 and ϵ_2. We have

$$\begin{aligned}
\epsilon_1 &= \mp\frac{x^{\mathrm{H}}\Delta Bx}{2(x^{\mathrm{H}}Ax)} + \frac{1}{2(x^{\mathrm{H}}Ax)}\left(\frac{\psi(x)}{\varsigma(x)} + \frac{\varsigma(x)^2\phi(x)-\psi(x)^2}{2[f(\xi;x)]^3}\right) \\
&= \mp\frac{x^{\mathrm{H}}\Delta Bx}{2(x^{\mathrm{H}}Ax)} + \frac{(x^{\mathrm{H}}Bx)(x^{\mathrm{H}}\Delta Bx)}{2(x^{\mathrm{H}}Ax)\varsigma(x)} - \frac{x^{\mathrm{H}}\Delta Cx}{\varsigma(x)} - \frac{x^{\mathrm{H}}Cx}{\varsigma(x)}\frac{x^{\mathrm{H}}\Delta Ax}{x^{\mathrm{H}}Ax} + \frac{\varsigma(x)^2\phi(x)-\psi(x)^2}{4(x^{\mathrm{H}}Ax)[f(\xi;x)]^3} \\
&= -\frac{\pm\varsigma(x)-(x^{\mathrm{H}}Bx)}{2(x^{\mathrm{H}}Ax)}\frac{x^{\mathrm{H}}\Delta Bx}{\varsigma(x)} - \frac{x^{\mathrm{H}}\Delta Cx}{\varsigma(x)} - \frac{x^{\mathrm{H}}Cx}{\varsigma(x)}\frac{x^{\mathrm{H}}\Delta Ax}{x^{\mathrm{H}}Ax} + \frac{\varsigma(x)^2\phi(x)-\psi(x)^2}{4(x^{\mathrm{H}}Ax)[f(\xi;x)]^3} \\
&= -\frac{\rho_\pm(x)(x^{\mathrm{H}}\Delta Bx)}{\varsigma(x)} - \frac{x^{\mathrm{H}}\Delta Cx}{\varsigma(x)} - \frac{x^{\mathrm{H}}Cx}{\varsigma(x)}\frac{x^{\mathrm{H}}\Delta Ax}{x^{\mathrm{H}}Ax} + \frac{x^{\mathrm{H}}\widetilde{A}x}{x^{\mathrm{H}}Ax}\frac{\varsigma(x)^2\phi(x)-\psi(x)^2}{4(x^{\mathrm{H}}\widetilde{A}x)[f(\xi;x)]^3} \\
&= -\delta_2^\pm(x) + \frac{\rho_\pm(x)^2(x^{\mathrm{H}}\Delta Ax)}{\varsigma(x)} - \frac{x^{\mathrm{H}}Cx}{\varsigma(x)}\frac{x^{\mathrm{H}}\Delta Ax}{x^{\mathrm{H}}Ax} + \frac{x^{\mathrm{H}}\widetilde{A}x}{x^{\mathrm{H}}Ax}\delta_3(x,\xi)
\end{aligned}$$

and

$$\epsilon_2 = -\frac{[\tilde{\varsigma}(x)\mp x^{\mathrm{H}}\widetilde{B}x](x^{\mathrm{H}}\Delta Ax)}{2(x^{\mathrm{H}}\widetilde{A}x)(x^{\mathrm{H}}Ax)} = \frac{\mp\tilde{\rho}_\pm(x)(x^{\mathrm{H}}\Delta Ax)}{x^{\mathrm{H}}Ax} = -[\pm\rho_\pm(x)\pm\Delta\rho_\pm(x)]\frac{x^{\mathrm{H}}\Delta Ax}{x^{\mathrm{H}}Ax}.$$

Noticing

$$\begin{aligned}
\frac{x^{\mathrm{H}}Cx}{\varsigma(x)} \pm \rho_\pm(x) &= \frac{x^{\mathrm{H}}Cx}{\varsigma(x)} \pm \frac{-x^{\mathrm{H}}Bx\pm\varsigma(x)}{2(x^{\mathrm{H}}Ax)} \\
&= \frac{2(x^{\mathrm{H}}Ax)(x^{\mathrm{H}}Cx)\mp x^{\mathrm{H}}Bx\varsigma(x)+\varsigma(x)^2}{2\varsigma(x)(x^{\mathrm{H}}Ax)} \\
&= \frac{(x^{\mathrm{H}}Bx)^2-\varsigma(x)^2\mp 2(x^{\mathrm{H}}Bx)\varsigma(x)+2\varsigma(x)^2}{4\varsigma(x)(x^{\mathrm{H}}Ax)} \\
&= \frac{[x^{\mathrm{H}}Bx\mp\varsigma(x)]^2}{4\varsigma(x)(x^{\mathrm{H}}Ax)} = \frac{\rho_\pm(x)^2(x^{\mathrm{H}}Ax)}{\varsigma(x)},
\end{aligned}$$

we have

$$\pm\Delta\rho_\pm(x) = \epsilon_1+\epsilon_2 = -\delta_2^\pm(x) + \frac{x^{\mathrm{H}}\widetilde{A}x}{x^{\mathrm{H}}Ax}\delta_3(x,\xi) - [\pm\Delta\rho_\pm(x)]\frac{x^{\mathrm{H}}\Delta Ax}{x^{\mathrm{H}}Ax}$$

solving which for $\pm\Delta\rho_\pm(x)$ leads to $\Delta\rho_\pm(x) = \delta^\pm(x,\xi)$. □

Lemma 4.4. *Suppose* (4.4) *and* (4.14) *hold. Let* $\delta_{\mathrm{lb}}^\pm(x)$*,* $\delta_{\mathrm{ub}}^\pm(x)$*,* $\tilde{\delta}_{\mathrm{lb}}^\pm(x)$*, and* $\tilde{\delta}_{\mathrm{ub}}^\pm(x)$ *be functions satisfying*

$$\delta_{\mathrm{lb}}^\pm(x) \le \delta^\pm(x,\xi) \le \delta_{\mathrm{ub}}^\pm(x), \quad \tilde{\delta}_{\mathrm{lb}}^\pm(x) \le \tilde{\delta}^\pm(x,\xi) \le \tilde{\delta}_{\mathrm{ub}}^\pm(x) \tag{4.37}$$

for all $x\in\mathbb{C}^n$*,* $\xi\in[0,1]$*, where* $\delta^\pm(x,\xi)$ *is defined as in* Lemma 4.3*. Write*

$$\begin{aligned}
\gamma_{\mathrm{uu}}^\pm &= \max_{x\ne 0}\{\ \ \delta_{\mathrm{ub}}^\pm(x), \tilde{\delta}_{\mathrm{ub}}^\pm(x)\}, &\quad \gamma_{\mathrm{ll}}^\pm &= \max_{x\ne 0}\{-\delta_{\mathrm{lb}}^\pm(x), -\tilde{\delta}_{\mathrm{lb}}^\pm(x)\}, \\
\gamma_{\mathrm{lu}}^\pm &= \max_{x\ne 0}\{-\delta_{\mathrm{lb}}^\pm(x), \delta_{\mathrm{ub}}^\pm(x)\}, &\quad \tilde{\gamma}_{\mathrm{lu}}^\pm &= \max_{x\ne 0}\{-\tilde{\delta}_{\mathrm{lb}}^\pm(x),\ \ \tilde{\delta}_{\mathrm{ub}}^\pm(x)\}.
\end{aligned}$$

Then

$$\|\Delta\Lambda_\pm\|_2 = \max_{1\le i\le n}|\Delta\lambda_i^\pm| \le \min\{\gamma_{\mathrm{uu}}^\pm, \gamma_{\mathrm{ll}}^\pm, \gamma_{\mathrm{lu}}^\pm, \tilde{\gamma}_{\mathrm{lu}}^\pm\}. \tag{4.38}$$

Proof. We only consider the "+" case below; the "−" case is similar. In fact simply replacing "+" with "−" gives a proof for the "−" case.

By Lemma 4.3,

$$\delta_{\mathrm{lb}}^{+}(x) \le \Delta\rho_{+}(x) = \delta^{+}(x,\xi) \le \delta_{\mathrm{ub}}^{+}(x).$$

Let $\mathcal{S}_i = \operatorname{span}\{u_1^+, \cdots, u_i^+\}, \mathcal{T}_i = \operatorname{span}\{u_i^+, \cdots, u_n^+\}$ and similarly define $\widetilde{\mathcal{S}}_i$ and $\widetilde{\mathcal{T}}_i$. By the Courant-Fischer type min-max principles in Theorem 3.2,

$$\begin{aligned}
\lambda_i^+ &= \min_{\dim \mathcal{X}=i} \max_{0\ne x\in\mathcal{X}} \rho_+(x) = \max_{0\ne x\in\mathcal{S}_i} \rho_+(x) = \rho_+(u_i^+),\\
\tilde{\lambda}_i^+ &= \min_{\dim \mathcal{X}=i} \max_{0\ne x\in\mathcal{X}} \tilde{\rho}_+(x) = \max_{0\ne x\in\widetilde{\mathcal{S}}_i} \tilde{\rho}_+(x) = \widetilde{\rho}_+(\widetilde{u}_i^+),\\
\lambda_i^+ &= \max_{\operatorname{codim} \mathcal{X}=i-1} \min_{0\ne x\in\mathcal{X}} \rho_+(x) = \min_{0\ne x\in\mathcal{T}_i} \rho_+(x) = \rho_+(u_i^+),\\
\tilde{\lambda}_i^+ &= \max_{\operatorname{codim} \mathcal{X}=i-1} \min_{0\ne x\in\mathcal{X}} \tilde{\rho}_+(x) = \min_{0\ne x\in\widetilde{\mathcal{T}}_i} \tilde{\rho}_+(x) = \widetilde{\rho}_+(\widetilde{u}_i^+).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\tilde{\lambda}_i^+ = \min_{\dim \mathcal{X}=i} \max_{0\ne x\in\mathcal{X}} \tilde{\rho}_+(x) &\le \max_{0\ne x\in\mathcal{S}_i} \tilde{\rho}_+(x)\\
&\le \max_{0\ne x\in\mathcal{S}_i} \left[\rho_+(x) + \delta_{\mathrm{ub}}^+(x)\right]\\
&\le \max_{0\ne x\in\mathcal{S}_i} \rho_+(x) + \max_{0\ne x\in\mathcal{S}_i} \delta_{\mathrm{ub}}^+(x)\\
&= \lambda_i^+ + \max_{0\ne x\in\mathcal{S}_i} \delta_{\mathrm{ub}}^+(x),\\
\tilde{\lambda}_i^+ = \max_{\operatorname{codim} \mathcal{X}=i-1} \min_{0\ne x\in\mathcal{X}} \tilde{\rho}_+(x) &\ge \min_{0\ne x\in\mathcal{T}_i} \tilde{\rho}_+(x)\\
&\ge \min_{0\ne x\in\mathcal{T}_i} \left[\rho_+(x) + \delta_{\mathrm{lb}}^+(x)\right]\\
&\ge \min_{0\ne x\in\mathcal{T}_i} \rho_+(x) + \min_{0\ne x\in\mathcal{T}_i} \delta_{\mathrm{lb}}^+(x)\\
&= \lambda_i^+ + \min_{0\ne x\in\mathcal{T}_i} \delta_{\mathrm{lb}}^+(x).
\end{aligned}$$

They give (4.39a) below and (4.39b) as well, by switching the roles of $\boldsymbol{Q}$ and $\widetilde{\boldsymbol{Q}}$:

$$\min_{0\ne x\in\mathcal{T}_i} \delta_{\mathrm{lb}}^+(x) \le \tilde{\lambda}_i^+ - \lambda_i^+ \le \max_{0\ne x\in\mathcal{S}_i} \delta_{\mathrm{ub}}^+(x), \tag{4.39a}$$

$$\min_{0\ne x\in\widetilde{\mathcal{T}}_i} \tilde{\delta}_{\mathrm{lb}}^+(x) \le \lambda_i^+ - \tilde{\lambda}_i^+ \le \max_{0\ne x\in\widetilde{\mathcal{S}}_i} \tilde{\delta}_{\mathrm{ub}}^+(x). \tag{4.39b}$$

It follows from (4.39) that

$$\begin{aligned}
|\Delta\lambda_i^+| &\le \max\left\{\max_{0\ne x\in\mathcal{S}_i} \delta_{\mathrm{ub}}^+(x), \max_{0\ne x\in\widetilde{\mathcal{S}}_i} \tilde{\delta}_{\mathrm{ub}}^+(x)\right\}\\
&\le \max_{x\ne 0}\{\delta_{\mathrm{ub}}^+(x), \tilde{\delta}_{\mathrm{ub}}^+(x)\} = \gamma_{\mathrm{uu}}^+,\\
|\Delta\lambda_i^+| &\le \max\left\{-\min_{0\ne x\in\mathcal{T}_i} \delta_{\mathrm{lb}}^+(x), -\min_{0\ne x\in\widetilde{\mathcal{T}}_i} \tilde{\delta}_{\mathrm{lb}}^+(x)\right\}
\end{aligned}$$

$$\le \max_{x\ne 0}\{-\delta^+_{\rm lb}(x), -\tilde\delta^+_{\rm lb}(x)\} = \gamma^+_{\rm ll},$$

$$|\Delta\lambda^+_i| \le \max\left\{-\min_{0\ne x\in\mathcal{T}_i}\delta^+_{\rm lb}(x), \max_{0\ne x\in\mathcal{S}_i}\delta^+_{\rm ub}(x)\right\}$$

$$\le \max_{x\ne 0}\{-\delta^+_{\rm lb}(x), \delta^+_{\rm ub}(x)\} = \gamma^+_{\rm lu},$$

$$|\Delta\lambda^+_i| \le \max\left\{-\min_{0\ne x\in\widetilde{\mathcal{T}}_i}\tilde\delta^+_{\rm lb}(x), \max_{0\ne x\in\widetilde{\mathcal{S}}_i}\tilde\delta^+_{\rm ub}(x)\right\}$$

$$\le \max_{x\ne 0}\{-\tilde\delta^+_{\rm lb}(x), \tilde\delta^+_{\rm ub}(x)\} = \tilde\gamma^+_{\rm lu}.$$

This completes the proof of (4.38) for the "+" case. □

Proof of Theorem 4.1. We only prove the perturbation results for Λ_+. The case for Λ_- can be turned into one for Λ_+ by considering the pos-type quadratic eigenvalues of $\boldsymbol{Q}(-\lambda)$ and $\widetilde{\boldsymbol{Q}}(-\lambda)$.

For any $\alpha > 0, x \ne 0$, we have

$$\epsilon_a < \alpha \qquad \Rightarrow \qquad |x^{\rm H}\Delta Ax| < \alpha x^{\rm H}Ax, \tag{4.40a}$$

$$\epsilon_a < \alpha\frac{\chi_\varsigma^2}{4\|A\|_2\|C\|_2} \qquad \Rightarrow \qquad |x^{\rm H}\Delta Ax| < \alpha\frac{\varsigma(x)^2}{4|x^{\rm H}Cx|}, \tag{4.40b}$$

$$\epsilon_c < \alpha\frac{\chi_\varsigma^2}{4\|A\|_2\|C\|_2} \qquad \Rightarrow \qquad |x^{\rm H}\Delta Cx| < \alpha\frac{\varsigma(x)^2}{4x^{\rm H}Ax}, \tag{4.40c}$$

$$\epsilon_b < \alpha\frac{\chi_\varsigma^2}{\|B\|_2(\|B\|_2+2\sqrt{\|A\|_2\|C\|_2})} \qquad \Rightarrow \qquad |x^{\rm H}\Delta Bx| < \alpha|x^{\rm H}Bx|, \tag{4.40d}$$

where (4.40a) and (4.40b) hold because

$$\left|\frac{x^{\rm H}\Delta Ax}{x^{\rm H}Ax}\right| = \left|\frac{x^{\rm H}A^{1/2}(A^{-1/2}\Delta AA^{-1/2})A^{1/2}x}{x^{\rm H}A^{1/2}A^{1/2}x}\right| \le \|A^{-1/2}\Delta AA^{-1/2}\|_2 = \epsilon_a,$$

and (4.40d) holds because the left part tells

$$|x^{\rm H}\Delta Bx| < \alpha\frac{\varsigma(x)^2}{|x^{\rm H}Bx| + \sqrt{4(x^{\rm H}Ax)|x^{\rm H}Cx|}} = \alpha\left(|x^{\rm H}Bx| - \sqrt{4(x^{\rm H}Ax)|x^{\rm H}Cx|}\right). \tag{4.41}$$

For item 1: $\Delta A = \Delta B = 0$, $\phi(x) = \tilde\phi(x) = 0, \psi(x) = -2(x^{\rm H}Ax)(x^{\rm H}\Delta Cx)$ and (4.14) holds. Under the assumption (4.18), (4.40c) holds with $\alpha = 1$. Thus $g(1) = \varsigma(x)^2 + 2\psi(x) + \phi(x) > 0$, or equivalently the perturbed quadratic polynomial is still hyperbolic. Note (4.4) holds for $\phi(x) = 0$. Thus $\delta_3(x,\xi) \le 0$ and $\tilde\delta_3(x,\xi) \le 0$. We can take, in (4.37),

$$\delta^+_{\rm ub}(x) = -\delta^+_2(x) = -\frac{x^{\rm H}\Delta Cx}{\varsigma(x)}, \quad \tilde\delta^+_{\rm ub}(x) = -\tilde\delta^+_2(x) = \frac{x^{\rm H}\Delta Cx}{\tilde\varsigma(x)} \tag{4.42}$$

to give

$$|\delta^+_{\rm ub}(x)| \le \frac{\|\Delta C\|_2}{\min_{x\ne 0}\varsigma_0(x)}, \quad |\tilde\delta^+_{\rm ub}(x)| \le \frac{\|\Delta C\|_2}{\min_{x\ne 0}\tilde\varsigma_0(x)}.$$

Using (4.38), we have $\|\Delta\Lambda_+\|_2 \le \gamma_{\rm uu}^+$ to get (4.19).

For item 2: $\Delta B = \Delta C = 0$, $\phi(x) = \tilde\phi(x) = 0, \psi(x) = -2(x^{\rm H}Cx)(x^{\rm H}\Delta Ax)$. Under the assumption (4.20), (4.14) holds; (4.40a) and (4.40b) hold with $\alpha = 1$. Thus $g(1) = \varsigma(x)^2 + 2\psi(x) + \phi(x) > 0$, or equivalently the perturbed quadratic polynomial is still hyperbolic. Note (4.4) holds for $\phi(x) = 0$. Thus $\delta_3(x,\xi) \le 0$ and $\tilde\delta_3(x,\xi) \le 0$. We can take, in (4.37),

$$\begin{aligned}\delta_{\rm ub}^+(x) &= -\frac{x^{\rm H}Ax}{x^{\rm H}\widetilde Ax}\,\delta_2^+(x) = -\frac{x^{\rm H}Ax}{x^{\rm H}\widetilde Ax}\,\frac{\rho_+(x)^2(x^{\rm H}\Delta Ax)}{\varsigma(x)},\\ \tilde\delta_{\rm ub}^+(x) &= -\frac{x^{\rm H}\widetilde Ax}{x^{\rm H}Ax}\,\tilde\delta_2^+(x) = \frac{x^{\rm H}\widetilde Ax}{x^{\rm H}Ax}\,\frac{\tilde\rho_+(x)^2(x^{\rm H}\Delta Ax)}{\tilde\varsigma(x)},\end{aligned}$$

along with (4.32), to give

$$|\delta_{\rm ub}^+(x)| \le \frac{1}{1-\epsilon_a}\frac{(\lambda_{\max}^+)^2\|\Delta A\|_2}{\min_{x\ne0}\varsigma_0(x)}, \quad |\tilde\delta_{\rm ub}^+(x)| \le (1+\epsilon_a)\frac{(\tilde\lambda_{\max}^+)^2\|\Delta A\|_2}{\min_{x\ne0}\tilde\varsigma_0(x)}.$$

Using (4.38), we have $\|\Delta\Lambda_+\|_2 \le \gamma_{\rm uu}^+$ to get (4.21).

For item 3: $\Delta A = \Delta C = 0$, $\phi(x) = \tilde\phi(x) = (x^{\rm H}Bx)(x^{\rm H}\Delta Bx), \psi(x) = (x^{\rm H}\Delta Bx)^2$ and (4.14) holds. Under the assumption (4.22), (4.40d) and (4.41) hold with $\alpha = 1$. (4.41) tells

$$\sqrt{4(x^{\rm H}Ax)|x^{\rm H}Cx|} < |x^{\rm H}Bx| - |x^{\rm H}\Delta Bx| \le |x^{\rm H}Bx + x^{\rm H}\Delta Bx|.$$

Thus

$$\begin{aligned}g(1) &= \varsigma(x)^2 + 2\psi(x) + \phi(x)\\ &= (x^{\rm H}\Delta Bx)^2 + 2(x^{\rm H}\Delta Bx)(x^{\rm H}Bx) + (x^{\rm H}Bx)^2 - 4(x^{\rm H}Ax)(x^{\rm H}Cx)\\ &\ge \left[x^{\rm H}\Delta Bx + x^{\rm H}Bx - \sqrt{4(x^{\rm H}Ax)|x^{\rm H}Cx|}\right]\left[x^{\rm H}\Delta Bx + x^{\rm H}Bx + \sqrt{4(x^{\rm H}Ax)|x^{\rm H}Cx|}\right]\\ &> 0,\end{aligned}$$

or equivalently the perturbed quadratic polynomial is still hyperbolic. (4.40d) tells $|\psi(x)| = |x^{\rm H}Bx| > |x^{\rm H}\Delta Bx| = \phi(x)$. Thus (4.4) holds. Notice

$$\begin{aligned}\varsigma(x)^2\phi(x) - \psi(x)^2 &= \varsigma(x)^2(x^{\rm H}\Delta Bx)^2 - [(x^{\rm H}Bx)(x^{\rm H}\Delta Bx)]^2\\ &= -4(x^{\rm H}Ax)(x^{\rm H}Cx)(x^{\rm H}\Delta Bx)^2\end{aligned}$$

to get

$$\delta_3(x,\xi) = -\frac{(x^{\rm H}Cx)(x^{\rm H}\Delta Bx)^2}{[f(\xi;x)]^3},$$

where $f(\xi;x) = \left[\varsigma(x)^2 + 2\psi(x)\xi + \phi(x)\xi^2\right]^{1/2}$. Since[6]

$$\min_{0\le\xi\le1} f(\xi;x) = \min\{f(0), f(1)\} = \min\{\varsigma(x), \tilde\varsigma(x)\}, \tag{4.43}$$

[6] For the quadratic function $h(t) = a(t-c)^2 + b$ with $a > 0$, if $|c| \ge 1$, i.e., c, the minimal point of $h(t)$ for $t \in \mathbb{R}$, is not in the interval $(0,1)$, then the minimal point of $h(t)$ on $[0,1]$ must be either 0 or 1. For the case here, $c = \psi(x)/\phi(x)$.

we can take, in (4.37),

$$\delta^+_{\rm ub}(x) = -\delta^+_2(x) + \frac{|x^{\rm H}Cx||x^{\rm H}\Delta Bx|^2}{\min\{\varsigma(x),\tilde\varsigma(x)\}^3} = -\frac{\rho_+(x)(x^{\rm H}\Delta Bx)}{\varsigma(x)} + \frac{|x^{\rm H}Cx||x^{\rm H}\Delta Bx|^2}{\min\{\varsigma(x),\tilde\varsigma(x)\}^3},$$
$$\tilde\delta^+_{\rm ub}(x) = -\tilde\delta^+_2(x) + \frac{|x^{\rm H}\widetilde Cx||x^{\rm H}\Delta Bx|^2}{\min\{\varsigma(x),\tilde\varsigma(x)\}^3} = \frac{\tilde\rho_+(x)(x^{\rm H}\Delta Bx)}{\tilde\varsigma(x)} + \frac{|x^{\rm H}\widetilde Cx||x^{\rm H}\Delta Bx|^2}{\min\{\varsigma(x),\tilde\varsigma(x)\}^3}$$

to give

$$|\delta^+_{\rm ub}(x)| \le \frac{\lambda^+_{\max}}{\min_{x\ne 0}\varsigma_0(x)}\,\|\Delta B\|_2 + \frac{\|C\|_2}{\chi_\varsigma^3}\,\|\Delta B\|_2^2,$$
$$|\tilde\delta^+_{\rm ub}(x)| \le \frac{\tilde\lambda^+_{\max}}{\min_{x\ne 0}\tilde\varsigma_0(x)}\,\|\Delta B\|_2 + \frac{\|\widetilde C\|_2}{\chi_\varsigma^3}\,\|\Delta B\|_2^2.$$

Using (4.38), we have $\|\Delta\Lambda_+\|_2 \le \gamma^+_{\rm uu}$ to get (4.23).

For item 4: $\Delta A = \Delta C = 0$, consider the shifted $\boldsymbol{Q}_{\lambda_0}(\lambda)$. By item 2 of Lemma 4.2, $\boldsymbol{Q}_{\lambda_0}(\lambda)$ and $\widetilde{\boldsymbol{Q}}_{\lambda_0}(\lambda)$ are overdamped for $\lambda_0 \in (-\infty, \min\{\lambda_1^-,\tilde\lambda_1^-\}] \cup [\max\{\lambda_n^+,\tilde\lambda_n^+\},+\infty)$. In particular, $B_{\lambda_0} \succ 0, C_{\lambda_0} \succeq 0; \widetilde B_{\lambda_0} \succ 0, \widetilde C_{\lambda_0} \succeq 0$. Note $\varsigma_{\lambda_0}(x) \equiv \varsigma(x), \widetilde\varsigma_{\lambda_0}(x) \equiv \widetilde\varsigma(x)$. Under the assumption (4.24), like[7] in item 3, $|\psi_{\lambda_0}(x)| > \phi_{\lambda_0}(x)$. Thus (4.4) for $\boldsymbol{Q}_{\lambda_0}(\lambda)$ and $\widetilde{\boldsymbol{Q}}_{\lambda_0}(\lambda)$ holds. Just as in item 3 (note $\Delta B_{\lambda_0} = \Delta B$ since $\Delta A = 0$),

$$\varsigma_{\lambda_0}(x)^2\phi_{\lambda_0}(x) - \psi_{\lambda_0}(x)^2 = -4(x^{\rm H}Ax)(x^{\rm H}C_{\lambda_0}x)(x^{\rm H}\Delta Bx)^2 < 0$$

which infers $\delta_{3;\lambda_0}(x,\xi) \le 0$ and thus we can take, in (4.37),

$$\delta^+_{{\rm ub};\lambda_0}(x) = -\delta^+_{2;\lambda_0}(x) = -\frac{\rho_{+;\lambda_0}(x)(x^{\rm H}\Delta Bx)}{\varsigma(x)},$$
$$\tilde\delta^+_{{\rm ub};\lambda_0}(x) = -\tilde\delta^+_{2;\lambda_0}(x) = -\frac{\tilde\rho_{+;\lambda_0}(x)(x^{\rm H}\Delta Bx)}{\tilde\varsigma(x)}$$

to give

$$|\delta^+_{{\rm ub};\lambda_0}(x)| \le \frac{\lambda^+_{\max;\lambda_0}}{\min_{x\ne 0}\varsigma_0(x)}\,\|\Delta B\|_2, \quad |\tilde\delta^+_{{\rm ub};\lambda_0}(x)| \le \frac{\tilde\lambda^+_{\max;\lambda_0}}{\min_{x\ne 0}\tilde\varsigma_0(x)}\,\|\Delta B\|_2.$$

Using (4.38), we have $\|\Delta\Lambda_{+;\lambda_0}\|_2 \le \gamma^+_{{\rm uu};\lambda_0}$ to get (4.25).

For item 5, under the assumption (4.26), $\epsilon_a < \gamma < 1$ and (4.40) holds with $\alpha = \gamma$. Then (4.14) holds, and

$$\begin{aligned}
|\psi(x)| &\le |x^{\rm H}Bx||x^{\rm H}\Delta Bx| + 2(x^{\rm H}Ax)|x^{\rm H}\Delta Cx| + 2|x^{\rm H}Cx||x^{\rm H}\Delta Ax|\\
&< |x^{\rm H}Bx|^2\gamma + \frac{\varsigma(x)^2}{2}\gamma + \frac{\varsigma(x)^2}{2}\gamma\\
&= [|x^{\rm H}Bx|^2 + \varsigma(x)^2]\gamma,\\
|\phi(x)| &\le |x^{\rm H}\Delta Bx|^2 + 4|x^{\rm H}\Delta Ax||x^{\rm H}\Delta Cx|
\end{aligned}$$

[7] We will use the same symbols as those for $\boldsymbol{Q}$ but with the subscript "λ_0" to represent the corresponding quantities for $\boldsymbol{Q}_{\lambda_0}$.

$$
\begin{aligned}
&< |x^{\mathrm{H}}Bx|^2\gamma^2 + |x^{\mathrm{H}}\Delta Ax|\frac{\varsigma(x)^2\gamma}{x^{\mathrm{H}}Ax} \\
&< |x^{\mathrm{H}}Bx|^2\gamma^2 + \varsigma(x)^2\gamma^2 \\
&= [|x^{\mathrm{H}}Bx|^2 + \varsigma(x)^2]\gamma^2,
\end{aligned}
$$

which infers

$$
\begin{aligned}
g(1) &= \varsigma(x)^2 + 2\psi(x) + \phi(x) \\
&> \varsigma(x)^2(1 - 2\gamma - \gamma^2) - |x^{\mathrm{H}}Bx|^2(2\gamma + \gamma^2) \\
&\geq (x^{\mathrm{H}}x)^2\left[\chi_\varsigma^2(1 - 2\gamma - \gamma^2) - \|B\|_2^2(2\gamma + \gamma^2)\right] \\
&= (x^{\mathrm{H}}x)^2\left[\chi_\varsigma^2 - (\|B\|_2^2 + \chi_\varsigma^2)(2\gamma + \gamma^2)\right] \\
&= 0,
\end{aligned}
$$

or equivalently the perturbed quadratic polynomial is still hyperbolic. By the same reasoning we had for items 1, 2 and 3, (4.4) holds and at the same time, we have (4.43). Note that

$$
\begin{aligned}
\varsigma(x)^2\phi(x) - \psi(x)^2 = &-4\big[(x^{\mathrm{H}}Ax)(x^{\mathrm{H}}\Delta Cx) - (x^{\mathrm{H}}Cx)(x^{\mathrm{H}}\Delta Ax)\big]^2 \\
&- 4\big[(x^{\mathrm{H}}Ax)(x^{\mathrm{H}}\Delta Bx) - (x^{\mathrm{H}}Bx)(x^{\mathrm{H}}\Delta Ax)\big]\times \\
&\qquad\big[(x^{\mathrm{H}}Cx)(x^{\mathrm{H}}\Delta Bx) - (x^{\mathrm{H}}Bx)(x^{\mathrm{H}}\Delta Cx)\big],
\end{aligned}
$$

and similarly

$$
\begin{aligned}
\tilde{\varsigma}(x)^2\tilde{\phi}(x) - \tilde{\psi}(x)^2 = &-4\big[-(x^{\mathrm{H}}\widetilde{A}x)(x^{\mathrm{H}}\Delta Cx) + (x^{\mathrm{H}}\widetilde{C}x)(x^{\mathrm{H}}\Delta Ax)\big]^2 \\
&- 4\big[-(x^{\mathrm{H}}\widetilde{A}x)(x^{\mathrm{H}}\Delta Bx) + (x^{\mathrm{H}}\widetilde{B}x)(x^{\mathrm{H}}\Delta Ax)\big]\times \\
&\qquad\big[-(x^{\mathrm{H}}\widetilde{C}x)(x^{\mathrm{H}}\Delta Bx) + (x^{\mathrm{H}}\widetilde{B}x)(x^{\mathrm{H}}\Delta Cx)\big] \\
= &\ \varsigma(x)^2\phi(x) - \psi(x)^2.
\end{aligned}
$$

Now take

$$
\begin{aligned}
\delta_{\mathrm{ub}}^{+}(x) &= -\frac{x^{\mathrm{H}}Ax}{x^{\mathrm{H}}\widetilde{A}x}\delta_2^{+}(x) + \frac{|\varsigma(x)^2\phi(x) - \psi(x)^2|}{(x^{\mathrm{H}}\widetilde{A}x)\min\{\varsigma(x), \tilde{\varsigma}(x)\}^3}, \\
\tilde{\delta}_{\mathrm{ub}}^{+}(x) &= -\frac{x^{\mathrm{H}}\widetilde{A}x}{x^{\mathrm{H}}Ax}\tilde{\delta}_2^{+}(x) + \frac{|\varsigma(x)^2\phi(x) - \psi(x)^2|}{(x^{\mathrm{H}}Ax)\min\{\varsigma(x), \tilde{\varsigma}(x)\}^3}
\end{aligned}
$$

in (4.37). Note

$$
\left|\frac{x^{\mathrm{H}}\Delta Ax}{x^{\mathrm{H}}Ax}\right| \leq \epsilon_a,
$$

we have

$$
|\varsigma(x)^2\phi(x) - \psi(x)^2| \leq 4(x^{\mathrm{H}}Ax)^2\|C\|_2^2[\epsilon_c + \epsilon_a]^2 + 4(x^{\mathrm{H}}Ax)\|B\|_2^2\|C\|_2[\epsilon_b + \epsilon_a][\epsilon_b + \epsilon_c].
$$

Using (4.38), we have $\|\Delta\Lambda_+\|_2 \leq \gamma_{\mathrm{uu}}^{+}$ to get (4.28). □

4.4 Perturbation bounds in unitarily invariant norms

Our main result of this subsection is Theorems 4.2 and 4.3. The proof of Theorem 4.2 is based on our new Wielandt-Lidskii min-max principles. Since it is rather long, we postpone it after stating both theorems.

Theorem 4.2. *Suppose $\Delta A = \Delta B = 0$ and (4.18) holds, and let*

$$\gamma = (\lambda_1^+ - \lambda_n^-)\lambda_{\min}(A), \quad \widetilde{\gamma} = (\tilde{\lambda}_1^+ - \tilde{\lambda}_n^-)\lambda_{\min}(A). \tag{4.44}$$

Then

$$\|\Delta\Lambda_\pm\|_{\rm ui} \le c \cdot \frac{\|\Delta C\|_{\rm ui}}{\min\{\gamma, \widetilde{\gamma}\}}, \tag{4.45}$$

where the constant $c = 1$ if ΔC is semidefinite and $c = 2$ in general.

The inequality (4.45) can be considered as an extension of (4.19), but a little bit less satisfying in that it does not become (4.19) after specializing the unitarily invariant norm to the spectral norm in two aspects: c is not always 1 and

$$\min_{x\neq 0} \varsigma_0(x) \ge \gamma$$

which can be a strict inequality. Thus it makes us wonder if the stronger version of (4.45) upon setting $c = 1$ always and replacing $\min\{\gamma, \widetilde{\gamma}\}$ by χ_ς holds. But how to settle this question eludes us for now.

Recall Theorem 2.5. The next theorem is a straightforward application of Theorem A.2, where $\|Z\|_2$ and $\|\widetilde{Z}\|_2$ can be bounded using item 5 of Theorem 2.5.

Theorem 4.3. *Let $\mathscr{A} - \lambda\mathscr{B} = \mathscr{L}_{\boldsymbol{Q}}(\lambda)$ and $\widetilde{\mathscr{A}} - \lambda\widetilde{\mathscr{B}} = \mathscr{L}_{\widetilde{\boldsymbol{Q}}}(\lambda)$, admitting the eigendecomposition in (2.16). Then*

$$\|\widetilde{\Lambda} - \Lambda\|_{\rm ui} \le \|Z\|_2\|\widetilde{Z}\|_2\Big(\|\widetilde{\mathscr{A}} - \mathscr{A}\|_{\rm ui} + \xi\|\widetilde{\mathscr{B}} - \mathscr{B}\|_{\rm ui}\Big), \tag{4.46}$$

where $\xi = \max\{|\lambda_{\max}^+|, |\lambda_{\max}^-|, |\tilde{\lambda}_{\max}^+|, |\tilde{\lambda}_{\max}^-|\}$, and $\lambda_{\max}^\pm$ and $\tilde{\lambda}_{\max}^\pm$ are defined by (4.16).

The rest of this subsection is devoted to the proof of Theorem 4.2.

Lemma 4.5. *Suppose $\Delta A = \Delta B = 0$ and (4.18) holds. Let $\varepsilon_1 \le \varepsilon_2 \le \cdots \le \varepsilon_n$ be the eigenvalues of ΔC, and γ and $\widetilde{\gamma}$ be given by (4.44).*

1. *Given $X \in \mathbb{C}^{n\times k}$ with $\operatorname{rank}(X) = k$, denote the quadratic eigenvalues of $X^{\rm H}\boldsymbol{Q}(\lambda)X$ by*

$$\lambda_{1,X}^- \le \cdots \le \lambda_{k,X}^- \le \lambda_{1,X}^+ \le \cdots \le \lambda_{k,X}^+,$$

and the quadratic eigenvalues of $X^{\rm H}\widetilde{\boldsymbol{Q}}(\lambda)X$ by $\tilde{\lambda}_{j,X}^\pm$ arranged in the same way. Then

$$-\sum_{i=1}^k \frac{\max\{0, -\varepsilon_1\} + \varepsilon_{n-1+i}}{\widetilde{\gamma}} \le \sum_{i=1}^k \Delta\lambda_{i,X}^+ \le -\sum_{i=1}^k \frac{\min\{0, -\varepsilon_n\} + \varepsilon_i}{\gamma}, \tag{4.47a}$$

$$\sum_{i=1}^k \frac{\min\{0, -\varepsilon_n\} + \varepsilon_i}{\gamma} \le \sum_{i=1}^k \Delta\lambda_{i,X}^- \le \sum_{i=1}^k \frac{\max\{0, -\varepsilon_1\} + \varepsilon_{n-1+i}}{\widetilde{\gamma}}. \tag{4.47b}$$

2. For any $1 \le i_1 < \cdots < i_k \le n$,

$$-\sum_{i=1}^{k} \frac{\max\{0,-\varepsilon_1\}+\varepsilon_{n+1-i}}{\widetilde{\gamma}} \le \sum_{i=1}^{k} \Delta\lambda_{i_k}^{+} \le -\sum_{i=1}^{k} \frac{\min\{0,-\varepsilon_n\}+\varepsilon_i}{\gamma}, \tag{4.48a}$$

$$\sum_{i=1}^{k} \frac{\min\{0,-\varepsilon_n\}+\varepsilon_i}{\gamma} \le \sum_{i=1}^{k} \Delta\lambda_{i_k}^{-} \le \sum_{i=1}^{k} \frac{\max\{0,-\varepsilon_1\}+\varepsilon_{n+1-i}}{\widetilde{\gamma}}, \tag{4.48b}$$

Proof. The assumption (4.18) guarantees that $\widetilde{\boldsymbol{Q}}(\lambda)$ is still hyperbolic. Without loss of generality, we may assume that X has orthonormal columns; otherwise, we consider $V^{\mathrm{H}}\boldsymbol{Q}(\lambda)V$ instead, where V is from a QR decomposition $X = VR$ of X, $V^{\mathrm{H}}V = I_k$ and $R \in \mathbb{C}^{k\times k}$. Evidently $X^{\mathrm{H}}\boldsymbol{Q}(\lambda)X$ and $V^{\mathrm{H}}\boldsymbol{Q}(\lambda)V$ have the same quadratic eigenvalues.

Recall the linearization (2.5) for $\boldsymbol{Q}(\lambda)$. We linearize

$$\boldsymbol{Q}_X(\lambda) := X^{\mathrm{H}}\boldsymbol{Q}(\lambda)X \equiv A_X\lambda^2 + B_X\lambda + C_X$$

in the same way to get

$$\mathscr{A}_X - \lambda\mathscr{B}_X \equiv \begin{bmatrix} -C_X & 0 \\ 0 & A_X \end{bmatrix} - \lambda \begin{bmatrix} B_X & A_X \\ A_X & 0 \end{bmatrix} = \mathscr{L}_{\boldsymbol{Q}_X}(\lambda).$$

Next we apply Theorem 2.5 to $\boldsymbol{Q}_X(\lambda)$ to obtain various associated eigen-decompositions and denote the corresponding quantities by the same symbols as those for $\boldsymbol{Q}(\lambda)$ but with the subscript X to indicate them for $\boldsymbol{Q}_X(\lambda)$. In particular, we will have

$$U_X = [u_{1,X}^{+}, \cdots, u_{k,X}^{+}], \quad \Lambda_{+,X} = \mathrm{diag}(\lambda_{1,X}^{+}, \lambda_{2,X}^{+}, \cdots, \lambda_{k,X}^{+}),$$

where $u_{i,X}^{+}$ are quadratic eigenvectors of $\boldsymbol{Q}_X(\lambda)$, $\varsigma_X(u_{i,X}^{+}) = 1$, and

$$S_X = \begin{bmatrix} U_X \\ U_X\Lambda_{+,X} \end{bmatrix}, \quad S_X^{\mathrm{H}}\mathscr{B}_X S_X = I_k.$$

Also $S_X^{\mathrm{H}}\widetilde{\mathscr{B}}_X S_X = I_k$ since $\widetilde{\mathscr{B}}_X = \mathscr{B}_X$. Note that $U_X \in \mathbb{C}^{k\times k}$ is nonsingular. By Theorems 2.2 and [37, Corollary 2.1],

$$\inf_{Z^{\mathrm{H}}\mathscr{B}_X Z = I_k} \mathrm{trace}(Z^{\mathrm{H}}\mathscr{A}_X Z) = \sum_{i=1}^{k} \lambda_{i,X}^{+} = \mathrm{trace}(S_X^{\mathrm{H}}\mathscr{A}_X S_X).$$

Let $\varepsilon_{1,X} \le \cdots \le \varepsilon_{k,X}$ be the eigenvalues of $\Delta C_X = X^{\mathrm{H}}\Delta C X$. Since X has orthonormal columns, we have $\varepsilon_i \le \varepsilon_{i,X} \le \varepsilon_{n-k+i}$ by the Cauchy interlacing theorem, and thus

$$\sum_{i=1}^{k} \varepsilon_i \le \sum_{i=1}^{k} \varepsilon_{i,X} \le \sum_{i=1}^{k} \varepsilon_{n+1-i}.$$

For the sake of presentation, we will drop the superscript "+" to $u_{i,X}^{+}$ in the rest of this proof. We have

$$\sum_{i=1}^{k} \tilde{\lambda}_{i,X}^{+} = \inf_{Z^{\mathrm{H}}\widetilde{\mathscr{B}}_X Z = I_k} \mathrm{trace}(Z^{\mathrm{H}}\widetilde{\mathscr{A}_X} Z)$$

$$
\begin{aligned}
&\le \operatorname{trace}(S_X^{\mathrm{H}}\widetilde{\mathscr{A}}_X S_X) \qquad (\text{since } S_X^{\mathrm{H}}\widetilde{\mathscr{B}}_X S_X = I_k)\\
&= \operatorname{trace}(S_X^{\mathrm{H}}\mathscr{A}_X S_X) + \operatorname{trace}(S_X^{\mathrm{H}}\Delta\mathscr{A}_X S_X)\\
&= \sum_{i=1}^{k}\lambda_{i,X}^{+} - \operatorname{trace}(U_X^{\mathrm{H}}\Delta C_X U_X). \qquad (4.49)
\end{aligned}
$$

Let $\mu = \min\{0, -\varepsilon_n\} \le 0$. For any scalar $\tau_0 \in (0,1)$, set $\tau^2 = \tau_0^2\gamma = \tau_0^2(\lambda_1^+ - \lambda_n^-)\lambda_{\min}(A)$, and

$$
\begin{aligned}
&E_X = -\mu U_X^{\mathrm{H}} U_X, && D_X = U_X^{\mathrm{H}}(U_X^{-\mathrm{H}} U_X^{-1} - \tau^2 I)U_X,\\
&\mathscr{C}_X = \begin{bmatrix} \tau^{-2}(\Delta C_X + \mu I) & 0\\ 0 & E_X\end{bmatrix} \in \mathbb{C}^{2k\times 2k}, && \mathscr{D}_X = \begin{bmatrix} I & 0\\ 0 & D_X\end{bmatrix} \in \mathbb{C}^{2k\times 2k}.
\end{aligned}
$$

Note that by (2.18a), (2.18e), and (2.24),

$$
U_X^{\mathrm{H}} A_X U_X \preceq (\lambda_{1,X}^+ - \lambda_{k,X}^-)^{-1} I \preceq (\lambda_1^+ - \lambda_n^-)^{-1} I
$$

which infers

$$
U_X^{-\mathrm{H}} U_X^{-1} \succeq (\lambda_1^+ - \lambda_n^-)A_X \succeq (\lambda_1^+ - \lambda_n^-)\lambda_{\min}(A_X) I \succeq (\lambda_1^+ - \lambda_n^-)\lambda_{\min}(A) I = \gamma I \succ \tau^2 I.
$$

Thus, $D_X \succ 0$, and so $\mathscr{D}_X \succ 0$. Hence the matrix pencil $\mathscr{C}_X - \lambda\mathscr{D}_X$ has $2k$ finite eigenvalues ν_i ($i = 1, \cdots, 2k$). By the choice of μ, $\Delta C_X + \mu I \preceq 0$ and $E_X \succeq 0$. Therefore these ν_i can be ordered as

$$
\nu_1 \le \cdots \le \nu_k \le 0 \le \nu_{k+1} \le \cdots \le \nu_{2k},
$$

where ν_i for $i = 1, \cdots, k$ are the eigenvalues of $\tau^{-2}(\Delta C_X + \mu I)$ and ν_i for $i = k+1, \cdots, 2k$ are the generalized eigenvalues of $E_X - \lambda D_X$. By the Courant-Fischer min-max principle, we have for $i = 1, \cdots, k$

$$
\begin{aligned}
\nu_i &= \min_{\dim \mathcal{X} = i}\ \max_{0\ne x\in\mathcal{X}} \frac{x^{\mathrm{H}}(\Delta C_X + \mu I)x}{\tau^2 x^{\mathrm{H}} x}\\
&= \frac{1}{\tau^2}\left[\mu + \min_{\dim \mathcal{X} = i}\ \max_{0\ne x\in\mathcal{X}} \frac{x^{\mathrm{H}}\Delta C_X x}{x^{\mathrm{H}} x}\right]\\
&= \frac{1}{\tau^2}\left[\mu + \varepsilon_{i,X}\right]\\
&\ge \frac{1}{\tau^2}\left[\mu + \varepsilon_i\right]\\
&= \frac{1}{\tau_0^2\gamma}\left[\mu + \varepsilon_i\right].
\end{aligned}
$$

By the arbitrary choice of $\tau_0 \in (0,1)$,

$$
\nu_i \ge \frac{\mu + \varepsilon_i}{\gamma}.
$$

For the matrix $T_X := \begin{bmatrix} \tau U_X\\ I\end{bmatrix}$, we have

$$
T_X^{\mathrm{H}}\mathscr{D}_X T_X = \tau^2 U_X^{\mathrm{H}} U_X + D_X = I,
$$

$$T_X^{\mathrm{H}}\mathscr{C}_X T_X = \tau^2\tau^{-2}U_X^{\mathrm{H}}(\varDelta C_X + \mu I)U_X + E_X = U_X^{\mathrm{H}}\varDelta C_X U_X.$$

Therefore

$$\begin{aligned}\operatorname{trace}(U_X^{\mathrm{H}}\varDelta C_X U_X) &= \operatorname{trace}(T_X^{\mathrm{H}}\mathscr{C}_X T_X)\\ &\geq \min_{Z^{\mathrm{H}}\mathscr{D}_X Z = I}\operatorname{trace}(Z^{\mathrm{H}}\mathscr{C}_X Z)\\ &= \sum_{i=1}^k \nu_i.\end{aligned}$$

Thus, (4.49) becomes

$$\sum_{i=1}^k \varDelta\lambda_{i,X}^+ \leq -\sum_{i=1}^k \nu_i \leq -\sum_{i=1}^k \frac{\mu+\varepsilon_i}{\gamma} = -\sum_{i=1}^k \frac{\min\{0,-\varepsilon_n\}+\varepsilon_i}{\gamma}. \tag{4.50}$$

Think of $\boldsymbol{Q}$ as obtained from perturbing $\widetilde{\boldsymbol{Q}}$ and apply (4.50) to get

$$-\sum_{i=1}^k \varDelta\lambda_{i,X}^+ \leq -\sum_{i=1}^k \frac{\min\{0,-(-\varepsilon_1)\}+(-\varepsilon_{n-1+i})}{\widetilde{\gamma}} \tag{4.51}$$

which, combined with (4.50), leads to (4.47a). Apply (4.47a) to $\boldsymbol{Q}(-\lambda)$ and $\widetilde{\boldsymbol{Q}}(-\lambda)$ to get (4.47b).

Now we prove (4.48). With all "sup" being taken over $\mathcal{X}_1 \subset \cdots \subset \mathcal{X}_k$ and $\operatorname{codim}\mathcal{X}_j = i_j - 1$, and all "inf" over $x_j \in \mathcal{X}_j$, $X = [x_1, \ldots, x_k]$, and $\operatorname{rank}(X) = k$, we have by Theorem 3.3

$$\begin{aligned}\sum_{j=1}^k \tilde{\lambda}_{i_k}^+ &= \sup\inf\sum_{j=1}^k \tilde{\lambda}_{k,X}^+\\ &\leq \sup\inf\left[\sum_{j=1}^k \lambda_{k,X}^+ - \sum_{i=1}^k \frac{\min\{0,-\varepsilon_n\}+\varepsilon_i}{\gamma}\right] \qquad \text{(by (4.50))}\\ &= \sup\inf\sum_{j=1}^k \lambda_{k,X}^+ - \sum_{i=1}^k \frac{\min\{0,-\varepsilon_n\}+\varepsilon_i}{\gamma}\\ &\leq \sum_{j=1}^k \lambda_{i_k}^+ - \sum_{i=1}^k \frac{\min\{0,-\varepsilon_n\}+\varepsilon_i}{\gamma}.\end{aligned} \tag{4.52}$$

Similarly,

$$\sum_{j=1}^k \lambda_{i_k}^+ \leq \sum_{j=1}^k \tilde{\lambda}_{i_k}^+ - \sum_{i=1}^k \frac{\min\{0,-(-\varepsilon_1)\}+(-\varepsilon_{n-1+i})}{\widetilde{\gamma}}. \tag{4.53}$$

The inequalities in (4.48a) is a consequence of (4.52) and (4.53). Apply (4.48a) to $\boldsymbol{Q}(-\lambda)$ and $\widetilde{\boldsymbol{Q}}(-\lambda)$ to get (4.48b). □

Lemma 4.6. *Suppose $\Delta A = \Delta B = 0$ and (4.18) holds. We have for $1 \le j \le n$*

$$\tilde{\lambda}_j^+ \le \lambda_j^+ \text{ and } \tilde{\lambda}_j^- \ge \lambda_j^- \text{ if } \Delta C \succeq 0, \tag{4.54a}$$
$$\tilde{\lambda}_j^+ \ge \lambda_j^+ \text{ and } \tilde{\lambda}_j^- \le \lambda_j^- \text{ if } \Delta C \preceq 0. \tag{4.54b}$$

Consequently $\widetilde{\gamma} \le \gamma$ if $\Delta C \succeq 0$, and $\widetilde{\gamma} \ge \gamma$ if $\Delta C \preceq 0$.

Proof. The assumption (4.18) guarantees that $\widetilde{\boldsymbol{Q}}(\lambda)$ is still hyperbolic. By (3.2), we see

$$\tilde{\rho}_+(x) \le \rho_+(x) \text{ and } \tilde{\rho}_-(x) \ge \tilde{\rho}_-(x) \text{ if } \Delta C \succeq 0,$$
$$\tilde{\rho}_+(x) \ge \rho_+(x) \text{ and } \tilde{\rho}_-(x) \le \tilde{\rho}_-(x) \text{ if } \Delta C \preceq 0.$$

Now use Theorem 3.2 to get (4.54). □

Proof of Theorem 4.2. The assumption (4.18) guarantees that $\widetilde{\boldsymbol{Q}}(\lambda)$ is still hyperbolic.

As in Lemma 4.5, let $\varepsilon_1 \le \varepsilon_2 \le \cdots \le \varepsilon_n$ be the eigenvalues of ΔC.

Consider first the case $\Delta C \succeq 0$. Then $0 \le \varepsilon_1$. Also $\Delta\lambda_i^+ \le 0$ for all i by Lemma 4.6. Therefore the leftmost inequality in (4.48a) gives

$$\sum_{i=1}^k |\Delta\lambda_{i_k}^+| \le \sum_{i=1}^k \frac{\varepsilon_{n+1-i}}{\widetilde{\gamma}}$$

for any $1 \le i_1 < \cdots < i_k \le n$. As a result of [56, Theorem II.3.6 and Theorem II.3.17], we have

$$\|\Delta\Lambda_+\|_{\rm ui} \le \frac{\|\Delta C\|_{\rm ui}}{\widetilde{\gamma}}. \tag{4.55}$$

Similarly, use the rightmost inequality in (4.48b) to get

$$\|\Delta\Lambda_-\|_{\rm ui} \le \frac{\|\Delta C\|_{\rm ui}}{\widetilde{\gamma}}. \tag{4.56}$$

Now we turn to the case $\Delta C \preceq 0$. Then $\varepsilon_n \le 0$. Also $\Delta\lambda_i^+ \ge 0$ for all i by Lemma 4.6. Therefore the rightmost inequality in (4.48a) gives

$$\sum_{i=1}^k |\Delta\lambda_{i_k}^+| \le \sum_{i=1}^k \frac{|\varepsilon_i|}{\gamma}$$

for any $1 \le i_1 < \cdots < i_k \le n$. Again as a result of [56, Theorem II.3.6 and Theorem II.3.17], we have

$$\|\Delta\Lambda_+\|_{\rm ui} \le \frac{\|\Delta C\|_{\rm ui}}{\gamma}. \tag{4.57}$$

Similarly, use the leftmost inequality in (4.48b) to get

$$\|\Delta\Lambda_-\|_{\rm ui} \le \frac{\|\Delta C\|_{\rm ui}}{\gamma}. \tag{4.58}$$

The inequalities (4.55) – (4.56) together give (4.45) for the case when ΔC is semidefinite.

For the general case when ΔC is indefinite, we can decompose $\Delta C = \Delta C_+ - \Delta C_-$, where $\Delta C_\pm \succeq 0$ and

$$\operatorname{eig}(\Delta C_+) = \{\max\{0, \varepsilon_i\}, 1 \le i \le n\}, \quad \operatorname{eig}(\Delta C_-) = \{\max\{0, -\varepsilon_i\}, 1 \le i \le n\}.$$

In particular, $\|\Delta C_\pm\|_{\rm ui} \le \|\Delta C\|_{\rm ui}$. Let $\widehat{C} = C - \Delta C_-$ and $\widehat{\boldsymbol{Q}}(\lambda) = \lambda^2 A + \lambda B + \widehat{C}$. We claim $\widehat{\boldsymbol{Q}}(\lambda)$ is hyperbolic. This is because $\widetilde{C} = C + \Delta C_+ - \Delta C_- \succeq C - \Delta C_- = \widehat{C}$ and thus for any $x \ne 0$

$$0 < (x^{\rm H} Bx)^2 - 4(x^{\rm H} Ax)(x^{\rm H} \widetilde{C} x) \le (x^{\rm H} Bx)^2 - 4(x^{\rm H} Ax)(x^{\rm H} \widehat{C} x),$$

where the first inequality holds because $\widetilde{\boldsymbol{Q}}(\lambda)$ is hyperbolic. Apply what we just proved to $\boldsymbol{Q}$ and $\widehat{\boldsymbol{Q}}$ to get

$$\|\widehat{\Lambda}_\pm - \Lambda_\pm\|_{\rm ui} \le \frac{\|\Delta C_-\|_{\rm ui}}{\gamma} \le \frac{\|\Delta C\|_{\rm ui}}{\gamma}, \tag{4.59}$$

where $\widehat{\Lambda}_\pm$ are similarly defined for $\widehat{\boldsymbol{Q}}$ to $\Lambda_\pm$ for $\boldsymbol{Q}$. Notice $\widetilde{C} = \widehat{C} + \Delta C_+$ and apply what we just proved to $\boldsymbol{Q}$ and $\widehat{\boldsymbol{Q}}$ to get

$$\|\widetilde{\Lambda}_\pm - \widehat{\Lambda}_\pm\|_{\rm ui} \le \frac{\|\Delta C_+\|_{\rm ui}}{\widetilde{\gamma}} \le \frac{\|\Delta C\|_{\rm ui}}{\widetilde{\gamma}}. \tag{4.60}$$

Finally

$$\|\widetilde{\Lambda}_\pm - \Lambda_\pm\|_{\rm ui} \le \|\widetilde{\Lambda}_\pm - \widehat{\Lambda}_\pm\|_{\rm ui} + \|\widehat{\Lambda}_\pm - \Lambda_\pm\|_{\rm ui} \le 2 \cdot \frac{\|\Delta C\|_{\rm ui}}{\min\{\gamma, \widetilde{\gamma}\}},$$

as was to be shown. $\square$

4.5 Perturbation bounds in the Frobenius norms

Theorem 4.4. *Suppose* (4.26) *holds and* $\lambda_0 \in (\lambda_n^-, \lambda_1^+) \cap (\tilde{\lambda}_n^-, \tilde{\lambda}_1^+)$. *Then*

$$\|\Delta\Lambda\|_{\rm F}^2 \le 2\Big(\chi_1^2\zeta_1^2\chi_2^4\|\Delta A\|_{\rm F}^2 + \chi_2^2\|\Delta B_{\lambda_0}\|_{\rm F}^2 + \chi_3^2\zeta_2^2\|\Delta C_{\lambda_0}\|_{\rm F}^2\Big), \tag{4.61}$$

where

$$\begin{aligned}
\chi_1 &= \sqrt{\|C_{\lambda_0}\|_2 + \|\widetilde{C}_{\lambda_0}\|_2 + \Big(\|\widetilde{A}^{-1/2}\widetilde{B}_{\lambda_0}\|_2 + \|A^{-1/2}B_{\lambda_0}\|_2\Big)^2},\\
\chi_2 &= \sqrt{\|A^{-1}\|_2\|\widetilde{A}^{-1}\|_2},\\
\chi_3 &= \sqrt{\|A^{-1}\|_2 + \|\widetilde{A}^{-1}\|_2},\\
\zeta_1 &= \frac{1}{\|A\|_2^{-1/2} + \|\widetilde{A}\|_2^{-1/2}},\\
\zeta_2 &= \frac{1}{\|C_{\lambda_0}^{-1}\|_2^{-1/2} + \|\widetilde{C}_{\lambda_0}^{-1}\|_2^{-1/2}}.
\end{aligned}$$

In particular, if $\Delta A = 0$, *then the scalar* 2 *in* (4.61) *can be replaced by* 1 *to give*

$$\|\Delta\Lambda\|_{\rm F}^2 \le \|A^{-1}\|_2^2\|\Delta B\|_{\rm F}^2 + 2\|A^{-1}\|_2\zeta_2^2\|\lambda_0\Delta B + \Delta C\|_{\rm F}^2. \tag{4.62}$$

Proof. The assumptions in (4.26) guarantee that $\widetilde{\boldsymbol{Q}}(\lambda)$ is still hyperbolic. The assumption $\lambda_0 \in (\lambda_n^-, \lambda_1^+) \cap (\tilde{\lambda}_n^-, \tilde{\lambda}_1^+)$ ensures $\widetilde{C}_{\lambda_0} \succeq 0$ and $C_{\lambda_0} \succeq 0$. Without loss of generality, we may assumed that both $\boldsymbol{Q}(\lambda)$ and $\widetilde{\boldsymbol{Q}}(\lambda)$ have already been shifted, or equivalently $\lambda_0 = 0$. This allows to drop the potential subscript "λ_0" to B_{λ_0}, C_{λ_0}, etc.

Note that $\mathscr{K}_{\boldsymbol{Q}}(\lambda) = \mathscr{A} - \lambda\mathscr{B}$ as in (2.6), where $\mathscr{B} = \operatorname{diag}(-C, A) \succ 0$. Thus, the hyperbolic eigenvalue problem $\boldsymbol{Q}(\lambda)$ is equivalent to the Hermitian eigenvalue problem $\boldsymbol{K}$ where

$$\boldsymbol{K} = \mathscr{B}^{-1/2}\mathscr{A}\mathscr{B}^{-1/2} = \begin{bmatrix} 0 & [-C]^{1/2}A^{-1/2} \\ A^{-1/2}[-C]^{1/2} & -A^{-1/2}BA^{-1/2} \end{bmatrix}.$$

By the Hoffman-Wielandt therorem [27, 56],

$$\|\varDelta\Lambda\|_{\mathrm{F}}^2 \le \|\varDelta\boldsymbol{K}\|_{\mathrm{F}}^2 = 2\left\|\varDelta\left([-C]^{1/2}A^{-1/2}\right)\right\|_{\mathrm{F}}^2 + \left\|\varDelta\left(A^{-1/2}BA^{-1/2}\right)\right\|_{\mathrm{F}}^2. \tag{4.63}$$

The rest of the proof is just to bound the two terms in the right-hand side of (4.63). To this end, we note that

$$\begin{aligned}\left\|\varDelta\left([-C]^{1/2}A^{-1/2}\right)\right\|_{\mathrm{F}} &= \left\|\big[-\widetilde{C}\,\big]^{1/2}\varDelta(A^{-1/2}) + \varDelta([-C]^{1/2})A^{-1/2}\right\|_{\mathrm{F}} \\ &\le \|\big[-\widetilde{C}\,\big]^{1/2}\|_2\|\varDelta(A^{-1/2})\|_{\mathrm{F}} + \|A^{-1/2}\|_2\left\|\varDelta\Big([-C]^{1/2}\Big)\right\|_{\mathrm{F}},\end{aligned} \tag{4.64}$$

and similarly

$$\begin{aligned}\left\|\varDelta\left([-C]^{1/2}A^{-1/2}\right)\right\|_{\mathrm{F}} &= \left\|\varDelta([-C]^{1/2})\widetilde{A}^{-1/2} + [-C]^{1/2}\varDelta(A^{-1/2})\right\|_{\mathrm{F}} \\ &\le \|\widetilde{A}^{-1/2}\|_2\left\|\varDelta\Big([-C]^{1/2}\Big)\right\|_{\mathrm{F}} + \|[-C]^{1/2}\|_2\left\|\varDelta(A^{-1/2})\right\|_{\mathrm{F}}.\end{aligned} \tag{4.65}$$

Also,

$$\begin{aligned}\left\|\varDelta\left(A^{-1/2}BA^{-1/2}\right)\right\|_{\mathrm{F}} &= \left\|\widetilde{A}^{-1/2}\widetilde{B}\varDelta(A^{-1/2}) + \widetilde{A}^{-1/2}\varDelta BA^{-1/2} + \varDelta(A^{-1/2})BA^{-1/2}\right\|_{\mathrm{F}} \\ &\le \Big(\|\widetilde{A}^{-1/2}\widetilde{B}\|_2 + \|A^{-1/2}B\|_2\Big)\|\varDelta(A^{-1/2})\|_{\mathrm{F}} \\ &\quad + \|\widetilde{A}^{-1/2}\|_2\|A^{-1/2}\|_2\|\varDelta B\|_{\mathrm{F}}.\end{aligned} \tag{4.66}$$

Combine[8] (4.63) – (4.66) to get

$$\|\varDelta\Lambda\|_{\mathrm{F}}^2 \le 2\left[\chi_1^2\|\varDelta(A^{-1/2})\|_{\mathrm{F}}^2 + \chi_2^2\|\varDelta B\|_{\mathrm{F}}^2 + \chi_3^2\left\|\varDelta\Big([-C]^{1/2}\Big)\right\|_{\mathrm{F}}^2\right]. \tag{4.67}$$

By [53],

$$\begin{aligned}\|\varDelta(A^{-1/2})\|_{\mathrm{F}} &\le \zeta_1\|\varDelta(A^{-1})\|_{\mathrm{F}} \\ &\le \zeta_1\|\widetilde{A}^{-1}\|_2\|\varDelta A\|_{\mathrm{F}}\|\widetilde{A}^{-1}\|_2,\end{aligned} \tag{4.68}$$

$$\left\|\varDelta\Big([-C]^{1/2}\Big)\right\|_{\mathrm{F}} \le \zeta_2\,\|\varDelta C\|_{\mathrm{F}}, \tag{4.69}$$

where the inequality sign in (4.68) is due to $\varDelta(A^{-1}) = -\widetilde{A}^{-1}\varDelta AA^{-1}$. Now substitute (4.68) and (4.69) into (4.67) to yield the desired inequality. □

[8] Actually we only use this: $(a+b)^2 \le 2(a^2+b^2)$ which results in the scalar 2 in (4.67).

Theorem 4.4 gives a perturbation result for all quadratic eigenvalues of $\boldsymbol{Q}(\lambda)$. However, using a different approach, we can obtain results in the Frobenius for only pos- or neg-type quadratic eigenvalues of $\boldsymbol{Q}(\lambda)$.

Following [20], we know the matrix equation

$$AX^2 + BX + C = 0$$

has two special solutions. One has all pos-type quadratic eigenvalues of $\boldsymbol{Q}(\lambda)$ as its eigenvalues while the other has all neg-type quadratic eigenvalues of $\boldsymbol{Q}(\lambda)$ as its eigenvalues. We call the first special solution the *pos-type* solution and the second special solution the *neg-type* solution.

Consider $\boldsymbol{Q}_{\lambda_0}(\lambda)$ and set

$$B_A = A^{-1/2} B_{\lambda_0} A^{-1/2}, \quad C_A = A^{-1/2} C_{\lambda_0} A^{-1/2}. \tag{4.70}$$

Because $A^{-1/2}\boldsymbol{Q}_{\lambda_0}(\lambda)A^{-1/2} = \lambda^2 I + \lambda B_A + C_A$ is hyperbolic, the following equation

$$X^2 + B_A X + C_A = 0, \tag{4.71}$$

has the pos- and neg-type solutions. Denote them by $R_\pm$, respectively, in the rest of this section. Both $R_\pm$ can be expressed explicitly by the quantities defined in Theorem 2.5. In fact,

$$R_\pm := A^{1/2} U_\pm (\Lambda_\pm - \lambda_0 I) U_\pm^{-1} A^{-1/2}. \tag{4.72}$$

Lemma 4.7. *Suppose* (4.26) *holds and* $\lambda_0 \ge \max\{\lambda_n^+, \tilde{\lambda}_n^+\}$*. Let* $\mathrm{typ} \in \{+,-\}$*. If*

$$\eta := 2\lambda_0 - \tilde{\lambda}_n^+ - \tilde{\lambda}_n^- - \|\widetilde{R}_{\mathrm{typ}}\|_2 - \|R_{\mathrm{typ}}\|_2 > 0, \tag{4.73}$$

then

$$\|\Delta R_{\mathrm{typ}}\|_{\mathrm{F}} \le \frac{\chi_4 \zeta_1 \chi_2^2}{\eta} \|\Delta A\|_{\mathrm{F}} + \frac{\chi_2}{\eta} \left(\|R_{\mathrm{typ}}\|_2 \|\Delta B_{\lambda_0}\|_{\mathrm{F}} + \|\Delta C_{\lambda_0}\|_{\mathrm{F}} \right), \tag{4.74}$$

where

$$\chi_4 = \|R_{\mathrm{typ}}\|_2 (\|\widetilde{A}^{-1/2} \widetilde{B}_{\lambda_0}\|_2 + \|A^{-1/2} B_{\lambda_0}\|_2) + \|\widetilde{A}^{-1/2} \widetilde{C}_{\lambda_0}\|_2 + \|A^{-1/2} C_{\lambda_0}\|_2,$$

and χ_2, ζ_1 *are as in* Theorem 4.4.

Proof. The assumptions in (4.26) guarantee that $\widetilde{\boldsymbol{Q}}(\lambda)$ is still hyperbolic. By (4.8) and (4.9),

$$\operatorname{eig}(B_A) \in [2\lambda_0 - \lambda_n^- - \lambda_n^+, 2\lambda_0 - \lambda_1^- - \lambda_1^+], \tag{4.75}$$
$$\operatorname{eig}(C_A) \in [(\lambda_0 - \lambda_n^-)(\lambda_0 - \lambda_n^+), (\lambda_0 - \lambda_1^-)(\lambda_0 - \lambda_1^+)]. \tag{4.76}$$

Subtract $\widetilde{R}_{\mathrm{typ}}^2 + \widetilde{B}_A \widetilde{R}_{\mathrm{typ}} + \widetilde{C}_A = 0$ from $R_{\mathrm{typ}}^2 + B_A R_{\mathrm{typ}} + C_A = 0$ to get

$$(\widetilde{R}_{\mathrm{typ}} + \widetilde{B}_A)\Delta R_{\mathrm{typ}} + (\Delta R_{\mathrm{typ}}) R_{\mathrm{typ}} = -(\Delta B_A) R_{\mathrm{typ}} - \Delta C_A,$$

or equivalently

$$\left[I \otimes (\widetilde{R}_{\mathrm{typ}} + \widetilde{B}_A) + R_{\mathrm{typ}}^{\mathrm{T}} \otimes I \right] \operatorname{vec}(\Delta R_{\mathrm{typ}}) = -\operatorname{vec}\left((\Delta B_A) R_{\mathrm{typ}} - \Delta C_A \right), \tag{4.77}$$

where vec(·) turns a matrix to a vector by appending the columns of the matrix one after another with the first column followed by the second column and so on. The equation (4.77) yields

$$\begin{aligned}\|\Delta R_{\rm typ}\|_{\rm F} &\le \left\|\left[I\otimes(\widetilde{R}_{\rm typ}+\widetilde{B}_A)+R_{\rm typ}^{\rm T}\otimes I\right]^{-1}\right\|_2 \|(\Delta B_A)R_{\rm typ}-\Delta C_A\|_{\rm F}\\ &\le \left\|\left[I\otimes(\widetilde{R}_{\rm typ}+\widetilde{B}_A)+R_{\rm typ}^{\rm T}\otimes I\right]^{-1}\right\|_2 (\|R_{\rm typ}\|_2\|\Delta B_A\|_{\rm F}+\|\Delta C_A\|_{\rm F}). \quad (4.78)\end{aligned}$$

Choose a $\tau \le \tilde{\lambda}_1^+ + \tilde{\lambda}_1^- - 2\lambda_0 \le \tilde{\lambda}_n^+ + \tilde{\lambda}_n^- - 2\lambda_0 = -\eta - \|\widetilde{R}_{\rm typ}\|_2 - \|R_{\rm typ}\|_2 < 0$. Then

$$\begin{aligned}&\|I\otimes\widetilde{R}_{\rm typ}+I\otimes(\widetilde{B}_A+\tau I)+R_{\rm typ}^{\rm T}\otimes I\|_2\\ &\qquad\le \|\widetilde{R}_{\rm typ}\|_2+\|R_{\rm typ}^{\rm T}\|_2+\|\widetilde{B}_A+\tau I\|_2\\ &\qquad\le \|\widetilde{R}_{\rm typ}\|_2+\|R_{\rm typ}\|_2+\tilde{\lambda}_n^+ + \tilde{\lambda}_n^- - 2\lambda_0-\tau\\ &\qquad< -\eta-\tau < -\tau = |\tau|\end{aligned}$$

from which we infer

$$\begin{aligned}&\left\|\left(I\otimes(\widetilde{R}_{\rm typ}+\widetilde{B}_A)+R_{\rm typ}^{\rm T}\otimes I\right)^{-1}\right\|_2\\ &\qquad= \left\|\tau^{-1}\left(\frac{I\otimes\widetilde{R}_{\rm typ}+I\otimes(\widetilde{B}_A+\tau I)+R_{\rm typ}^{\rm T}\otimes I}{\tau}-I\otimes I\right)^{-1}\right\|_2\\ &\qquad\le \frac{|\tau|^{-1}}{1-|\tau|^{-1}\|I\otimes\widetilde{R}_{\rm typ}+I\otimes(\widetilde{B}_A+\tau I)+R_{\rm typ}^{\rm T}\otimes I\|_2}\\ &\qquad= \frac{1}{-\tau-\|I\otimes\widetilde{R}_{\rm typ}+I\otimes(\widetilde{B}_A+\tau I)+R_{\rm typ}^{\rm T}\otimes I\|_2}\\ &\qquad\le \frac{1}{-\tau-(-\eta-\tau)}=\frac{1}{\eta}. \quad (4.79)\end{aligned}$$

Like (4.66), (4.68) and (4.69), we can obtain the estimates of $\|\Delta B_A\|_{\rm F}$ and $\|\Delta C_A\|_{\rm F}$. Then (4.74) follows. □

$R_{\rm typ}$ is diagonalizable by (4.72). By [7, Theorem 3.1], we have

$$\|\Delta\Lambda_{\rm typ}\|_{\rm F} \le \kappa\|\Delta R_{\rm typ}\|_{\rm F}, \quad (4.80)$$

where

$$\kappa = \sqrt{\kappa_2(A^{1/2}U_{\rm typ})\kappa_2(\widetilde{A}^{1/2}\widetilde{U}_{\rm typ})}. \quad (4.81)$$

Theorem 4.5. *Suppose $\lambda_0 \ge \max\{\lambda_n^+, \tilde{\lambda}_n^+\}$. If (4.73) holds, then*

$$\|\Delta\Lambda_{\rm typ}\|_{\rm F} \le \frac{\kappa\chi_4\zeta_1\chi_2^2}{\eta}\|\Delta A\|_{\rm F} + \frac{\kappa\chi_2}{\eta}(\|R_{\rm typ}\|_2\|\Delta B_{\lambda_0}\|_{\rm F}+\|\Delta C_{\lambda_0}\|_{\rm F}) \quad (4.82)$$

where $\kappa, \eta, \chi_4, \chi_2, \zeta_1$ are as in Lemma 4.7 *and* (4.81).

5 Best approximations from a subspace and Rayleigh-Ritz procedure

Two most important aspects in solving a large scale eigenvalue problem are

1. building subspaces to which the desired eigenvectors (or invariant subspaces) are close, and

2. seeking "*best possible*" approximations from the suitably built subspaces.

In this section, we shall address the second aspect for our current problem at hand, i.e., seeking "*best possible*" approximations to a few quadratic eigenvalues of $\boldsymbol{Q}(\lambda)$ and their associated quadratic eigenvectors from a given subspace of $\mathbb{C}^n$. We leave the first aspect to the later sections when we present our computational algorithms.

The concept of "*best possible*" comes with a quantitative measure as to what constitutes "*best possible*". There may not be such a measure in general. In [47, section 11.4], Parlett uses three different ways to justify the use of the Rayleigh-Ritz procedure for the symmetric eigenvalue problem. For the HQEP here, each of the minimization principles in section 3 provides a quantitative measure.

Let $\boldsymbol{Q}(\lambda) = \lambda^2 A + \lambda B + C \in \mathbb{C}^{n\times n}$ be a hyperbolic quadratic matrix polynomial, and let $\mathcal{Y} \subset \mathbb{C}^n$ be a subspace of dimension m. We are seeking "*best possible*" approximations to a few quadratic eigenvalues of $\boldsymbol{Q}(\lambda)$ using $\mathcal{Y}$. Let $Y \in \mathbb{C}^{n\times m}$ be a basis matrix of $\mathcal{Y}$.

According to (3.7a) which says (upon substituting $i = n - j + 1$)

$$\lambda^+_{n-j+1} = \max_{\substack{\mathcal{X}\subseteq\mathbb{C}^n\\ \dim\mathcal{X}=j}} \min_{\substack{x\in\mathcal{X}\\ x\neq 0}} \rho_+(x), \tag{3.7a$'$}$$

it is natural to approximate λ^+_{n-j+1}, given $\mathcal{Y} \subset \mathbb{C}^n$, by

$$\mu^+_{m-j+1} := \max_{\substack{\mathcal{X}\subseteq\mathcal{Y}\\ \dim\mathcal{X}=j}} \min_{\substack{x\in\mathcal{X}\\ x\neq 0}} \rho_+(x), \tag{5.1}$$

via replacing $\mathcal{X} \subseteq \mathbb{C}^n$ in (3.7a$'$) by $\mathcal{X} \subseteq \mathcal{Y}$. Any $x \in \mathcal{X} \subseteq \mathcal{Y}$ can be written as $x = Yy$ for some $y \in \mathbb{C}^m$, and thus

$$\rho_+(x) = \rho_+(Yy) = \frac{-(y^{\mathrm{H}}Y^{\mathrm{H}}BYy) + \left[(y^{\mathrm{H}}Y^{\mathrm{H}}BYy)^2 - 4(y^{\mathrm{H}}Y^{\mathrm{H}}AYy)(y^{\mathrm{H}}Y^{\mathrm{H}}CYy)\right]^{1/2}}{2(y^{\mathrm{H}}Y^{\mathrm{H}}AYy)}.$$

Combined with (3.7a$'$) and this expression for $\rho_+(x)$, (5.1) implies that $\mu^+_1, \dots, \mu^+_m$ are the m pos-type quadratic eigenvalues of $Y^{\mathrm{H}}\boldsymbol{Q}(\lambda)Y$. What this means is that μ^+_j for $1 \le j \le m$ provide the best approximations to the m largest λ^+_j, given $\mathcal{Y}$, in the sense of (3.7a). Of course, some approximations $\mu^+_j \approx \lambda^+_{n-m+j}$ are more accurate than others.

Similarly, given $\mathcal{Y}$, μ^+_j for $1 \le j \le m$ provide the best approximations to the m smallest λ^+_j in the sense of (3.7b).

Let $\mu^-_1, \dots, \mu^-_m$ are the m neg-type quadratic eigenvalues of $Y^{\mathrm{H}}\boldsymbol{Q}(\lambda)Y$. The same argument shows, given $\mathcal{Y}$, μ^-_j for $1 \le j \le m$ provide the best approximations to the m largest λ^-_j in the sense of (3.7c), and the best approximations to the m smallest λ^-_j in the sense of (3.7d).

Algorithm 5.1 Rayleigh-Ritz procedure

Given $Y \in \mathbb{C}^{n\times m}$ which is a basis matrix of $\mathcal{Y} \subset \mathbb{C}^n$, this algorithm returns approximations to k extreme quadratic eigenpairs (of pos- or neg-type) of $\boldsymbol{Q}(\lambda)$.

1: solve the QEP for $Y^{\mathrm{H}}\boldsymbol{Q}(\lambda)Y$ to get its quadratic eigenvalues $\mu_j^{\pm}$ and associated quadratic eigenvectors $y_j^{\pm}$.

2: **return**

- $(\mu_i^{\pm}, Yy_i^{\pm})$ for $1 \le i \le k$ as approximations to $(\lambda_i^{\pm}, u_i^{\pm})$ for $1 \le i \le k$, or
- $(\mu_i^{\pm}, Yy_i^{\pm})$ for $m-k+1 \le i \le m$ as approximations to $(\lambda_i^{\pm}, u_i^{\pm})$ for $n-k+1 \le i \le n$,

depending on what kind of extreme quadratic eigenpairs are desired.

In summary, we have justified that the quadratic eigenvalues of $Y^{\mathrm{H}}\boldsymbol{Q}(\lambda)Y$ yield the best approximations to some of the largest or smallest pos- or neg-type quadratic eigenvalues of $\boldsymbol{Q}(\lambda)$ in certain respective senses. This statement could sound confusing: how could the same set of values be the best approximations to some of both the largest and smallest eigenvalues at the same time? But we point out this is not what the statement is saying. The key to understand the subtlety is not to forget that they provide the best approximations under the mentioned senses, and being the best approximations (under a particular sense) does not necessarily imply that the approximates are good, just that they are the best (under that particular sense). In practice, $\mathcal{Y}$ is built to approximate either the largest or smallest eigenvalues well, but unlikely both.

Theorems 3.3, 3.4, and 3.5, generalizing Amir-Moéz's min-max principles and the Wielandt-Lidskii min-max principles, can also be used to justify that the quadratic eigenvalues of $Y^{\mathrm{H}}\boldsymbol{Q}(\lambda)Y$ are candidates for best approximating the largest or smallest pos- or neg-type quadratic eigenvalues of $\boldsymbol{Q}(\lambda)$, too. For example, according to (3.13a) with any pre-chosen $\varPhi$, we should seek best approximations to λ_i^+ for $1 \le i \le k$ by

$$\text{minimizing } \varPhi(\lambda_{1,X}^+, \cdots, \lambda_{k,X}^+) \text{ subject to } \mathcal{R}(X) \subseteq \mathcal{Y} \text{ and } \operatorname{rank}(X) = k. \tag{3.13a$'$}$$

Noticing that any $X \in \mathbb{C}^{n\times k}$ satisfying $\mathcal{R}(X) \subseteq \mathcal{Y}$ and $\operatorname{rank}(X) = k$ can be written as $X = Y\hat{X}$ for some $\hat{X} \in \mathbb{C}^{m\times k}$ with $\operatorname{rank}(\hat{X}) = k$, we see that $\lambda_{j,X}^+$ are pos-type quadratic eigenvalues of $[Y\widehat{X}]^{\mathrm{H}}\boldsymbol{Q}(\lambda)[Y\widehat{X}] = \widehat{X}^{\mathrm{H}}\, Y^{\mathrm{H}}\boldsymbol{Q}(\lambda)Y\,\widehat{X}$. Varying X subject to $\mathcal{R}(X) \subseteq \mathcal{Y}$ and $\operatorname{rank}(X) = k$ is transferred to varying $\hat{X} \in \mathbb{C}^{m\times k}$ subject to $\operatorname{rank}(\hat{X}) = k$. Consequently,

$$\min_X \varPhi(\lambda_{1,X}^+, \cdots, \lambda_{k,X}^+) = \min_{\widehat{X}} \varPhi(\mu_{1,\widehat{X}}^+, \cdots, \mu_{k,\widehat{X}}^+), \tag{5.2}$$

where $\mu_{j,\widehat{X}}^+$ are pos-type quadratic eigenvalues of $\widehat{X}^{\mathrm{H}}\, Y^{\mathrm{H}}\boldsymbol{Q}(\lambda)Y\,\widehat{X}$. Apply Theorem 3.3 to see the right-hand side of (5.2) is $\varPhi(\mu_1^+, \cdots, \mu_k^+)$, indicating μ_j^+ for $1 \le j \le k$ provide the best approximations to the k smallest λ_j^+, as expected.

The same statement can be made about μ_j^+ as approximations to the largest λ_j^+, μ_j^- as approximations to the smallest λ_j^- or as approximations to the largest λ_j^-, using other min-max principles in Theorems 3.3, 3.4, and 3.5.

In summary, our discussion so far lead to a Rayleigh-Ritz type procedure detailed in Algorithm 5.1 to compute the best approximations to the desired quadratic eigenpairs of $\boldsymbol{Q}(\lambda)$, given a pre-built subspace $\mathcal{Y}$.

6 The steepest descent/ascent method

A common approach to solve a quadratic eigenvalue problem in general, as well as any polynomial eigenvalue problem, is through *linearization* which converts the problem into a linear generalized eigenvalue problem of a matrix pencil [25, 42, 41]. The latter can be either solved by some iterative methods for a large scale problem or by the QZ algorithm [2, 44] for a problem of small to modest size (n up to around a few thousands for example). This approach is usually adopted for QEP without much structure to exploit. For HQEP, however, it is a different story – there is much to exploit. Most recent development includes the solvent approach [10, 21, 24, 61] for certain kinds of QEPs among which is HQEP [20]. Numerical evidence indicates that this solvent approach is rather efficient for QEP of small to modest sizes.

In this paper, we focus on optimization approaches based on various min-max principles previously established and the new ones established here. They are iterative methods and intended for solving large scale HQEP.

The equations in (3.8):

$$\lambda_1^+ = \min_{x\neq 0} \rho_+(x), \quad \lambda_n^+ = \max_{x\neq 0} \rho_+(x), \tag{3.8a}$$

$$\lambda_1^- = \min_{x\neq 0} \rho_-(x), \quad \lambda_n^- = \max_{x\neq 0} \rho_-(x). \tag{3.8b}$$

naturally suggest using some optimization techniques, including the steepest descent/ascent or CG-type method, to compute the first or last quadratic eigenpair $(\lambda_j^\pm, u_j^\pm)$ as in the case of the standard Hermitian eigenvalue problem [3, 14]. Block variations can also be devised to simultaneously compute the first or last few quadratic eigenpairs $(\lambda_j^\pm, u_j^\pm)$ again as in the case of the standard Hermitian eigenvalue problem [3, 40].

6.1 Gradients

To apply any of optimization techniques, we need to compute the gradients of $\rho_\pm(x)$. To this end, we use $\rho(x)$ for either $\rho_+(x)$ or $\rho_-(x)$. As x is perturbed to $x+p$, where p is assumed small in magnitude, $\rho(x+p)$ is changed to $\rho(x+p) = \rho(x) + \eta$, where the magnitude η is comparable to $\|p\|$. We have by (3.1)

$$[\rho(x)+\eta]^2 \,(x+p)^{\mathrm{H}} A(x+p) + [\rho(x)+\eta]\,(x+p)^{\mathrm{H}} B(x+p) + (x+p)^{\mathrm{H}} C(x+p) = 0$$

which gives, upon noticing $f(\rho(x), x) = 0$, that

$$\begin{aligned}[2\rho(x)\, x^{\mathrm{H}}Ax + x^{\mathrm{H}}Bx]\eta + p^{\mathrm{H}}[\rho(x)^2 Ax + \rho(x)Bx + Cx] \\ + [\rho(x)^2Ax + \rho(x)Bx + Cx]^{\mathrm{H}}p + O(\|p\|^2) = 0\end{aligned}$$

and thus

$$\eta = -\frac{p^{\mathrm{H}}[\rho(x)^2Ax + \rho(x)Bx + Cx] + [\rho(x)^2Ax + \rho(x)Bx + Cx]^{\mathrm{H}}p}{2\rho(x)\, x^{\mathrm{H}}Ax + x^{\mathrm{H}}Bx}.$$

Therefore the gradient of $\rho(x)$ at x is

$$\nabla\rho(x) = -\frac{2[\rho(x)^2A + \rho(x)B + C]x}{2\rho(x)\, x^{\mathrm{H}}Ax + x^{\mathrm{H}}Bx},$$

or equivalently

$$\nabla\rho_{\pm}(x) = \mp\frac{2\,\boldsymbol{Q}(\rho_{\pm}(x))x}{\varsigma(x)}, \tag{6.1}$$

where we have used (3.5).

It is important to notice that the gradient $\nabla\rho_{\pm}(x)$ is parallel to the residual vector

$$r_{\pm}(x) := [\rho_{\pm}(x)^2 A + \rho_{\pm}(x)B + C]x = \boldsymbol{Q}(\rho_{\pm}(x))x \tag{6.2}$$

whose normalized norm is commonly used to determine if the approximate eigenpair $(\rho_{\pm}(x), x)$ meets a pre-set tolerance `rtol`:

$$\frac{\|r_{\pm}(x)\|}{|\rho_{\pm}(x)|^2\|Ax\| + |\rho_{\pm}(x)|\,\|Bx\| + \|Cx\|} < \texttt{rtol}. \tag{6.3}$$

If (6.3) holds for $(\rho_{+}(x), x)$, then it is accepted as a converged pos-type quadratic eigenpairs, and similarly for $(\rho_{-}(x), x)$. Here which vector norm $\|\cdot\|$ to use is usually inconsequential, but for the sake of convenience. More conservatively, $\|Ax\|$ in the denominator should be replaced by $\|A\|\,\|x\|$, and likewise for $\|Bx\|$ and $\|Cx\|$ there. For large sparse matrices, the use of $\|Ax\|$, $\|Bx\|$, and $\|Cx\|$ is more economical because of their availability.

Beside being easily implementable, the use of (6.3) can also be rationalized by the existing backward error analysis of approximate eigenpairs for polynomial eigenvalue problems [25, 36, 62].

6.2 The steepest descent/ascent method

Now the steepest descent/ascent method for computing one of $\lambda_{\ell}^{\pm}$ for $\ell \in \{1, n\}$ can be readily given. For this purpose, we fix two parameters "typ" and ℓ with varying ranges as

$$\text{typ} \in \{+, -\}, \quad \ell \in \{1, n\} \tag{6.4}$$

to mean that we are to compute the quadratic eigenpair $(\lambda_{\ell}^{\text{typ}}, u_{\ell}^{\text{typ}})$. A key step of the method is the following line-search problem

$$t_{\text{opt}} = \underset{t\in\mathbb{C}}{\text{argopt}}\ \rho_{\text{typ}}(x + t\,p), \tag{6.5}$$

where x is the current approximation to u_{ℓ}^{typ} (thus no reason to let $x = 0$), p is the search direction, and

$$\text{argopt} = \begin{cases} \text{argmin}, & \text{for } \ell = 1, \\ \text{argmax}, & \text{for } \ell = n. \end{cases} \tag{6.6}$$

The next approximate quadratic eigenvector is

$$y = \begin{cases} x + t_{\text{opt}}\,p, & \text{if } t_{\text{opt}} \text{ is finite}, \\ p, & \text{otherwise}. \end{cases} \tag{6.7}$$

But the line-search problem (6.5) doesn't seem to be solvable straightforwardly by simple calculus as for the standard symmetric eigenvalue problem (see, e.g., [3, 14, 40, 70]), given the (complicated) expressions for ρ_{typ} in (3.2). Fortunately, the theory we developed in

Algorithm 6.1 Steepest descent/ascent method

Given an initial approximation $\boldsymbol{x}_0$ to u_ℓ^{typ}, and a relative tolerance `rtol`, the algorithm computes an approximate pair to $(\lambda_\ell^{\text{typ}}, u_\ell^{\text{typ}})$ with the prescribed `rtol`.

$\boldsymbol{x}_0 = \boldsymbol{x}_0/\|\boldsymbol{x}_0\|$, $\boldsymbol{\rho}_0 = \rho_{\text{typ}}(\boldsymbol{x}_0)$, $\boldsymbol{r}_0 = r_{\text{typ}}(\boldsymbol{x}_0)$;
for $i = 0, 1, \ldots$ **do**
  if $\|\boldsymbol{r}_i\|/(|\boldsymbol{\rho}_i|^2\|A\boldsymbol{x}_i\| + |\boldsymbol{\rho}_i|\,\|B\boldsymbol{x}_i\| + \|C\boldsymbol{x}_i\|) \le$ `rtol` **then**
    BREAK;
  else
    solve QEP for $Y_i^{\text{H}}\boldsymbol{Q}(\lambda)Y_i$, where $Y_i = [\boldsymbol{x}_i, \boldsymbol{r}_i]$ to get its quadratic eigenvalues $\mu_j^{\pm}$ as in (6.8) and corresponding quadratic eigenvectors $y_j^{\pm}$;
    select the next approximate quadratic eigenpair $(\mu, y) = (\mu_j^{\text{typ}}, Y_i y_j^{\text{typ}})$ according to the table (6.9);
    $\boldsymbol{x}_{i+1} = y/\|y\|$, $\boldsymbol{\rho}_{i+1} = \mu$, $\boldsymbol{r}_{i+1} = r_{\text{typ}}(\boldsymbol{x}_{i+1})$;
  end if
end for
return $(\boldsymbol{\rho}_i, \boldsymbol{x}_i)$ as an approximate quadratic eigenpair to $(\lambda_\ell^{\text{typ}}, u_\ell^{\text{typ}})$.

section 5 points us another way to look at it and thus solve it with ease. In fact, the problem is equivalent to find the best possible approximation within the subspace $\mathcal{Y} = \mathcal{R}([x, p])$. Suppose x and p are linearly independent[9] and let $Y = [x, p]$. Solve the 2-by-2 HQEP for $Y^{\text{H}}\boldsymbol{Q}(\lambda)Y$ to get its quadratic eigenvalues

$$\mu_1^- \le \mu_2^- < \mu_1^+ \le \mu_2^+ \tag{6.8}$$

and corresponding quadratic eigenvectors $y_j^{\pm} \in \mathbb{C}^2$. We then have the following table for selecting the next approximate quadratic eigenpair, according to the parameter pair (typ, ℓ).

(typ, ℓ)	current approx.	next approx.
$(+, 1)$	$(\rho_+(x), x)$	(μ_1^+, Yy_1^+)
$(+, n)$	$(\rho_+(x), x)$	(μ_2^+, Yy_2^+)
$(-, 1)$	$(\rho_-(x), x)$	(μ_1^-, Yy_1^-)
$(-, n)$	$(\rho_-(x), x)$	(μ_2^-, Yy_2^-)

(6.9)

In light of this alternative way to solve (6.5), the resulting steepest descent/ascent method is summarized in Algorithm 6.1.

Lemma 6.1. *For* (6.5) – (6.7), $p^{\text{H}}r_{\text{typ}}(y) = 0$.

Proof. If x and p are linearly dependent (the trivial case $p = 0$ included), than $p = \alpha x$ and $y = \beta x$ for some scalars α and β. Thus $\rho_{\text{typ}}(y) = \rho_{\text{typ}}(x)$, $r_{\text{typ}}(y) = \beta r_{\text{typ}}(x)$, and $p^{\text{H}}r_{\text{typ}}(y) = \alpha\beta x^{\text{H}}r_{\text{typ}}(x) = 0$ by the definition of $\rho_{\text{typ}}(x)$.

Suppose x and p are linearly independent. If $|t_{\text{opt}}| = \infty$, then $y = p$. Thus $p^{\text{H}}r_{\text{typ}}(y) = y^{\text{H}}r_{\text{typ}}(y) = 0$. Consider the case that t_{opt} is finite. Let $t = t_{\text{opt}} + s$. For tiny s, we have

$$\rho(y + sp) = \rho(y) - \frac{2\text{Re}\,\left(s[\rho(y)^2Ay + \rho(y)By + Cy]^{\text{H}}p\right)}{2\rho(y)\,y^{\text{H}}Ay + y^{\text{H}}By} + \mathcal{O}(s^2),$$

[9] Otherwise, no improvement is expected by optimizing $\rho_{\text{typ}}(x + tp)$ because then $\rho_{\text{typ}}(x + tp) \equiv \rho_{\text{typ}}(x)$ for all scalar t.

where we drop the subscript "typ" to $\rho_{\rm typ}(\,\cdot\,)$ for convenience. Since $\min_s \rho(y+sp)$ over $s \in \mathbb{C}$ is attained at $s=0$, it must hold that $[\rho(y)^2Ay + \rho(y)By + Cy]^{\rm H}p = 0$, as was to be shown. $\square$

6.3 The extended steepest descent/ascent method

In Algorithm 6.1, the search space is spanned by

$$\boldsymbol{x}_i,\ \boldsymbol{r}_i = \boldsymbol{Q}(\boldsymbol{\rho}_i)\boldsymbol{x}_i.$$

Thus it is the second order Krylov subspace $\mathcal{K}_2(\boldsymbol{Q}(\boldsymbol{\rho}_i), \boldsymbol{x}_i)$ of $\boldsymbol{Q}(\boldsymbol{\rho}_i)$ on $\boldsymbol{x}_i$. Inspired by the inverse free Krylov subspace method [18] which seeks to improve the steepest descent method for the Hermitian generalized eigenvalue problem by extending the search space to a Krylov subspace, we may improve Algorithm 6.1 in the same way, i.e., using a high order Krylov subspace

$$\mathcal{K}_m(\boldsymbol{Q}(\boldsymbol{\rho}_i), \boldsymbol{x}_i) = \operatorname{span}\{\boldsymbol{x}_i, \boldsymbol{Q}(\boldsymbol{\rho}_i)\boldsymbol{x}_i, \ldots, [\boldsymbol{Q}(\boldsymbol{\rho}_i)]^{m-1}\boldsymbol{x}_i\} \tag{6.10}$$

as the search space. Let Y_i be a basis matrix of this Krylov subspace. We then solve[10] the m-by-m HQEP for $Y_i^{\rm H}\boldsymbol{Q}(\lambda)Y_i$ to get its quadratic eigenvalues

$$\mu_1^- \le \cdots \le \mu_m^- < \mu_1^+ \le \cdots \le \mu_m^+ \tag{6.11}$$

and corresponding quadratic eigenvectors $y_j^{\pm}$. We then have the following table for selecting the next approximate quadratic eigenpair, according to the parameter pair (typ, ℓ).

(typ, ℓ)	current approx.	next approx.
$(+,1)$	$(\rho_+(\boldsymbol{x}_i), \boldsymbol{x}_i)$	$(\mu_1^+, Y_i y_1^+)$
$(+,n)$	$(\rho_+(\boldsymbol{x}_i), \boldsymbol{x}_i)$	$(\mu_m^+, Y_i y_m^+)$
$(-,1)$	$(\rho_-(\boldsymbol{x}_i), \boldsymbol{x}_i)$	$(\mu_1^-, Y_i y_1^-)$
$(-,n)$	$(\rho_-(\boldsymbol{x}_i), \boldsymbol{x}_i)$	$(\mu_m^-, Y_i y_m^-)$

(6.12)

We summarize the resulting method, called the *Extended Steepest Descent/Ascent method*, into Algorithm 6.2.

When $m=2$, Algorithm 6.2 reduces to the steepest descent/ascent method given in Algorithm 6.1.

6.4 Convergence analysis

While our convergent results are stated for all four possible $(\mathrm{typ}, \ell) \in \{(\pm,1),(\pm,n)\}$, our proofs will be presented mostly for one (typ, ℓ)

$$(\mathrm{typ}, \ell) = (+,1), \quad \text{and thus argopt} = \text{argmin in (6.6)} \tag{6.13}$$

to save space. Proofs for other (typ, ℓ) can be obtained with minor changes accordingly. For convenience, in our proofs we will drop the pos-type sign "+" in $r_+(\,\cdot\,)$, $\rho_+(\,\cdot\,)$, and

[10] Often $Y_i \in \mathbb{C}^{n\times m}$, but there is a possibility that $\dim \mathcal{K}_m(\boldsymbol{Q}(\boldsymbol{\rho}_i), \boldsymbol{x}_i) < m$. When this occurs, Y_i will have fewer columns than m, and the rest of the development is still valid with minor changes. This is rare, especially in actual computations. For simplicity of presentation, we will assume that Y_i has m columns.

Algorithm 6.2 Extended steepest descent/ascent method

Given an initial approximation $\boldsymbol{x}_0$ to $u_\ell^{\rm typ}$, and a relative tolerance `rtol`, and the search space dimension m, the algorithm computes an approximate pair to $(\lambda_\ell^{\rm typ}, u_\ell^{\rm typ})$ with the prescribed `rtol`.

1: $\boldsymbol{x}_0 = \boldsymbol{x}_0/\|\boldsymbol{x}_0\|$, $\boldsymbol{\rho}_0 = \rho_{\rm typ}(\boldsymbol{x}_0)$, $\boldsymbol{r}_0 = r_{\rm typ}(\boldsymbol{x}_0)$;
2: **for** $i = 0, 1, \ldots$ **do**
3:    **if** $\|\boldsymbol{r}_i\|/(|\boldsymbol{\rho}_i|^2\|A\boldsymbol{x}_i\| + |\boldsymbol{\rho}_i|\,\|B\boldsymbol{x}_i\| + \|C\boldsymbol{x}_i\|) \le$ `rtol` **then**
4:       **BREAK**;
5:    **else**
6:       compute a basis matrix Y_i for the Krylov subspace $\mathcal{K}_m(\boldsymbol{Q}(\boldsymbol{\rho}_i), \boldsymbol{x}_i)$ in (6.10);
7:       solve QEP for $Y_i^{\rm H}\boldsymbol{Q}(\lambda)Y_i$ to get its quadratic eigenvalues $\mu_j^{\pm}$ as in (6.11) and corresponding quadratic eigenvectors $y_j^{\pm}$;
8:       select the next approximate quadratic eigenpair $(\mu, y) = (\mu_j^{\rm typ}, Yy_j^{\rm typ})$ according to the table in (6.12);
9:       $\boldsymbol{x}_{i+1} = y/\|y\|$, $\boldsymbol{\rho}_{i+1} = \mu$, $\boldsymbol{r}_{i+1} = r_{\rm typ}(\boldsymbol{x}_{i+1})$;
10:    **end if**
11: **end for**
12: **return** $(\boldsymbol{\rho}_i, \boldsymbol{x}_i)$ as an approximate quadratic eigenpair to $(\lambda_\ell^{\rm typ}, u_\ell^{\rm typ})$.

u_j^+ with an understanding that they are all for the pos-type, even though occasionally, the sign is still written out at critical places.

By Theorem 2.5, $\boldsymbol{Q}(\lambda)$ has n linearly independent pos-type quadratic eigenvectors u_j^+ for $1 \le j \le n$ and n linearly independent neg-type quadratic eigenvectors u_j^- for $1 \le j \le n$. Define for each (pos/neg-type) quadratic eigenvalue μ its corresponding quadratic eigenspace

$$\mathcal{U}_\mu = \{x \in \mathbb{C}^n \mid \boldsymbol{Q}(\mu)x = 0\} = \bigoplus_{\lambda_i^{\rm typ} = \mu} \operatorname{span}\{u_i^{\rm typ}\}.$$

We'll use the angle $\theta(\boldsymbol{x}_i, \mathcal{U}_\mu)$ from $\boldsymbol{x}_i$ to an eigenspace $\mathcal{U}_\mu$:

$$\cos\theta(\boldsymbol{x}_i, \mathcal{U}_\mu) := \min_{0 \ne u \in \mathcal{U}_\mu} \frac{|u^{\rm H}\boldsymbol{x}_i|}{\|\boldsymbol{x}_i\|_2\|u\|_2}$$

to measure the convergence of $\boldsymbol{x}_i$ towards $\mathcal{U}_\mu$. Note $0 \le \theta(\boldsymbol{x}_i, \mathcal{U}_\mu) \le \pi/2$.

For the sake of our convergence analysis, it is convenient for us to execute Algorithms 6.1 and 6.2 without their Lines 3 and 4 so that $\boldsymbol{x}_i$, $\boldsymbol{r}_i$, and $\boldsymbol{\rho}_i$ are defined for all $i \ge 0$. But without the two lines, we need to be clear about the case when $\boldsymbol{r}_i = 0$ for some i. When it occurs, $\mathcal{K}_m(\boldsymbol{Q}(\boldsymbol{\rho}_i), \boldsymbol{x}_i) = \operatorname{span}\{\boldsymbol{x}_i\}$ for any $m \ge 2$. For Algorithm 6.2, all subsequent $\boldsymbol{x}_j$, $\boldsymbol{\rho}_j$, and $\boldsymbol{r}_j$ for $j > i$ are well-defined. In fact, we will have

$$\boldsymbol{\rho}_i = \boldsymbol{\rho}_{i+1} = \cdots, \ \boldsymbol{x}_i = \boldsymbol{x}_{i+1} = \cdots, \ \boldsymbol{r}_i = \boldsymbol{r}_{i+1} = \cdots = 0. \tag{6.14}$$

But for Algorithm 6.1, all we have to do is to modify its Line 6 to "$Y_i = \boldsymbol{x}_i$ if $\boldsymbol{r}_i = 0$" and then $\boldsymbol{x}_j$, $\boldsymbol{\rho}_j$, and $\boldsymbol{r}_j$ for $j > i$ are again well-defined and they again satisfy (6.14).

Theorem 6.1. *Let the sequences* $\{\boldsymbol{\rho}_i\}, \{\boldsymbol{r}_i\}, \{\boldsymbol{x}_i\}$ *be produced by* Algorithm 6.1/6.2.

1. *Only one of the following two mutually exclusive situations can occur:*

   (a) *For some i, (6.14) holds, and $(\boldsymbol{\rho}_i, \boldsymbol{x}_i)$ is a quadratic eigenpair of $\boldsymbol{Q}(\lambda)$.*

   (b) *$\boldsymbol{\rho}_i$ is strictly monotonically decreasing for $(\mathrm{typ}, \ell) \in \{(\pm, 1)\}$ or strictly monotonically increasing for $(\mathrm{typ}, \ell) \in \{(\pm, n)\}$, $\boldsymbol{r}_i \neq 0$ for all i, and no two $\boldsymbol{x}_i$ are linearly dependent.*

2. $\boldsymbol{x}_i^{\mathrm{H}} \boldsymbol{r}_i = 0$, $\boldsymbol{r}_i^{\mathrm{H}} \boldsymbol{r}_{i+1} = 0$, $\boldsymbol{x}_i^{\mathrm{H}} \boldsymbol{r}_{i+1} = 0$ *for* Algorithm 6.1*;*

3. $\boldsymbol{x}_i^{\mathrm{H}} \boldsymbol{r}_i = 0$, $Y_i^{\mathrm{H}} \boldsymbol{r}_{i+1} = 0$ *for* Algorithm 6.2*;*

4. *In the case of 1(b),*

   (a) $\boldsymbol{\rho}_i \to \hat{\rho} \in [\lambda_1^{\mathrm{typ}}, \lambda_n^{\mathrm{typ}}]$ *as* $i \to \infty$,

   (b) $\boldsymbol{r}_i \neq 0$ *for all i but* $\boldsymbol{r}_i \to 0$ *as* $i \to \infty$,

   (c) *$\hat{\rho}$ is a quadratic eigenvalue of $\boldsymbol{Q}(\lambda)$, and any limit point $\hat{x}$ of $\{\boldsymbol{x}_i\}$ is a corresponding quadratic eigenvector, i.e.,* $\boldsymbol{Q}(\hat{\rho})\hat{x} = 0$,

   (d) $\theta(\boldsymbol{x}_i, \mathcal{U}_{\hat{\rho}}) \to 0$ *as* $i \to \infty$.

Proof. As we remarked at the beginning of this subsection, we will prove the claims only for $(\mathrm{typ}, \ell) = (+, 1)$.

There are only two possibilities: either $\boldsymbol{r}_i = 0$ for some i or $\boldsymbol{r}_i \neq 0$ for all i. If $\boldsymbol{r}_i = 0$ for some i, then $\boldsymbol{\rho}_i = \boldsymbol{\rho}_{i+1}$ and $\boldsymbol{x}_i = \boldsymbol{x}_{i+1}$ because $\rho(\boldsymbol{x}_i + t\boldsymbol{r}_i) \equiv \rho(\boldsymbol{x}_i)$. Consequently $\boldsymbol{r}_{i+1} = 0$, and the equations in (6.14) hold. Consider now $\boldsymbol{r}_i \neq 0$ for all i. Note that $\boldsymbol{r}_i \neq 0$ implies $\nabla\boldsymbol{\rho}_i \neq 0$, and so $\rho(\boldsymbol{x}_i - s\nabla\boldsymbol{\rho}_i) < \rho(\boldsymbol{x}_i)$ for some s with sufficiently tiny $|s|$. This in turn implies $\rho(\boldsymbol{x}_i + t\boldsymbol{r}_i) < \rho(\boldsymbol{x}_i)$ for some t with sufficiently tiny $|t|$ and thus

$$\boldsymbol{\rho}_{i+1} = \inf_t \rho(\boldsymbol{x}_i + t\boldsymbol{r}_i) < \rho(\boldsymbol{x}_i).$$

Therefore $\boldsymbol{\rho}_i$ is strictly monotonically decreasing. No two $\boldsymbol{x}_i$ are linear dependent because linear dependent $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ produce $\boldsymbol{\rho}_i = \boldsymbol{\rho}_j$. This proves item 1.

For item 2, $\boldsymbol{x}_i^{\mathrm{H}} \boldsymbol{r}_i = \boldsymbol{x}_i^{\mathrm{H}} \boldsymbol{Q}(\boldsymbol{\rho}_i)\boldsymbol{x}_i = 0$. Since $\rho(\boldsymbol{x}_{i+1}) = \min_t \rho(\boldsymbol{x}_i + t\boldsymbol{r}_i)$, by Lemma 6.1, $\boldsymbol{r}_i^{\mathrm{H}} \boldsymbol{r}_{i+1} = 0$. We now prove $\boldsymbol{x}_i^{\mathrm{H}} \boldsymbol{r}_{i+1} = 0$. If $\boldsymbol{r}_i = 0$, then all $\boldsymbol{r}_j = 0$ for $j > i$ – no proof is necessary. Consider $\boldsymbol{r}_i \neq 0$. Then $\boldsymbol{\rho}_{i+1} < \boldsymbol{\rho}_i$. Note $\boldsymbol{x}_{i+1}$ is a linear combination of $\boldsymbol{x}_i$ and $\boldsymbol{r}_i$; so we write $\boldsymbol{x}_{i+1} = \alpha_i \boldsymbol{x}_i + \beta_i \boldsymbol{r}_i$ for some scalar α_i and β_i. We know $\beta_i \neq 0$; otherwise $\boldsymbol{x}_{i+1} = \alpha_i \boldsymbol{x}_i$ to yield $\boldsymbol{\rho}_{i+1} = \boldsymbol{\rho}_i$ which contradicts $\boldsymbol{\rho}_{i+1} < \boldsymbol{\rho}_i$. Therefore

$$\boldsymbol{\rho}_{i+1} = \rho(\boldsymbol{r}_i + (\alpha_i/\beta_i)\boldsymbol{x}_i) = \inf_t \rho(\boldsymbol{r}_i + t\boldsymbol{x}_i).$$

Apply Lemma 6.1 with $x = \boldsymbol{r}_i$ and $p = \boldsymbol{x}_i$ to get $\boldsymbol{x}_i^{\mathrm{H}} \boldsymbol{r}_{i+1} = 0$.

For item 3, again $\boldsymbol{x}_i^{\mathrm{H}} \boldsymbol{r}_i = \boldsymbol{x}_i^{\mathrm{H}} \boldsymbol{Q}(\boldsymbol{\rho}_i)\boldsymbol{x}_i = 0$. Let $\boldsymbol{x}_{i+1} = Y_i y$. Then for each column z of Y_i, we have

$$\boldsymbol{\rho}_{i+1} = \rho(Y_i y) = \inf_t \rho(Y_i y + tz).$$

Apply Lemma 6.1 with $x = Y_i y$ and $p = z$ to get $z^{\mathrm{H}} \boldsymbol{r}_{i+1} = 0$. Since z is any column of Y_i, we conclude $Y_i^{\mathrm{H}} \boldsymbol{r}_{i+1} = 0$.

Now for item 4(a), since $\boldsymbol{\rho}_i$ is strictly monotonically decreasing and bounded from below since $\boldsymbol{\rho}_i \geq \lambda_1^+$, it is convergent and $\boldsymbol{\rho}_i \to \hat{\rho} \in [\lambda_1^+, \lambda_n^+]$ because $\boldsymbol{\rho}_i = \rho(\boldsymbol{x}_i) \in [\lambda_1^+, \lambda_n^+]$ for all i by Theorem 3.1.

For item 4(b), we have $\|\boldsymbol{r}_i\| = \|(A\boldsymbol{\rho}_i^2 + B\boldsymbol{\rho}_i + C)\boldsymbol{x}_i\| \leq \|A\|(\lambda_n^+)^2 + \|B\|\,|\lambda_n^+| + \|C\|$ since $\|\boldsymbol{x}_i\| = 1$; so both $\{\boldsymbol{r}_i\}$ and $\{\boldsymbol{x}_i\}$ are bounded sequences. It suffices to show that any limit point of $\{\boldsymbol{r}_i\}$ is the zero vector. Assume, to the contrary, $\{\boldsymbol{r}_i\}$ has a nonzero limit point $\hat{r}$, i.e., $\boldsymbol{r}_{i_j} \to \hat{r}$, where $\{\boldsymbol{r}_{i_j}\}$ is a subsequence of $\{\boldsymbol{r}_i\}$. Since $\{\boldsymbol{x}_{i_j}\}$ is bounded, it has a convergent subsequence. Without loss of generality, we may assume $\boldsymbol{x}_{i_j}$ itself is convergent and $\boldsymbol{x}_{i_j} \to \hat{x}$ as $j \to \infty$. We have $\hat{r}^{\mathrm{H}}\hat{x} = 0$ and $\|\hat{x}\| = 1$ because $\boldsymbol{r}_{i_j}^{\mathrm{H}}\boldsymbol{x}_{i_j} = 0$ and $\|\boldsymbol{x}_{i_j}\| = 1$. Now consider the quadratic eigenvalue problem for

$$\boldsymbol{Q}_{i_j}(\lambda) := Y_{i_j}^{\mathrm{H}}\boldsymbol{Q}(\lambda)Y_{i_j} = \begin{bmatrix} \boldsymbol{x}_{i_j}^{\mathrm{H}}\boldsymbol{Q}(\lambda)\boldsymbol{x}_{i_j} & \boldsymbol{x}_{i_j}^{\mathrm{H}}\boldsymbol{Q}(\lambda)\boldsymbol{r}_{i_j} \\ \boldsymbol{r}_{i_j}^{\mathrm{H}}\boldsymbol{Q}(\lambda)\boldsymbol{x}_{i_j} & \boldsymbol{r}_{i_j}^{\mathrm{H}}\boldsymbol{Q}(\lambda)\boldsymbol{r}_{i_j} \end{bmatrix}, \tag{6.15}$$

where $Y_{i_j} = [\boldsymbol{x}_{i_j}, \boldsymbol{r}_{i_j}]$. Since $\boldsymbol{r}_{i_j}^{\mathrm{H}}\boldsymbol{x}_{i_j} = 0$, $\mathrm{rank}(Y_{i_j}) = 2$, and thus $\boldsymbol{Q}_{i_j}(\lambda)$ is hyperbolic. Denote by $\mu_{j;k}^{\pm}$ its quadratic eigenvalues. It can be seen that

$$\lambda_1^- \leq \mu_{j;1}^- \leq \mu_{j;2}^- \leq \lambda_n^- < \lambda_1^+ \leq \mu_{j;1}^+ \leq \mu_{j;2}^+ \leq \lambda_n^+. \tag{6.16}$$

Then[11] $\lambda_1^+ \leq \boldsymbol{\rho}_{i_j+1} \leq \mu_{j;1}^+$. Let

$$\widehat{\boldsymbol{Q}}(\lambda) = \lim_{j\to\infty} \boldsymbol{Q}_{i_j}(\lambda)$$

whose quadratic eigenvalues are denoted by $\hat{\mu}_i^{\pm}$. By the continuity of the quadratic eigenvalues with respect to the entries of coefficient matrices of a quadratic polynomial with a nonsingular leading coefficient matrix, we know $\mu_{j;i}^{\pm} \to \hat{\mu}_i^{\pm}$ as $j \to \infty$, and thus

$$\lambda_1^- \leq \hat{\mu}_1^- \leq \hat{\mu}_2^- \leq \lambda_n^- < \lambda_1^+ \leq \hat{\mu}_1^+ \leq \hat{\mu}_2^+ \leq \hat{\lambda}_n^+. \tag{6.17}$$

Notice by (6.16) and (6.17)

$$\lambda_1^+ \leq \boldsymbol{\rho}_{i_j+1} \leq \mu_{j;1}^+ \quad \Rightarrow \quad \hat{\mu}_2^- < \lambda_1^+ \leq \hat{\rho} \leq \hat{\mu}_1^+. \tag{6.18}$$

On the other hand, by (6.16), we have

$$\widehat{\boldsymbol{Q}}(\hat{\rho}) = \lim_{j\to\infty} \boldsymbol{Q}_{i_j}(\boldsymbol{\rho}_{i_j}) = \lim_{j\to\infty} \begin{bmatrix} 0 & \boldsymbol{r}_{i_j}^{\mathrm{H}}\boldsymbol{r}_{i_j} \\ \boldsymbol{r}_{i_j}^{\mathrm{H}}\boldsymbol{r}_{i_j} & \boldsymbol{r}_{i_j}^{\mathrm{H}}\boldsymbol{Q}(\boldsymbol{\rho}_{i_j})\boldsymbol{r}_{i_j} \end{bmatrix} = \begin{bmatrix} 0 & \hat{r}^{\mathrm{H}}\hat{r} \\ \hat{r}^{\mathrm{H}}\hat{r} & \hat{r}^{\mathrm{H}}\boldsymbol{Q}(\hat{\rho})\hat{r} \end{bmatrix}$$

which is indefinite because $\hat{r}^{\mathrm{H}}\hat{r} > 0$. But by (6.18) and Theorem 2.1, $\widehat{\boldsymbol{Q}}(\hat{\rho}) \preceq 0$, a contradiction. So $\hat{r} = 0$, as was to be shown.

For item 4(c), since $\|\boldsymbol{x}_i\| = 1$, $\{\boldsymbol{x}_i\}$ has at least one limit point. Let $\hat{x}$ be any limit point of $\boldsymbol{x}_i$, i.e., $\boldsymbol{x}_{i_j} \to \hat{x}$. Take limit at the both sides of $\boldsymbol{Q}(\boldsymbol{\rho}_{i_j})\boldsymbol{x}_{i_j} = \boldsymbol{r}_{i_j}$ to get $\boldsymbol{Q}(\hat{\rho})\hat{x} = 0$, i.e., $(\hat{\rho}, \hat{x})$ is a quadratic eigenpair.

For item 4(d), write $\theta_i = \theta(\boldsymbol{x}_i, \mathcal{U}_{\hat{\rho}})$ for convenience and write[12] $\boldsymbol{x}_i = \hat{u}_i \cos\theta_i + \hat{v}_i \sin\theta_i$, where $\hat{u}_i \in \mathcal{U}_{\hat{\rho}}$, $\hat{v}_i \in \mathcal{U}_{\hat{\rho}}^{\perp}$ (the orthogonal complement of $\mathcal{U}_{\hat{\rho}}$), and $\|\hat{u}_i\|_2 = \|\hat{v}_i\|_2 = 1$. Then

$$\boldsymbol{r}_i = \boldsymbol{Q}(\boldsymbol{\rho}_i)\boldsymbol{x}_i = (\boldsymbol{\rho}_i - \hat{\rho})\,[(\boldsymbol{\rho}_i + \hat{\rho})A + B]\,\hat{u}_i \cos\theta_i + \boldsymbol{Q}(\boldsymbol{\rho}_i)\hat{v}_i \sin\theta_i. \tag{6.19}$$

[11]For Algorithm 6.1, $\boldsymbol{\rho}_{i_j+1} = \mu_{j;1}^+$.

[12]Without loss of generality, we may assume $\|\cdot\|_2$ is used in the algorithms.

We claim that $\boldsymbol{Q}(\boldsymbol{\rho}_i)\hat{v}_i \sin\theta_i \to 0$. To see this, we notice

$$\|(\boldsymbol{\rho}_i + \hat{\rho})A + B\|_2 \le 2\max\{|\lambda_1^+|, |\lambda_n^+|\}\, \|A\|_2 + \|B\|_2,$$

$\boldsymbol{r}_i \to 0$, and $\boldsymbol{\rho}_i - \hat{\rho} \to 0$. Thus $\boldsymbol{Q}(\boldsymbol{\rho}_i)\hat{v}_i \sin\theta_i \to 0$ by (6.19). The null space of $\boldsymbol{Q}(\hat{\rho})$ is $\mathcal{U}_{\hat{\rho}}$. Since $\boldsymbol{Q}(\hat{\rho})$ is Hermitian,

$$\|\boldsymbol{Q}(\hat{\rho})v\|_2 \ge \gamma\|v\|_2 \quad \text{for any } v \in \mathcal{U}_{\hat{\rho}}^{\perp},$$

where $\gamma = \min|\xi|$ taken over all nonzero $\xi \in \operatorname{eig}(\boldsymbol{Q}(\hat{\rho}))$. Therefore $\|\boldsymbol{Q}(\hat{\rho})\hat{v}_i\|_2 \ge \gamma$. Because $\boldsymbol{\rho}_i \to \hat{\rho}$, for sufficiently large i we have $\|\boldsymbol{Q}(\boldsymbol{\rho}_i)\hat{v}_i\|_2 \ge \gamma/2$ and thus

$$\|\boldsymbol{Q}(\boldsymbol{\rho}_i)\hat{v}_i \sin\theta_i\|_2 \ge (\gamma/2)\sin\theta_i,$$

implying $\sin\theta_i \to 0$ which leads to $\theta_i \to 0$ because $0 \le \theta_i \le \pi/2$. □

Theorem 6.1 ensures us the global convergence of Algorithm 6.1/6.2, but gives no indication as how fast the convergence may be. For that, we turn to our next theorem – Theorem 6.2 – which provides an asymptotic rate of the sequences $\{\boldsymbol{\rho}_i\}$ generated by the algorithms. Both theorems are reminiscent of [18, Theorem 3.2] and [18, Theorem 3.4], respectively. But Theorem 6.2 about the rate of convergence is much more difficult to prove than [18, Theorem 3.4]. Because of that, we will devote the entire subsection 6.5 for its proof.

We introduce a few new notations: for any $x \ne 0$,

$$a(x) = \frac{x^{\mathrm{H}}Ax}{x^{\mathrm{H}}x}, \quad b(x) = \frac{x^{\mathrm{H}}Bx}{x^{\mathrm{H}}x}, \quad c(x) = \frac{x^{\mathrm{H}}Cx}{x^{\mathrm{H}}x}. \tag{6.20}$$

Also recall $\boldsymbol{Q}_{\lambda_0}(\lambda) := \boldsymbol{Q}(\lambda + \lambda_0)$ in (4.5) for a given shift λ_0. Accordingly,

$$b_0(x) = \frac{x^{\mathrm{H}}B_{\lambda_0}x}{x^{\mathrm{H}}x} = \frac{x^{\mathrm{H}}(2\lambda_0 A + B)x}{x^{\mathrm{H}}x}, \quad c_0(x) = \frac{x^{\mathrm{H}}C_{\lambda_0}x}{x^{\mathrm{H}}x} = \frac{x^{\mathrm{H}}\boldsymbol{Q}(\lambda_0)x}{x^{\mathrm{H}}x}. \tag{6.21}$$

Theorem 6.2. *Suppose $\lambda_1^{\mathrm{typ}} \le \boldsymbol{\rho}_0 < \lambda_2^{\mathrm{typ}}$ if $\ell = 1$ or $\lambda_{n-1}^{\mathrm{typ}} < \boldsymbol{\rho}_0 \le \lambda_n^{\mathrm{typ}}$ if $\ell = n$, and let the sequences $\{\boldsymbol{\rho}_i\}, \{\boldsymbol{r}_i\}, \{\boldsymbol{x}_i\}$ be produced by* Algorithm 6.2. *Given a shift $\lambda_0 \ge \lambda_n^+$, define B_{λ_0}, C_{λ_0} by* (4.5)*.*

1. *As $i \to \infty$, $\boldsymbol{\rho}_i$ monotonically converges to $\hat{\rho} = \lambda_\ell^{\mathrm{typ}}$, and $\boldsymbol{x}_i$ converges to u_ℓ^{typ} in direction, i.e., $\theta(\boldsymbol{x}_i, u_\ell^{\mathrm{typ}}) \to 0$.*

2. *The eigenvalues[13] ω_j of $\boldsymbol{Q}(\boldsymbol{\rho}_i)$ can be ordered as*

$$\omega_1 > 0 > \omega_2 \ge \cdots \ge \omega_n \quad \text{if } (\mathrm{typ}, \ell) \in \{(+,1), (-,n)\}, \quad \text{or}, \tag{6.22a}$$
$$\omega_1 < 0 < \omega_2 \le \cdots \le \omega_n \quad \text{if } (\mathrm{typ}, \ell) \in \{(+,n), (-,1)\}. \tag{6.22b}$$

Denote by v_1 the eigenvector of $\boldsymbol{Q}(\boldsymbol{\rho}_i)$ associated with its eigenvalue ω_1. If $\boldsymbol{\rho}_i$ is sufficiently close to $\lambda_\ell^{\mathrm{typ}}$, then

$$|\boldsymbol{\rho}_{i+1} - \lambda_\ell^{\mathrm{typ}}| \le \varepsilon_m^2|\boldsymbol{\rho}_i - \lambda_\ell^{\mathrm{typ}}| + (1-\varepsilon_m^2)\varepsilon_m\eta(v_1)|\boldsymbol{\rho}_i - \lambda_\ell^{\mathrm{typ}}|^{3/2} + O(|\boldsymbol{\rho}_i - \lambda_\ell^{\mathrm{typ}}|^2), \tag{6.23}$$

[13]Their dependency upon i is suppressed for clarity.

where

$$\varepsilon_m = \min_{g\in\mathbb{P}_{m-1},g(\omega_1)\neq 0} \max_{i\neq 1} \frac{|g(\omega_i)|}{|g(\omega_1)|}, \tag{6.24}$$

$$\tau_A = \frac{1}{|\omega_2|}\frac{\|A\|_2}{a(v_1)}, \quad \tau_B = \frac{1}{|\omega_2|}\frac{\|B_{\lambda_0}\|_2}{b_0(v_1)}, \quad \tau_C = \frac{1}{|\omega_2|}\frac{\|C_{\lambda_0}\|_2}{c_0(v_1)}, \tag{6.25}$$

$$\eta(v_1) = 3\tau_A^{1/2} + 2\frac{(b_0(v_1))^2\tau_B^{1/2} + 2a(v_1)c_0(v_1)(\tau_A^{1/2}+\tau_C^{1/2})}{\varsigma_0(v_1)^2}, \tag{6.26}$$

and $\mathbb{P}_{m-1}$*, the set of polynomials of degree no higher than* $m-1$*.*

3. *Denote*[14] *by* γ *and* Γ *the smallest and largest positive eigenvalue of*

$$\begin{cases} -\boldsymbol{Q}(\lambda_\ell^{\mathrm{typ}}) & \text{for } (\mathrm{typ},\ell)\in\{(+,1),(-,n)\}, \\ \boldsymbol{Q}(\lambda_\ell^{\mathrm{typ}}) & \text{for } (\mathrm{typ},\ell)\in\{(+,n),(-,1)\}. \end{cases}$$

If $\boldsymbol{\rho}_i$ *is sufficiently close to* $\lambda_\ell^{\mathrm{typ}}$*, then*

$$|\boldsymbol{\rho}_{i+1} - \lambda_\ell^{\mathrm{typ}}| \le \varepsilon^2|\boldsymbol{\rho}_i - \lambda_\ell^{\mathrm{typ}}| + (1-\varepsilon^2)\varepsilon\eta|\boldsymbol{\rho}_i - \lambda_\ell^{\mathrm{typ}}|^{3/2} + O(|\boldsymbol{\rho}_i - \lambda_\ell^{\mathrm{typ}}|^2), \tag{6.27}$$

where

$$\varepsilon = 2\left[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^{m-1} + \left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^{-(m-1)}\right]^{-1}, \qquad \kappa = \frac{\Gamma}{\gamma}, \tag{6.28}$$

$$\eta = \sqrt{\frac{1}{|\gamma|}}\left[3\sqrt{\frac{\|A\|_2}{a(u)}} + 2\frac{b_0(u)^2}{\varsigma_0(u)^2}\sqrt{\frac{\|B_{\lambda_0}\|_2}{b_0(u)}} + 4\frac{a(u)c_0(u)}{\varsigma_0(u)^2}\left(\sqrt{\frac{\|A\|_2}{a(u)}} + \sqrt{\frac{\|C_{\lambda_0}\|_2}{c_0(u)}}\right)\right] \tag{6.29}$$

$$\le \sqrt{\frac{1}{|\gamma|}}\left[3\sqrt{\frac{\|A\|_2}{a(u)}} + 2\frac{\|B_{\lambda_0}\|_2^2 + 4\|A\|_2\|C_{\lambda_0}\|_2}{b(u)^2 - 4a(u)c(u)}\right], \tag{6.30}$$

and $u = u_\ell^{\mathrm{typ}}$ *for short.*

6.5 Proof of Theorem 6.2

We recall (3.5) to see

$$\begin{aligned} \varsigma(x) :&= \left[(x^{\mathrm{H}}Bx)^2 - 4(x^{\mathrm{H}}Ax)(x^{\mathrm{H}}Cx)\right]^{1/2} \\ &= \pm x^{\mathrm{H}}[2\rho_\pm(x)A + B]x \\ &= \pm x^{\mathrm{H}}\boldsymbol{Q}'(\rho_\pm(x))x, \end{aligned} \tag{6.31}$$

and $\varsigma_0(x) = \varsigma(x)/\|x\|_2^2$. For a perturbation $E\in\mathbb{C}^{n\times n}$ which is assumed Hermitian, we define

$$\boldsymbol{Q}_E(\lambda) := \boldsymbol{Q}(\lambda) + E = \lambda^2 A + \lambda B + C + E. \tag{6.32}$$

[14] $\boldsymbol{Q}(\lambda_\ell^{\mathrm{typ}})$ is singular and, by Theorem 2.1, negative semidefinite if $(\mathrm{typ},\ell)\in\{(+,1),(-,n)\}$ or positive semidefinite if $(\mathrm{typ},\ell)\in\{(+,n),(-,1)\}$.

When $\boldsymbol{Q}_E(\lambda)$ is also hyperbolic, the pos- and neg-type Rayleigh quotients, denoted by $\rho_{E;\pm}$, can be defined for $\boldsymbol{Q}_E(\lambda)$. Accordingly, we will define ς_E and $\varsigma_{E;0}$, too. Specifically,

$$\rho_{E;\pm}(x) = \frac{-(x^{\rm H}Bx) \pm \left\{(x^{\rm H}Bx)^2 - 4(x^{\rm H}Ax)(x^{\rm H}[C+E]x)\right\}^{1/2}}{2(x^{\rm H}Ax)}, \tag{6.33}$$

and

$$\begin{aligned} \varsigma_E(x) :&= \left\{(x^{\rm H}Bx)^2 - 4(x^{\rm H}Ax)(x^{\rm H}[C+E]x)\right\}^{1/2} && (6.34a) \\ &= \pm x^{\rm H}[2\rho_{E;\pm}(x)\,A + B]x, \\ \varsigma_{E;0}(x) :&= \frac{\varsigma_E(x)}{\|x\|_2^2}. && (6.34b) \end{aligned}$$

Lemma 6.2. *Suppose $\boldsymbol{Q}_E(\lambda)$ in (6.32) is also hyperbolic.*

1. *Let (λ_1^+, u_1^+) and (μ_1^+, v_1^+) be the smallest quadratic eigenpair*[15] *with the pos-type of $\boldsymbol{Q}(\lambda)$ and $\boldsymbol{Q}_E(\lambda)$, respectively. Then*
$$\frac{\lambda_{\min}(E)}{\varsigma_0(u_1^+)} \le \lambda_1^+ - \mu_1^+ \le \frac{\lambda_{\max}(E)}{\varsigma_{E;0}(v_1^+)}. \tag{6.35}$$

2. *Let (λ_n^+, u_n^+) and (μ_n^+, v_n^+) be the largest quadratic eigenpair with the pos-type of $\boldsymbol{Q}(\lambda)$ and $\boldsymbol{Q}_E(\lambda)$, respectively. Then*
$$\frac{\lambda_{\min}(E)}{\varsigma_0(v_n^+)} \le \lambda_n^+ - \mu_n^+ \le \frac{\lambda_{\max}(E)}{\varsigma_{E;0}(u_n^+)}. \tag{6.36}$$

3. *Let (λ_1^-, u_1^-) and (μ_1^-, v_1^-) be the smallest quadratic eigenpair with the neg-type of $\boldsymbol{Q}(\lambda)$ and $\boldsymbol{Q}_E(\lambda)$, respectively. Then*
$$\frac{\lambda_{\min}(E)}{\varsigma_0(v_1^-)} \le \mu_1^- - \lambda_1^- \le \frac{\lambda_{\max}(E)}{\varsigma_{E;0}(u_1^-)}. \tag{6.37}$$

4. *Let (λ_n^-, u_n^-) and (μ_n^-, v_n^-) be the largest quadratic eigenpair with the neg-type of $\boldsymbol{Q}(\lambda)$ and $\boldsymbol{Q}_E(\lambda)$, respectively. Then*
$$\frac{\lambda_{\min}(E)}{\varsigma_0(u_n^-)} \le \mu_n^- - \lambda_n^- \le \frac{\lambda_{\max}(E)}{\varsigma_{E;0}(v_n^-)}. \tag{6.38}$$

Proof. As in the proof of Lemma 4.4, we have

$$\mu_1^+ = \min_x \rho_{E;+}(x) \le \rho_{E;+}(u_1^+) \le \rho_+(u_1^+) + \delta_{\rm ub}^+(u_1^+) = \lambda_1^+ + \delta_{\rm ub}^+(u_1^+)$$

which gives

$$\mu_1^+ - \lambda_1^+ \le \delta_{\rm ub}^+(u_1^+), \quad \lambda_1^+ - \mu_1^+ \le \tilde{\delta}_{\rm ub}^+(v_1^+), \tag{6.39}$$

[15] By the smallest (largest) pos/neg-type quadratic eigenpair, we mean the quadratic eigenvalue in question is the smallest (largest) of that given type. The same naming is used for the usual linear eigenpair, too.

where the second inequality is actually obtained from the first one there by switching the roles of $\boldsymbol{Q}(\lambda)$ and $\boldsymbol{Q}_E(\lambda)$. Now use (4.42) in the proof of Theorem 4.1 for $\Delta A = \Delta B = 0$ and $\Delta C = E$ to get item 1.

Similarly, we have

$$\lambda_n^+ = \max_x \rho_+(x) \geq \rho_+(v_n^+) \geq \rho_{E;+}(v_n^+) - \delta_{\rm ub}^+(v_n^+) = \mu_n^+ - \delta_{\rm ub}^+(v_n^+)$$

which gives

$$\mu_n^+ - \lambda_n^+ \leq \delta_{\rm ub}^+(v_n^+), \quad \lambda_n^+ - \mu_n^+ \leq \tilde{\delta}_{\rm ub}^+(u_n^+), \tag{6.40}$$

where the second inequality is actually obtained from switching the roles of $\boldsymbol{Q}(\lambda)$ and $\boldsymbol{Q}_E(\lambda)$. Now use (4.42) in the proof of Theorem 4.1 for $\Delta A = \Delta B = 0$ and $\Delta C = E$ to get item 2.

Items 3 and 4 are corollaries of items 2 and 1 applied to $\boldsymbol{Q}(-\lambda)$ and $\boldsymbol{Q}_E(-\lambda)$. □

Lemma 6.3. *$\boldsymbol{Q}_E(\lambda)$ with $E = -\sigma I$ is hyperbolic if*

$$\sigma > -\frac{(\lambda_1^+ - \lambda_n^-)^2 \lambda_{\min}(A)}{4}. \tag{6.41}$$

Proof. For any vector $x \neq 0$, we have

$$\begin{aligned}(x^{\rm H}Bx)^2 - 4(x^{\rm H}Ax)(x^{\rm H}[C-\sigma I]x) &= (x^{\rm H}Bx)^2 - 4(x^{\rm H}Ax)(x^{\rm H}Cx) + 4\sigma(x^{\rm H}Ax)(x^{\rm H}x) \\ &= [\rho_+(x) - \rho_-(x)]^2 (x^{\rm H}Ax)^2 + 4\sigma(x^{\rm H}Ax)(x^{\rm H}x) \\ &\geq (x^{\rm H}Ax)(x^{\rm H}x)\left[(\lambda_1^+ - \lambda_n^-)^2 \frac{x^{\rm H}Ax}{x^{\rm H}x} + 4\sigma\right] \\ &\geq (x^{\rm H}Ax)(x^{\rm H}x)\left[(\lambda_1^+ - \lambda_n^-)^2 \lambda_{\min}(A) + 4\sigma\right] \\ &> 0,\end{aligned}$$

where the last inequality holds because of (6.41). □

So ς_E and $\varsigma_{E;0}$ are well-defined for any $E = -\sigma I$ satisfying (6.41). To emphasize such special $E = -\sigma I$, we introduce notations

$$\varsigma_\sigma(x) := \varsigma_E(v), \quad \varsigma_{\sigma;0}(v) := \varsigma_{E;0}(v) \quad \text{for } E = -\sigma I. \tag{6.42}$$

For $\boldsymbol{\rho} \in (\lambda_1^{\rm typ}, \lambda_n^{\rm typ})$, it follows from Theorem 2.1 that the largest eigenvalue, denoted by ω_1, of $\boldsymbol{Q}(\boldsymbol{\rho})$ is nonnegative, and thus this $\sigma = \omega_1$ automatically satisfies (6.41). But the smallest eigenvalue, denoted also by ω_1, of $\boldsymbol{Q}(\boldsymbol{\rho})$ is non-positive and (6.41) may fail for $\sigma = \omega_1$ unless $|\omega_1|$ is sufficiently tiny.

Lemma 6.4. *Given $\lambda_1^{\rm typ} \leq \boldsymbol{\rho} \leq \lambda_n^{\rm typ}$, let (ω_1, v_1) be the largest eigenpair $\boldsymbol{Q}(\boldsymbol{\rho})$ if $({\rm typ}, \ell) \in \{(+,1), (-,n)\}$ or the smallest eigenpair $\boldsymbol{Q}(\boldsymbol{\rho})$ if $({\rm typ}, \ell) \in \{(+,n), (-,1)\}$. If (6.41) holds with $\sigma = \omega_1$, then for the four different $({\rm typ}, \ell)$*

$$\frac{\varsigma_0(u_1^+)}{\varsigma_{\omega_1;0}(v_1)}(\boldsymbol{\rho} - \lambda_1^+) \leq \frac{\omega_1}{\varsigma_{\omega_1;0}(v_1)} \leq \boldsymbol{\rho} - \lambda_1^+ \quad \textit{for } ({\rm typ}, \ell) = (+,1), \tag{6.43a}$$

$$\frac{\varsigma_{\omega_1;0}(u_n^+)}{\varsigma_0(v_1)}(\lambda_n^+ - \boldsymbol{\rho}) \leq \frac{-\omega_1}{\varsigma_0(v_1)} \leq \lambda_n^+ - \boldsymbol{\rho} \quad \textit{for } ({\rm typ}, \ell) = (+,n), \tag{6.43b}$$

$$\frac{\varsigma_{\omega_1;0}(u_1^-)}{\varsigma_0(v_1)}(\boldsymbol{\rho}-\lambda_1^-) \le \frac{-\omega_1}{\varsigma_0(v_1)} \le \boldsymbol{\rho}-\lambda_1^- \quad \textit{for } (\mathrm{typ},\ell)=(-,1), \tag{6.43c}$$

$$\frac{\varsigma_0(u_n^-)}{\varsigma_{\omega_1;0}(v_1)}(\lambda_n^- -\boldsymbol{\rho}) \le \frac{\omega_1}{\varsigma_{\omega_1;0}(v_1)} \le \lambda_n^- -\boldsymbol{\rho} \quad \textit{for } (\mathrm{typ},\ell)=(-,n). \tag{6.43d}$$

Moreover, for $\boldsymbol{\rho}$ sufficiently close to $\lambda_\ell^{\mathrm{typ}}$,

$$\frac{\omega_1}{\varsigma_{\omega_1;0}(v_1)} = \boldsymbol{\rho}-\lambda_1^+ + O([\boldsymbol{\rho}-\lambda_1^+]^2) \quad \textit{for } (\mathrm{typ},\ell)=(+,1), \tag{6.44a}$$

$$\frac{-\omega_1}{\varsigma_0(v_1)} = \lambda_n^+ -\boldsymbol{\rho} + O([\lambda_n^+ -\boldsymbol{\rho}]^2) \quad \textit{for } (\mathrm{typ},\ell)=(+,n), \tag{6.44b}$$

$$\frac{-\omega_1}{\varsigma_0(v_1)} = \boldsymbol{\rho}-\lambda_1^- + O([\boldsymbol{\rho}-\lambda_1^-]^2) \quad \textit{for } (\mathrm{typ},\ell)=(-,1), \tag{6.44c}$$

$$\frac{\omega_1}{\varsigma_{\omega_1;0}(v_1)} = \lambda_n^- -\boldsymbol{\rho} + O([\lambda_n^- -\boldsymbol{\rho}]^2) \quad \textit{for } (\mathrm{typ},\ell)=(-,n). \tag{6.44d}$$

Proof. Consider the case $(\mathrm{typ},\ell)=(+,1)$. We have $\omega_1 \ge 0$ and $[\boldsymbol{Q}(\boldsymbol{\rho})-\omega_1 I]\,v_1=0$. Since ω_1 is the largest eigenvalue of $\boldsymbol{Q}(\boldsymbol{\rho})$, $\boldsymbol{Q}(\boldsymbol{\rho})-\omega_1 I \preceq 0$. Thus, $(\boldsymbol{\rho},v_1)$ is the smallest pos-type quadratic eigenpair of $\boldsymbol{Q}_E(\lambda)$ with $E=-\omega_1 I$. By Lemma 6.2,

$$\frac{\omega_1}{\varsigma_{E;0}(v_1)} \le \boldsymbol{\rho}-\lambda_1^+ \le \frac{\omega_1}{\varsigma_0(u_1)}$$

which gives (6.43a). To prove (6.44a), we denote by $\alpha(t)$ the largest eigenvalue of $\boldsymbol{Q}(t)$ near $t=\lambda_1^+$. Then $\alpha(\lambda_1^+)=0$ and $\alpha(\boldsymbol{\rho})=\omega_1$. Note that

$$\boldsymbol{Q}(\boldsymbol{\rho})v_1=\omega_1 v_1 \quad\Rightarrow\quad v_1^{\mathrm{H}}\boldsymbol{Q}(\boldsymbol{\rho})v_1=\omega_1 v_1^{\mathrm{H}}v_1 \quad\Rightarrow\quad v_1^{\mathrm{H}}[\boldsymbol{Q}(\boldsymbol{\rho})-\omega_1 I]v_1=0,$$

i.e., $\boldsymbol{\rho}$ is a Rayleigh quotient of $\boldsymbol{Q}_E(\lambda)$ with $E=-\omega_1 I$. Therefore

$$\alpha'(\boldsymbol{\rho})=\frac{v_1^{\mathrm{H}}\boldsymbol{Q}'(\boldsymbol{\rho})v_1}{v_1^{\mathrm{H}}v_1}=\frac{v_1^{\mathrm{H}}\boldsymbol{Q}_E'(\boldsymbol{\rho})v_1}{v_1^{\mathrm{H}}v_1}=\varsigma_{\omega_1;0}(v_1),$$

where the first equality is due to [56, p.183], and the third equality due to (6.31). Finally $\alpha(\lambda_1^+)=\alpha(\boldsymbol{\rho})+\varsigma_{\omega_1;0}(v_1)(\lambda_1^+-\boldsymbol{\rho})+O(|\lambda_1^+-\boldsymbol{\rho}|^2)$ which leads to (6.44a). □

Remark 6.1. There is a different proof of Lemma 6.4, without using Lemma 6.2. For the case $(\mathrm{typ},\ell)=(+,1)$, $(\boldsymbol{\rho},v_1)$ is the smallest pos-type quadratic eigenpair of $\boldsymbol{Q}_E(\lambda)=$

$\lambda^2 A+\lambda B+C-\omega_1 I$. By direct calculations[16],

$$\omega_1=\omega_1-\frac{u_1^{\rm H}\boldsymbol{Q}(\boldsymbol{\rho})u_1}{u_1^{\rm H}u_1}+\varsigma_0(u_1)(\boldsymbol{\rho}-\lambda_1^+)+\frac{u_1^{\rm H}Au_1}{u_1^{\rm H}u_1}(\boldsymbol{\rho}-\lambda_1^+)^2, \tag{6.45a}$$

$$\omega_1=\frac{v_1^{\rm H}\boldsymbol{Q}(\lambda_1^+)v_1}{v_1^{\rm H}v_1}\qquad+\varsigma_{\omega_1;0}(v_1)(\boldsymbol{\rho}-\lambda_1^+)-\frac{v_1^{\rm H}Av_1}{v_1^{\rm H}v_1}(\boldsymbol{\rho}-\lambda_1^+)^2. \tag{6.45b}$$

Along with $\boldsymbol{Q}(\boldsymbol{\rho})-\omega_1 I\preceq 0$, $\boldsymbol{Q}(\lambda_1^+)\preceq 0$, they yield

$$\frac{\omega_1}{\varsigma_{\omega_1;0}(v_1)}\le\boldsymbol{\rho}-\lambda_1^+\le\frac{\omega_1}{\varsigma_0(u_1)}$$

and then

$$\frac{\varsigma_0(u_1)}{\varsigma_{\omega_1;0}(v_1)}(\boldsymbol{\rho}-\lambda_1^+)\le\frac{\omega_1}{\varsigma_{\omega_1;0}(v_1)}\le\boldsymbol{\rho}-\lambda_1^+$$

which is (6.43a).

While Lemmas 6.5 and 6.6 are stated for any $g\in\mathbb{P}_{m-1}$ with the specified conditions satisfied, in their eventual application, it will be taken to be the one that minimizes ε_g.

Lemma 6.5. *Given $x\in\mathbb{C}^n$, assign $\boldsymbol{\rho}_\pm=\rho_\pm(x)$ and $\boldsymbol{\rho}_{g;\pm}=\rho_\pm(g(\boldsymbol{Q}(\boldsymbol{\rho}_+))x)$ for any $g\in\mathbb{P}_{m-1}$. Suppose $\lambda_1^{\rm typ}\le\boldsymbol{\rho}_{\rm typ}<\lambda_2^{\rm typ}$ if $\ell=1$ or $\lambda_{n-1}^{\rm typ}<\boldsymbol{\rho}_{\rm typ}\le\lambda_n^{\rm typ}$ if $\ell=n$, and let the eigenvalues of $\boldsymbol{Q}(\boldsymbol{\rho}_{\rm typ})$ be ω_j for $1\le j\le n$ which can be arranged as*

$$\begin{aligned}&\omega_1>0>\omega_2\ge\cdots\ge\omega_n\quad\text{if } ({\rm typ},\ell)\in\{(+,1),(-,n)\},\quad\text{or,}\\&\omega_1<0<\omega_2\le\cdots\le\omega_n\quad\text{if } ({\rm typ},\ell)\in\{(+,n),(-,1)\}.\end{aligned}$$

Denote by v_1 the eigenvector of $\boldsymbol{Q}(\boldsymbol{\rho}_{\rm typ})$ associated with its eigenvalue ω_1. Then for a $g\in\mathbb{P}_{m-1}$ such that $g(\omega_1)\ne0$ and

$$\varepsilon_g:=\max_{i\ne1}\frac{|g(\omega_i)|}{|g(\omega_1)|}<1, \tag{6.46}$$

we have

$$|\boldsymbol{\rho}_{g;{\rm typ}}-\lambda_\ell^{\rm typ}|\le|\boldsymbol{\rho}_{\rm typ}-\lambda_\ell^{\rm typ}|-\frac{|\omega_1|}{|\boldsymbol{\rho}_{\rm typ}-\boldsymbol{\rho}_{g;{\rm typ}'}|\,a(v_1)}+\frac{|\omega_1|}{|\boldsymbol{\rho}_{\rm typ}-\boldsymbol{\rho}_{g;{\rm typ}'}|\,a(v_1)}h(\varepsilon_g,\omega_1), \tag{6.47}$$

---

[16] In fact,

$$\begin{aligned}u_1^{\rm H}Au_1(\boldsymbol{\rho}-\lambda_1^+)^2+\varsigma(u_1)(\boldsymbol{\rho}-\lambda_1^+)&=u_1^{\rm H}Au_1\big[\boldsymbol{\rho}^2-2\boldsymbol{\rho}\lambda_1^++(\lambda_1^+)^2\big]+(2\lambda_1^+u_1^{\rm H}Au_1+u_1^{\rm H}Bu_1)(\boldsymbol{\rho}-\lambda_1^+)\\&=\boldsymbol{\rho}^2u_1^{\rm H}Au_1+\boldsymbol{\rho}u_1^{\rm H}Bu_1-(\lambda_1^+)^2u_1^{\rm H}Au_1-\lambda_1^+u_1^{\rm H}Bu_1\\&=u_1^{\rm H}\boldsymbol{Q}(\boldsymbol{\rho})u_1-u_1^{\rm H}\boldsymbol{Q}(\lambda_1^+)u_1\\&=u_1^{\rm H}\boldsymbol{Q}(\boldsymbol{\rho})u_1,\\v_1^{\rm H}Av_1(\boldsymbol{\rho}-\lambda_1^+)^2-\varsigma_{\omega_1}(v_1)(\boldsymbol{\rho}-\lambda_1^+)&=v_1^{\rm H}Av_1\big[\boldsymbol{\rho}^2-2\boldsymbol{\rho}\lambda_1^++(\lambda_1^+)^2\big]-(2\boldsymbol{\rho}v_1^{\rm H}Av_1+v_1^{\rm H}Bv_1)(\boldsymbol{\rho}-\lambda_1^+)\\&=(\lambda_1^+)^2v_1^{\rm H}Av_1+\lambda_1^+v_1^{\rm H}Bv_1-\boldsymbol{\rho}^2v_1^{\rm H}Av_1-\boldsymbol{\rho}v_1^{\rm H}Bv_1\\&=v_1^{\rm H}\boldsymbol{Q}(\lambda_1^+)v_1-v_1^{\rm H}\boldsymbol{Q}(\boldsymbol{\rho})v_1\\&=v_1^{\rm H}\boldsymbol{Q}(\lambda_1^+)v_1-\omega_1v_1^{\rm H}v_1.\end{aligned}$$

They lead to the equations in (6.45) right away.

where typ′ *is the opposite type of* typ, *and*

$$h(\varepsilon_g, \omega_1) = 1 - \frac{1-\varepsilon_g^2}{\left(1+\varepsilon_g|\omega_1|^{1/2}\tau_A^{1/2}\right)^2}, \quad \tau_A = \frac{1}{|\omega_2|}\frac{\|A\|_2}{a(v_1)}. \tag{6.48}$$

Proof. Consider the case $(\mathrm{typ}, \ell) = (+, 1)$, and write $\boldsymbol{\rho} = \boldsymbol{\rho}_+$. Without loss of generality, we may assume $\|v_1\|_2 = 1$. Let the eigenvalue decomposition of $\boldsymbol{Q}(\boldsymbol{\rho})$ be

$$\boldsymbol{Q}(\boldsymbol{\rho}) = V\Sigma V^{\mathrm{H}}, \quad V = [v_1, \cdots, v_n], \quad \Sigma = \mathrm{diag}(\omega_1, \cdots, \omega_n),$$

where $\omega_1 > 0 > \omega_2 \geq \cdots \geq \omega_n$ and $V^{\mathrm{H}}V = I_n$. Set

$$\hat{x} = V^{\mathrm{H}}x = \begin{bmatrix}\xi_1\\ \xi_2\\ \vdots\\ \xi_n\end{bmatrix}, \quad \hat{x}_2 = \hat{x} - \xi_1 e_1 = \begin{bmatrix}0\\ \xi_2\\ \vdots\\ \xi_n\end{bmatrix}.$$

Then

$$0 = x^{\mathrm{H}}\boldsymbol{Q}(\boldsymbol{\rho})x = \hat{x}^{\mathrm{H}}\Sigma\hat{x} = \omega_1|\xi_1|^2 + \sum_{i\neq 1}\omega_i|\xi_i|^2. \tag{6.49}$$

Note that for any vector z, $z^{\mathrm{H}}\boldsymbol{Q}(\lambda)z = z^{\mathrm{H}}Az\,[\lambda - \rho_+(z)][\lambda - \rho_-(z)]$. Substitute $\lambda = \boldsymbol{\rho}$ and $z = g(\boldsymbol{Q}(\boldsymbol{\rho}))x$ to get

$$\begin{aligned}
\boldsymbol{\rho}_g - \lambda_1^+ &= \boldsymbol{\rho} - \lambda_1^+ - \frac{1}{\boldsymbol{\rho} - \boldsymbol{\rho}_{g;-}} \cdot \frac{x^{\mathrm{H}}g(\boldsymbol{Q}(\boldsymbol{\rho}))^{\mathrm{H}}\boldsymbol{Q}(\boldsymbol{\rho})g(\boldsymbol{Q}(\boldsymbol{\rho}))x}{x^{\mathrm{H}}g(\boldsymbol{Q}(\boldsymbol{\rho}))^{\mathrm{H}}Ag(\boldsymbol{Q}(\boldsymbol{\rho}))x}\\
&= \boldsymbol{\rho} - \lambda_1^+ - \frac{1}{\boldsymbol{\rho} - \boldsymbol{\rho}_{g;-}} \cdot \frac{\hat{x}^{\mathrm{H}}g(\Sigma)^{\mathrm{H}}\Sigma g(\Sigma)\hat{x}}{\hat{x}^{\mathrm{H}}g(\Sigma)^{\mathrm{H}}\widehat{A}g(\Sigma)\hat{x}},
\end{aligned} \tag{6.50}$$

where $\widehat{A} = V^{\mathrm{H}}AV$ and $\boldsymbol{\rho}_g = \boldsymbol{\rho}_{g;+}$. We need to estimate the right-hand side of (6.50). We have

$$\begin{aligned}
\hat{x}^{\mathrm{H}}g(\Sigma)^{\mathrm{H}}\Sigma g(\Sigma)\hat{x} &= \omega_1|g(\omega_1)|^2|\xi_1|^2 + \sum_{i\neq 1}\omega_i|g(\omega_i)|^2|\xi_i|^2\\
&\geq \omega_1|g(\omega_1)|^2|\xi_1|^2 + \varepsilon_g^2|g(\omega_1)|^2\sum_{i\neq 1}\omega_i|\xi_i|^2\\
&= \omega_1|g(\omega_1)|^2|\xi_1|^2 - \varepsilon_g^2|g(\omega_1)|^2\omega_1|\xi_1|^2 \qquad \text{(by (6.49))}\\
&= (1-\varepsilon_g^2)\omega_1|g(\omega_1)|^2|\xi_1|^2, \qquad\qquad (6.51)\\
\hat{x}^{\mathrm{H}}g(\Sigma)^{\mathrm{H}}\widehat{A}g(\Sigma)\hat{x} &= \|g(\Sigma)\hat{x}\|_{\widehat{A}}^2\\
&= \|g(\omega_1)\xi_1 e_1 + g(\Sigma)\hat{x}_2\|_{\widehat{A}}^2\\
&\leq \left[|g(\omega_1)|\,|\xi_1|\,\|e_1\|_{\widehat{A}} + \|g(\Sigma)\hat{x}_2\|_{\widehat{A}}\right]^2\\
&\leq \left[|g(\omega_1)|\,|\xi_1|\,\|e_1\|_{\widehat{A}} + \varepsilon_g|g(\omega_1)|\|\hat{x}_2\|_{\widehat{A}}\right]^2\\
&\leq \left[|g(\omega_1)|\,|\xi_1|\,\|e_1\|_{\widehat{A}} + \varepsilon_g|g(\omega_1)|\left(\|A\|_2\frac{\omega_1}{-\omega_2}|\xi_1|^2\right)^{1/2}\right]^2 \qquad (6.52)
\end{aligned}$$

$$= |g(\omega_1)|^2|\xi_1|^2 v_1^{\rm H}Av_1\left[1+\varepsilon_g\left(\frac{\omega_1}{-\omega_2}\frac{\|A\|_2}{v_1^{\rm H}Av_1}\right)^{1/2}\right]^2, \tag{6.53}$$

where the inequality sign at (6.52) holds because

$$\|\hat{x}_2\|_{\widehat{A}}^2 \le \|\widehat{A}\|_2\|\hat{x}_2\|_2^2 = \|V^{\rm H}AV\|_2\sum_{i\ne1}|\xi_i|^2 \le \|A\|_2\frac{\sum_{i\ne1}\omega_i|\xi_i|^2}{\omega_2} = \|A\|_2\frac{\omega_1}{-\omega_2}|\xi_1|^2$$

by (6.49). Thus, from (6.50), (6.51), and (6.53),

$$\boldsymbol{\rho}_g-\lambda_1^+ \le \boldsymbol{\rho}-\lambda_1^+ - \frac{\omega_1}{(\boldsymbol{\rho}-\boldsymbol{\rho}_{g;-})v_1^{\rm H}Av_1}\frac{1-\varepsilon_g^2}{\left[1+\varepsilon_g\left(\frac{\omega_1}{-\omega_2}\frac{\|A\|_2}{v_1^{\rm H}Av_1}\right)^{1/2}\right]^2} \tag{6.54}$$

which gives (6.47) for the case $(\mathrm{typ},\ell)=(+,1)$. □

Lemma 6.6. *Under the conditions of* Lemma 6.5*, we have*

$$|\boldsymbol{\rho}_{g;\mathrm{typ}}-\lambda_\ell^{\mathrm{typ}}| \le \frac{|\omega_1|}{\varsigma_0(v_1)}\varepsilon_g^2 + \frac{1-\varepsilon_g^2}{\varsigma_0(v_1)}\left(3\tau_A^{1/2}+2\chi_1\right)\varepsilon_g|\omega_1|^{3/2}+O(\omega_1^2), \tag{6.55}$$

provided

$$\varepsilon_g|\omega_1|^{1/2}\max\{\tau_A^{1/2},\zeta\chi_1\}<1, \quad 4a(v_1)|\omega_1|<\varsigma_0(v_1)^2, \tag{6.56}$$

where τ_A, τ_B, and τ_C are defined in (6.25)*, and*

$$\chi_1 = \frac{b_0(v_1)^2\tau_B^{1/2}+2a(v_1)c_0(v_1)(\tau_A^{1/2}+\tau_C^{1/2})}{\varsigma_0(v_1)^2}, \tag{6.57}$$

$$\zeta = 4+6\varepsilon_g\omega_1^{1/2}\tau_B^{1/2}+4\varepsilon_g^2\omega_1\tau_B+\varepsilon_g^3\omega_1^{3/2}\tau_B^{3/2}, \tag{6.58}$$

and the shift $\lambda_0\ge\lambda_n^+$ in defining $b_0(\,\cdot\,)$ and $c_0(\,\cdot\,)$ in (6.21)*. Alternatively,*

$$|\boldsymbol{\rho}_{g;\mathrm{typ}}-\lambda_\ell^{\mathrm{typ}}| \le \varepsilon_g^2|\boldsymbol{\rho}_{\mathrm{typ}}-\lambda_\ell^{\mathrm{typ}}|+(1-\varepsilon_g^2)(3\tau_A^{1/2}+2\chi_1)\varepsilon_g|\boldsymbol{\rho}_{\mathrm{typ}}-\lambda_\ell^{\mathrm{typ}}|^{3/2}+O(|\boldsymbol{\rho}_{\mathrm{typ}}-\lambda_\ell^{\mathrm{typ}}|^2), \tag{6.59}$$

provided

$$|\boldsymbol{\rho}_{\mathrm{typ}}-\lambda_\ell^{\mathrm{typ}}| < \max\left\{\frac{\varsigma_0(v_1)}{4a(v_1)}, \frac{1}{\varsigma_0(v_1)\varepsilon_g^2\max\{\tau_A,\zeta^2\chi_1^2\}}\right\}. \tag{6.60}$$

Proof. Consider the case $(\mathrm{typ},\ell)=(+,1)$, and write $\boldsymbol{\rho}=\boldsymbol{\rho}_+$. Without loss of generality, we may assume $\|v_1\|_2=1$. Write $x_g=g(\boldsymbol{Q}(\boldsymbol{\rho}))x$, and

$$t_M=\omega_1^{1/2}\tau_M^{1/2} \quad \text{for } M=A,B,C,$$
$$\boldsymbol{a}=a(v_1), \quad \boldsymbol{b}=b(v_1), \quad \boldsymbol{c}=c(v_1),$$
$$\boldsymbol{b}_0=b_0(v_1), \quad \boldsymbol{c}_0=c_0(v_1).$$

By Lemma 6.5, $\boldsymbol{\rho}_g\le\boldsymbol{\rho}$ (see (6.54)) and

$$\boldsymbol{\rho}_g-\lambda_1^+ \le \delta_0+\delta_1+\delta_2+\delta_3, \tag{6.61}$$

where

$$0 \le \delta_0 = \boldsymbol{\rho} - \lambda_1^+ - \frac{\omega_1}{\varsigma_{\omega_1;0}(v_1)} = O(|\boldsymbol{\rho} - \lambda_1^+|^2) = O(\omega_1^2), \tag{6.62}$$
$$\delta_1 = \frac{\omega_1}{\varsigma_{\omega_1;0}(v_1)} - \frac{\omega_1}{(\boldsymbol{\rho}_g - \boldsymbol{\rho}_{g;-})\boldsymbol{a}},$$
$$\delta_2 = \frac{\omega_1}{(\boldsymbol{\rho}_g - \boldsymbol{\rho}_{g;-})\boldsymbol{a}} - \frac{\omega_1}{(\boldsymbol{\rho} - \boldsymbol{\rho}_{g;-})\boldsymbol{a}},$$
$$\delta_3 = \frac{\omega_1}{(\boldsymbol{\rho} - \boldsymbol{\rho}_{g;-})\boldsymbol{a}} h(\varepsilon_g, \omega_1).$$

The rest of the proof is mainly to estimate δ_1, δ_2, and δ_3.

For δ_2, we have

$$0 \le \delta_2 = \frac{\omega_1}{\boldsymbol{a}} \frac{\boldsymbol{\rho} - \boldsymbol{\rho}_g}{(\boldsymbol{\rho}_g - \boldsymbol{\rho}_{g;-})(\boldsymbol{\rho} - \boldsymbol{\rho}_{g;-})} \le \frac{\omega_1}{\boldsymbol{a}} \frac{\boldsymbol{\rho} - \lambda_1^+}{(\boldsymbol{\rho}_g - \boldsymbol{\rho}_{g;-})(\boldsymbol{\rho} - \boldsymbol{\rho}_{g;-})} = O\left(\omega_1^2\right), \tag{6.63}$$

where we have used (6.44a).

Consider δ_1. If $4\boldsymbol{a}\omega_1 < \boldsymbol{b}^2 - 4\boldsymbol{ac}$ which holds for sufficiently tiny ω_1, then

$$\frac{1}{\varsigma_{\omega_1}(v_1)} = \frac{1}{\sqrt{\boldsymbol{b}^2 - 4\boldsymbol{a}(\boldsymbol{c} - \omega_1)}} = \frac{1}{\sqrt{\boldsymbol{b}^2 - 4\boldsymbol{ac}}} \left[1 - \frac{2\boldsymbol{a}}{\boldsymbol{b}^2 - 4\boldsymbol{ac}}\omega_1 + O(\omega_1^2)\right]. \tag{6.64}$$

By item 2 of Lemma 4.2, any shift $\lambda_0 \ge \lambda_n^+$ makes $\boldsymbol{Q}_{\lambda_0}(\lambda)$ overdamped, i.e., $B_{\lambda_0} \succ 0$ and $C_{\lambda_0} \succeq 0$. It can be verified that

$$\boldsymbol{b}_0^2 - 4\boldsymbol{ac}_0 = \boldsymbol{b}^2 - 4\boldsymbol{ac} = [\varsigma(v_1)]^2.$$

We get, similarly to (6.53),

$$\begin{aligned}
\boldsymbol{a}\,|g(\omega_1)|^2|\xi_1|^2(1 - 2\varepsilon_g t_A) &\le \ x_g^{\mathrm{H}} A x_g \ \le \boldsymbol{a}\,|g(\omega_1)|^2|\xi_1|^2(1 + \varepsilon_g t_A)^2,\\
\boldsymbol{b}_0|g(\omega_1)|^2|\xi_1|^2(1 - 2\varepsilon_g t_B) &\le x_g^{\mathrm{H}} B_{\lambda_0} x_g \le \boldsymbol{b}_0|g(\omega_1)|^2|\xi_1|^2(1 + \varepsilon_g t_B)^2,\\
\boldsymbol{c}_0|g(\omega_1)|^2|\xi_1|^2(1 - 2\varepsilon_g t_C) &\le x_g^{\mathrm{H}} C_{\lambda_0} x_g \le \boldsymbol{c}_0|g(\omega_1)|^2|\xi_1|^2(1 + \varepsilon_g t_C)^2.
\end{aligned}$$

Note that $\boldsymbol{\rho}_g - \lambda_0$ (recalling $\boldsymbol{\rho}_g$ is the shorthand for $\boldsymbol{\rho}_{g;+}$) and $\boldsymbol{\rho}_{g;-} - \lambda_0$ are two distinct roots of $x_g^{\mathrm{H}} A x_g \lambda^2 + x_g^{\mathrm{H}} B_{\lambda_0} x_g \lambda + x_g^{\mathrm{H}} C_{\lambda_0} x_g = 0$ in λ. So

$$\begin{aligned}
\frac{1}{(\boldsymbol{\rho}_g - \boldsymbol{\rho}_{g;-})\boldsymbol{a}} &= \frac{x_g^{\mathrm{H}} A x_g}{\boldsymbol{a}\sqrt{(x_g^{\mathrm{H}} B_{\lambda_0} x_g)^2 - 4(x_g^{\mathrm{H}} A x_g)(x_g^{\mathrm{H}} C_{\lambda_0} x_g)}}\\
&\ge \frac{1 - 2\varepsilon_g t_A}{\sqrt{\boldsymbol{b}_0^2(1 + \varepsilon_g t_B)^4 - 4\boldsymbol{ac}_0(1 - 2\varepsilon_g t_A)(1 - 2\varepsilon_g t_C)}}\\
&= \frac{1 - 2\varepsilon_g t_A}{\sqrt{\boldsymbol{b}_0^2 - 4\boldsymbol{ac}_0 + 4\varepsilon_g(\boldsymbol{b}_0^2 t_B + 2\boldsymbol{ac}_0 t_A + 2\boldsymbol{ac}_0 t_C) + 2\varepsilon_g^2(3\boldsymbol{b}_0^2 t_B^2 - 8\boldsymbol{ac}_0 t_A t_C) + 4\varepsilon_g^3\boldsymbol{b}_0^2 t_B^3 + \varepsilon_g^4 \boldsymbol{b}_0^2 t_B^4}}\\
&= \frac{1 - 2\varepsilon_g t_A}{\sqrt{(\boldsymbol{b}_0^2 - 4\boldsymbol{ac}_0)(1 + 4\varepsilon_g\chi_1\omega_1^{1/2} + 2\varepsilon_g^2\chi_2\omega_1) + 4\varepsilon_g^3\boldsymbol{b}_0^2 t_B^3 + \varepsilon_g^4\boldsymbol{b}_0^2 t_B^4}}\\
&= \frac{1}{\sqrt{\boldsymbol{b}_0^2 - 4\boldsymbol{ac}_0}}(1 - 2\varepsilon_g\omega_1^{1/2}\tau_A^{1/2})\left[1 - 2\varepsilon_g\chi_1\omega_1^{1/2} + \varepsilon_g^2(6\chi_1^2 - \chi_2)\omega_1 + \cdots\right]
\end{aligned} \tag{6.65}$$

$$= \frac{1}{\sqrt{\boldsymbol{b}^2 - 4\boldsymbol{ac}}} \left[1 - 2\varepsilon_g(\tau_A^{1/2} + \chi_1)\omega_1^{1/2} + \varepsilon_g^2(6\chi_1^2 - \chi_2 + 4\tau_A^{1/2}\chi_1)\omega_1 + O(\omega_1^{3/2}) \right], \tag{6.66}$$

where

$$\chi_1 = \frac{\boldsymbol{b}_0^2\tau_B^{1/2} + 2\boldsymbol{ac}_0(\tau_A^{1/2} + \tau_C^{1/2})}{\boldsymbol{b}^2 - 4\boldsymbol{ac}}, \quad \chi_2 = \frac{3\boldsymbol{b}_0^2\tau_B - 8\boldsymbol{ac}_0\tau_A^{1/2}\tau_C^{1/2}}{\boldsymbol{b}^2 - 4\boldsymbol{ac}}.$$

In obtaining (6.65), we need[17] $\zeta\varepsilon_g\chi_1\omega_1^{1/2} < 1$, where $\zeta = 4 + 6\varepsilon_g t_B + 4\varepsilon_g^2 t_B^2 + \varepsilon_g^3 t_B^3$. Using (6.66), we have for δ_1

$$\begin{aligned}
\delta_1 &= \frac{\omega_1}{\varsigma_{\omega_1;0}(v_1)} - \frac{\omega_1}{(\boldsymbol{\rho}_g - \boldsymbol{\rho}_{g;-})\boldsymbol{a}} \\
&= \frac{\omega_1}{\sqrt{\boldsymbol{b}^2 - 4\boldsymbol{ac}}} \left[1 - \frac{2\boldsymbol{a}}{\boldsymbol{b}^2 - 4\boldsymbol{ac}}\omega_1 + O(\omega_1^2) \right] \\
&\quad - \frac{\omega_1}{\sqrt{\boldsymbol{b}^2 - 4\boldsymbol{ac}}} \left[1 - 2\varepsilon_g(\tau_A^{1/2} + \chi_1)\omega_1^{1/2} + \varepsilon_g^2(6\chi_1^2 - \chi_2 + 4\tau_A^{1/2}\chi_1)\omega_1 + O(\omega_1^{3/2}) \right] \\
&= \frac{2\varepsilon_g(\tau_A^{1/2} + \chi_1)\omega_1^{3/2}}{\sqrt{\boldsymbol{b}^2 - 4\boldsymbol{ac}}} + O(\omega_1^2).
\end{aligned} \tag{6.67}$$

Now we turn to δ_3. If $\varepsilon_g t_A < 1$, then

$$\begin{aligned}
h(\varepsilon_g, \omega_1) &= 1 - (1 - \varepsilon_g^2)\left(1 + \varepsilon_g t_A\right)^{-2} \\
&= 1 - (1 - \varepsilon_g^2)(1 - \varepsilon_g t_A + 2\varepsilon_g^2 t_A^2 - 3\varepsilon_g^3 t_A^3 + \cdots) \\
&= \varepsilon_g^2 + (1 - \varepsilon_g^2)(\varepsilon_g t - 2\varepsilon_g^2 t_A^2 + \cdots) \\
&= \varepsilon_g^2 + \varepsilon_g(1 - \varepsilon_g^2)t_A + O\left(t_A^2\right) \\
&= \varepsilon_g^2 + \varepsilon_g(1 - \varepsilon_g^2)\omega_1^{1/2}\tau_A^{1/2} + O\left(\omega_1\right), \\
h(\varepsilon_g, \omega_1) &= 1 - (1 - \varepsilon_g^2)\left(1 + t_A\varepsilon_g\right)^{-2} \\
&\geq 1 - (1 - \varepsilon_g^2) \\
&= \varepsilon_g^2 \geq 0.
\end{aligned}$$

Therefore

$$\begin{aligned}
\delta_3 &= \frac{\omega_1}{(\boldsymbol{\rho} - \boldsymbol{\rho}_{g;-})a} h(\varepsilon_g, \omega_1) \\
&= \frac{\omega_1\varepsilon_g^2 + \varepsilon_g(1 - \varepsilon_g^2)\omega_1^{3/2}\tau_A^{1/2}}{(\boldsymbol{\rho} - \boldsymbol{\rho}_{g;-})a} + O\left(\omega_1^2\right).
\end{aligned} \tag{6.68}$$

[17] For the expansion in (6.65), it is needed that

$$4\varepsilon_g\chi_1\omega_1^{1/2} + 2\varepsilon_g^2\chi_2\omega_1 + \frac{4\varepsilon_g^3\boldsymbol{b}_0^2 t_B^3}{\boldsymbol{b}^2 - 4\boldsymbol{ac}} + \frac{\varepsilon_g^4\boldsymbol{b}_0^2 t_B^4}{\boldsymbol{b}^2 - 4\boldsymbol{ac}} < 1.$$

However,

$$\frac{2\varepsilon_g^2\chi_2\omega_1 + \frac{4\varepsilon_g^3\boldsymbol{b}_0^2 t_B^3}{\boldsymbol{b}^2 - 4\boldsymbol{ac}} + \frac{\varepsilon_g^4\boldsymbol{b}_0^2 t_B^4}{\boldsymbol{b}^2 - 4\boldsymbol{ac}}}{4\varepsilon_g\chi_1\omega_1^{1/2}} \leq \frac{2\varepsilon_g^2 3\boldsymbol{b}_0^2 t_B^2 + 4\varepsilon_g^3\boldsymbol{b}_0^2 t_B^3 + \varepsilon_g^4\boldsymbol{b}_0^2 t_B^4}{4\varepsilon_g\boldsymbol{b}_0^2 t_B} = \frac{\varepsilon_g t_B}{4}(6 + 4\varepsilon_g t_B + \varepsilon_g^2 t_B^2).$$

We have finished estimating δ_i for $i = 0, 1, 2, 3$. Now, combine (6.61), (6.62), (6.63), (6.67), and (6.68) to get

$$\begin{aligned}
\boldsymbol{\rho}_g - \lambda_1^+ &\le \frac{2\varepsilon_g(\tau_A^{1/2} + \chi_1)\omega_1^{3/2}}{\sqrt{\boldsymbol{b}^2 - 4\boldsymbol{ac}}} + \frac{\omega_1\varepsilon_g^2 + \varepsilon_g(1-\varepsilon_g^2)\omega_1^{3/2}\tau_A^{1/2}}{(\boldsymbol{\rho} - \boldsymbol{\rho}_{g;-})\boldsymbol{a}} + O(\omega_1^2) \\
&= \frac{\varepsilon_g^2}{(\boldsymbol{\rho} - \boldsymbol{\rho}_{g;-})\boldsymbol{a}}\omega_1 + \left(\frac{2(\tau_A^{1/2} + \chi_1)}{\sqrt{\boldsymbol{b}^2 - 4\boldsymbol{ac}}} + \frac{(1-\varepsilon_g^2)\tau_A^{1/2}}{(\boldsymbol{\rho} - \boldsymbol{\rho}_{g;-})\boldsymbol{a}}\right)\varepsilon_g\omega_1^{3/2} + O(\omega_1^2),
\end{aligned}$$

which, along with

$$\frac{1}{(\boldsymbol{\rho} - \boldsymbol{\rho}_{g;-})\boldsymbol{a}} = \frac{1}{(\boldsymbol{\rho}_g - \boldsymbol{\rho}_{g;-})\boldsymbol{a}} - \frac{\delta_2}{\omega_1} = \frac{1}{\sqrt{\boldsymbol{b}^2 - 4\boldsymbol{ac}}}\left[1 - 2\varepsilon_g(\tau_A^{1/2} + \chi_1)\omega_1^{1/2}\right] + O(\omega_1),$$

yield (6.55). Use (6.64) to see

$$\frac{1}{\varsigma_0(v_1)} = \frac{1}{\varsigma_{\omega_1;0}(v_1)}\left[1 + \frac{2\boldsymbol{a}}{\boldsymbol{b}^2 - 4\boldsymbol{ac}}\omega_1 + O(\omega_1^2)\right]$$

substituting which and (6.44a) into (6.55) to get (6.59). $\square$

We are now ready to prove Theorem 6.2.

Proof of Theorem 6.2. Item 1 is a direct consequence of item 4 of Theorem 6.1.

Item 2 is a consequence of Lemma 6.6 upon letting g be the minimizer that gives the minimal ε_m and using $|\boldsymbol{\rho}_{i+1} - \lambda_\ell^{\text{typ}}| \le |\boldsymbol{\rho}_g - \lambda_\ell^{\text{typ}}|$.

For item 3, again let g be the minimizer that gives the minimal ε_m. As $i \to \infty$ in item 2, we have $\omega_1 \to 0$, $\omega_2 \to \gamma$, and $v_1 \to u_\ell^{\text{typ}}$ in direction, and thus

$$\lim_{i\to\infty} \eta(v_1) = \lim_{i\to\infty} 3\tau_A^{1/2} + 2\frac{(b_0(v_1))^2\tau_{B_{\lambda_0}}^{1/2} + 2a(v_1)c_0(v_1)(\tau_A^{1/2} + \tau_{C_{\lambda_0}}^{1/2})}{\varsigma_0(v_1)^2} = \eta$$

as given by (6.29). Now let

$$\hat{g}(t) = \mathscr{T}_{m-1}\left(\frac{2t - (\omega_n + \omega_2)}{\omega_n - \omega_2}\right)\Big/\mathscr{T}_{m-1}\left(-\frac{1+\hat{\kappa}}{1-\hat{\kappa}}\right), \quad \hat{\kappa} = \frac{\omega_2 - \omega_1}{\omega_n - \omega_1},$$

where $\mathscr{T}_{m-1}(t)$ is the $(m-1)$st Chebyshev polynomial of the first kind. Then [34, section 2]

$$\varepsilon_m \le \varepsilon_{\hat{g}} \le \max_{\omega_2 \le t \le \omega_n} |\hat{g}(t)| = 2\left[\left(\frac{1+\sqrt{\hat{\kappa}}}{1-\sqrt{\hat{\kappa}}}\right)^{m-1} + \left(\frac{1+\sqrt{\hat{\kappa}}}{1-\sqrt{\hat{\kappa}}}\right)^{-(m-1)}\right]^{-1}$$

which goes to ε as $i \to \infty$ because $\hat{\kappa} \to \kappa$. $\square$

7 Preconditioned steepest descent/ascent method

7.1 Preconditioning

We will explain the idea of preconditioning for computing (λ_1^+, u_1^+) only, via two different points of view. The same argument can be made for other extreme pos- and neg-quadratic eigenpairs.

It is well-known that when the contours of the objective function near its optimum are extremely elongated, at each step of the conventional steepest descent/ascent method, following the search direction which is the opposite of the gradient gets closer to the optimum on the line for a very short while and then starts to get away because the direction doesn't point "towards the optimum", resulting in a long zigzag path of a large number of steps. The ideal search direction p is therefore the one such that with its starting point at $\boldsymbol{x}$, p points to the optimum, i.e., the optimum is on the line $\{\boldsymbol{x} + tp \,:\, t \in \mathbb{C}\}$. Specifically, expand $\boldsymbol{x}$ as a linear combination of u_j^+

$$\boldsymbol{x} = \sum_{j=1}^{n} \alpha_j u_j^+ =: \alpha_1 u_1^+ + \boldsymbol{v}, \quad \boldsymbol{v} = \sum_{j=2}^{n} \alpha_j u_j^+. \tag{7.1}$$

Then the ideal search direction is

$$p = \alpha u_1^+ + \beta \boldsymbol{v}$$

for some scalar α and $\beta \neq 0$ such that $\alpha_1\beta - \alpha \neq 0$ (otherwise $p = \beta\boldsymbol{x}$). Of course, this is impractical because we don't know u_1^+ and $\boldsymbol{v}$. But we can construct one that is close to it. One such p is

$$p = [\boldsymbol{Q}(\sigma)]^{-1}\, r_+(\boldsymbol{x}) = [\boldsymbol{Q}(\sigma)]^{-1}\boldsymbol{Q}(\boldsymbol{\rho}_+)\boldsymbol{x},$$

where $\boldsymbol{\rho}_+ = \rho_+(\boldsymbol{x})$ and[18] σ is some shift near λ_1^+ but not equal to $\boldsymbol{\rho}_+$. Let us analyze this p. By (2.17a), we have

$$[\boldsymbol{Q}(\sigma)]^{-1}\boldsymbol{Q}(\boldsymbol{\rho}_+) = U_+(\sigma I - \Lambda_+)^{-1}(U_-^{\mathrm{H}}AU_+)^{-1}(\sigma I - \Lambda_-)^{-1}(\boldsymbol{\rho}_+ I - \Lambda_-)U_-^{\mathrm{H}}AU_+(\boldsymbol{\rho}_+ I - \Lambda_+)U_+^{-1}.$$

Suppose now that both σ and $\boldsymbol{\rho}_+$ are near λ_1^+. Then

$$(\sigma I - \Lambda_-)^{-1}(\boldsymbol{\rho}_+ I - \Lambda_-) = I + (\boldsymbol{\rho}_+ - \sigma)(\sigma I - \Lambda_-)^{-1} \approx I.$$

Therefore $[\boldsymbol{Q}(\sigma)]^{-1}\boldsymbol{Q}(\boldsymbol{\rho}_+) \approx U_+(\sigma I - \Lambda_+)^{-1}(\boldsymbol{\rho}_+ I - \Lambda_+)U_+^{-1}$, and thus

$$p = [\boldsymbol{Q}(\sigma)]^{-1}\boldsymbol{Q}(\boldsymbol{\rho}_+)\boldsymbol{x} \approx \sum_{j=1}^{n} \mu_j \alpha_j u_j^+, \quad \mu_j := \frac{\lambda_j^+ - \boldsymbol{\rho}_+}{\lambda_j^+ - \sigma}. \tag{7.2}$$

Now if $\lambda_1^+ \leq \boldsymbol{\rho}_+ < \lambda_2^+$ and if the gap $\lambda_2^+ - \lambda_1^+$ is reasonably modest, then

$$\mu_j \approx 1 \quad \text{for } j > 1$$

to give a $p \approx \alpha u_1^+ + \boldsymbol{v}$, resulting in fast convergence. This rough but intuitive analysis suggests that $K = [\boldsymbol{Q}(\sigma)]^{-1}$ with a suitably chosen shift σ can be used to serve as a

[18] We reasonably assume also $\sigma \neq \lambda_j^+$ for all j, too.

good preconditioner to improve the steepest descent/ascent method – Algorithm 6.1 by simply modifying $Y_i = [\boldsymbol{x}_i, \boldsymbol{r}_i]$ at Line 6 there to $Y_i = [\boldsymbol{x}_i, K\boldsymbol{r}_i]$. We caution the reader that implementing $K\boldsymbol{r}_i$ is amount to solving a linear system. This is usually done approximately by, e.g., some iterative methods such as the linear conjugate gradient method, MINRES [11, 17, 19].

The second view point is similar to the one proposed by Golub and Ye [18] for the generalized linear eigenvalue problem. Theorem 6.2 reveals that the rates of convergence for Algorithms 6.1 and 6.2 depend on the distribution of the eigenvalues ω_j of $\boldsymbol{Q}(\boldsymbol{\rho}_i)$, not the quadratic eigenvalues of of $\boldsymbol{Q}(\lambda)$. In particular, if all $\omega_2 = \cdots = \omega_n$, then $\epsilon_m = 0$ for $m \geq 2$ and thus

$$\boldsymbol{\rho}_{i+1} - \lambda_1^+ = \mathcal{O}(|\boldsymbol{\rho}_i - \lambda_1^+|^2),$$

suggesting quadratic convergence. Such an extreme case, though highly welcome, is unlikely to happen in practice, but it gives us an idea that if somehow we could transform an eigenvalue problem towards such an extreme case, the transformed problem would be easier to solve. Specifically we should seek equivalent transformations that change the eigenvalues of $\boldsymbol{Q}(\boldsymbol{\rho}_i)$ as much as possible to,

$$\boxed{\text{one isolated eigenvalue } \omega_1, \text{ and the rest } \omega_j \ (2 \leq j \leq n) \text{ tightly clustered,}} \tag{7.3}$$

but leave the quadratic eigenvalues of $\boldsymbol{Q}(\lambda)$ unchanged.

We would like to equivalently transform the QEP for $\boldsymbol{Q}(\lambda)$ to for $L^{-1}\boldsymbol{Q}(\lambda)L^{-\mathrm{H}}$ by some nonsingular L (whose inverse or any linear system with L is easy to solve) so that the eigenvalues of $L^{-1}\boldsymbol{Q}(\boldsymbol{\rho}_i)L^{-\mathrm{H}}$ distribute more or less like (7.3). Then apply one step of Algorithm 6.1 or 6.2 to the pencil $L^{-1}\boldsymbol{Q}(\lambda)L^{-\mathrm{H}}$ to find the next approximation $\boldsymbol{\rho}_{i+1}$. The process repeats, i.e., find a new L to transform the problem and apply one step of Algorithm 6.1 or 6.2 to the transformed problem.

Such an L may be constructed using the LDL^{H} decomposition of $\boldsymbol{Q}(\boldsymbol{\rho}_i)$ [17, p.139] if the decomposition exists: $\boldsymbol{Q}(\boldsymbol{\rho}_i) = LDL^{\mathrm{H}}$, where L is lower triangular and $D = \mathrm{diag}(\pm 1)$. Then $L^{-1}\boldsymbol{Q}(\boldsymbol{\rho}_i)L^{-\mathrm{H}} = D$ has the ideal eigenvalue distribution that gives $\epsilon_m = 0$ for any $m \geq 2$. Unfortunately, this simple solution is impractical in practice for the following reasons:

1. The decomposition may not exist at all. In theory, the decomposition exists if all the leading principle submatrices of $\boldsymbol{Q}(\boldsymbol{\rho}_i)$ are nonsingular.

2. If the decomposition does exist, it may not be numerically stable to compute, especially when $\boldsymbol{\rho}_i$ comes closer and closer to λ_1^+.

3. The sparsity in $\boldsymbol{Q}(\boldsymbol{\rho}_i)$ is most likely destroyed, leaving L significantly denser than $\boldsymbol{Q}(\boldsymbol{\rho}_i)$. This makes all ensuing computations much more expensive.

A more practical solution is, however, through an incomplete LDL^{H} factorization (see [51, Chapter 10]), to get

$$\boldsymbol{Q}(\boldsymbol{\rho}_i) \approx LDL^{\mathrm{H}},$$

where "≈" includes not only the usual "approximately equal", but also the case when $\boldsymbol{Q}(\boldsymbol{\rho}_i) - LDL^{\mathrm{H}}$ is approximately a low rank matrix, and $D = \mathrm{diag}(\pm 1)$. Such an L changes from one step of the algorithm to another. In practice, often we may use one fixed preconditioner for all or a number of consecutive iterative steps. Using a constant preconditioner is certainly not optimal: it likely doesn't give the best rate of convergence per step and thus increases the number of total iterative steps but it may reduce overall cost because it saves work in preconditioner constructions and thus reduces cost per step. The basic idea of using a step-independent preconditioner is to find a σ that is close to λ_1^+, and perform an incomplete LDL^{H} decomposition:

$$\boldsymbol{Q}(\sigma) \approx LDL^{\mathrm{H}}$$

and transform $\boldsymbol{Q}(\lambda)$ accordingly before applying Algorithm 6.1 or 6.2. Now the rate of convergence is determined by the eigenvalues of

$$L^{-1}\boldsymbol{Q}(\boldsymbol{\rho}_i)L^{-\mathrm{H}} = L^{-1}\boldsymbol{Q}(\sigma)L^{-\mathrm{H}} + (\boldsymbol{\rho}_i - \sigma)L^{-1}\boldsymbol{Q}'(\sigma)L^{-\mathrm{H}} + O\Big(|\boldsymbol{\rho}_i - \sigma|^2\Big)$$

which would have a better spectral distribution so long as the last two terms is small relative to $L^{-1}\boldsymbol{Q}(\boldsymbol{\rho}_i)L^{-\mathrm{H}}$. When $\lambda_n^- < \sigma < \lambda_1^+$, $-\boldsymbol{Q}(\sigma) \succ 0$ and the incomplete LDL^{H} factorization becomes incomplete Cholesky factorization.

7.2 Preconditioned steepest descent/ascent method

We have insisted so far about applying Algorithm 6.1 or 6.2 straightforwardly to the transformed problem. There is another way, perhaps, better: only symbolically applying Algorithm 6.1 or 6.2 to the transformed problem as a derivation tool for a preconditioned method that always projects the original pencil $\boldsymbol{Q}(\lambda)$ directly every step. The only difference is now the projecting subspaces are preconditioned. Again we will explain it for the case of computing the first pos-type quadratic eigenpair (λ_1^+, u_1^+).

Suppose $\boldsymbol{Q}(\lambda)$ is transformed to $\widehat{\boldsymbol{Q}}(\lambda) := L^{-1}\boldsymbol{Q}(\lambda)L^{-\mathrm{H}}$. Consider a typical step of Algorithm 6.2 applied to $\widehat{\boldsymbol{Q}}(\lambda)$. For the purpose of distinguishing notational symbols, we will put hats on all those for $\widehat{\boldsymbol{Q}}(\lambda)$. The typical step of Algorithm 6.2 on $\widehat{\boldsymbol{Q}}$ is

$$\boxed{\begin{array}{l}\text{compute the smallest pos-type quadratic eigenvalue } \mu \text{ and corresponding} \\ \text{quadratic eigenvector } \hat{v} \text{ of } \hat{Z}^{\mathrm{H}}\widehat{\boldsymbol{Q}}(\lambda)\hat{Z}, \text{ where } \hat{Z} \in \mathbb{C}^{n\times m} \text{ is a basis matrix} \\ \text{of Krylov subspace } \mathcal{K}_m(\widehat{\boldsymbol{Q}}(\hat{\boldsymbol{\rho}}), \hat{\boldsymbol{x}}).\end{array}} \tag{7.4}$$

Notice $\Big[\widehat{\boldsymbol{Q}}(\hat{\boldsymbol{\rho}})\Big]^j \hat{\boldsymbol{x}} = L^{\mathrm{H}}\left[(LL^{\mathrm{H}})^{-1}\boldsymbol{Q}(\hat{\boldsymbol{\rho}})\right]^j (L^{-\mathrm{H}}\hat{\boldsymbol{x}})$ to see

$$L^{-\mathrm{H}} \cdot \mathcal{K}_m(\widehat{\boldsymbol{Q}}(\hat{\boldsymbol{\rho}}), \hat{\boldsymbol{x}}) = \mathcal{K}_m(\,K\boldsymbol{Q}(\hat{\boldsymbol{\rho}}), \boldsymbol{x}),$$

where $\boldsymbol{x} = L^{-\mathrm{H}}\hat{\boldsymbol{x}}$ and $K = (LL^{\mathrm{H}})^{-1}$. So $Z = L^{-\mathrm{H}}\hat{Z}$ is a basis matrix of Krylov subspace $\mathcal{K}_m(\,K\boldsymbol{Q}(\hat{\boldsymbol{\rho}}), \boldsymbol{x})$. Since also

$$\hat{Z}^{\mathrm{H}}\widehat{\boldsymbol{Q}}(\lambda)\hat{Z} = (L^{-\mathrm{H}}\hat{Z})^{\mathrm{H}}\boldsymbol{Q}(\lambda)(L^{-\mathrm{H}}\hat{Z}),$$
$$\hat{\boldsymbol{\rho}} = \hat{\rho}_+(\hat{\boldsymbol{x}}) = \rho_+(\boldsymbol{x}) = \boldsymbol{\rho},$$

the typical step (7.4) can be reformulated equivalently to

$$\boxed{\begin{array}{l}\text{compute the smallest pos-type quadratic eigenvalue } \mu \text{ and corresponding} \\ \text{quadratic eigenvector } v \text{ of } Z^{\mathrm{H}}\boldsymbol{Q}(\lambda)Z, \text{ where } Z \in \mathbb{C}^{n\times m} \text{ is a basis matrix} \\ \text{of Krylov subspace } \mathcal{K}_m(\,K\boldsymbol{Q}(\boldsymbol{\rho}), \boldsymbol{x}), \text{ where } K = (LL^{\mathrm{H}})^{-1}.\end{array}} \tag{7.5}$$

Algorithm 7.1 Preconditioned extended steepest descent/ascent method

Given an initial approximation $\boldsymbol{x}_0$ to $u_\ell^{\rm typ}$, and a relative tolerance `rtol`, and the search space dimension m, the algorithm computes an approximate pair to $(\lambda_\ell^{\rm typ}, u_\ell^{\rm typ})$ with the prescribed `rtol`.

$\boldsymbol{x}_0 = \boldsymbol{x}_0/\|\boldsymbol{x}_0\|$, $\boldsymbol{\rho}_0 = \rho_{\rm typ}(\boldsymbol{x}_0)$, $\boldsymbol{r}_0 = r_{\rm typ}(\boldsymbol{x}_0)$;
for $i = 0, 1, \dots$ **do**
  if $\|\boldsymbol{r}_i\|/(|\boldsymbol{\rho}_i|^2\|A\boldsymbol{x}_i\| + |\boldsymbol{\rho}_i|\,\|B\boldsymbol{x}_i\| + \|C\boldsymbol{x}_i\|) \le$ `rtol` **then**
    BREAK;
  else
    construct a preconditioner K_i;
    compute a basis matrix Y_i for the Krylov subspace $\mathcal{K}_m(K_i\boldsymbol{Q}(\boldsymbol{\rho}_i), \boldsymbol{x}_i)$;
    solve HQEP for $Y_i^{\rm H}\boldsymbol{Q}(\lambda)Y_i$ to get its quadratic eigenvalues $\mu_j^{\pm}$ as in (6.11) and quadratic eigenvectors $y_j^{\pm}$;
    select the next approximate quadratic eigenpair $(\mu, y) = (\mu_j^{\rm typ}, Yy_j^{\rm typ})$ according to the table in (6.12);
    $\boldsymbol{x}_{i+1} = y/\|y\|$, $\boldsymbol{\rho}_{i+1} = \mu$, $\boldsymbol{r}_{i+1} = r_{\rm typ}(\boldsymbol{x}_{i+1})$;
  end if
end for
return $(\boldsymbol{\rho}_i, \boldsymbol{x}_i)$ as an approximate quadratic eigenpair to $(\lambda_\ell^{\rm typ}, u_\ell^{\rm typ})$.

We are now ready to state a version of the *preconditioned extended steepest descent/ascent method*. To make it be inclusive, in Algorithm 7.1 we use K_i to denote the preconditioner at the ith iterative step. Once again, they may all be the same or vary from one iterative step to another. Although the derivation of this algorithm was for the preconditioners obtained from the second view point above, its final form includes the preconditioners from the first view point.

7.3 Convergence analysis

If $K_i \succ 0$, the ith iterative step of Algorithm 7.1 is just one step of the extended steepest descent/ascent method applied to $K_i^{1/2}\boldsymbol{Q}(\lambda)K_i^{1/2}$. Therefore Theorem 6.2 implies the following theorem for Algorithm 7.1.

Theorem 7.1. *Suppose* $\lambda_1^{\rm typ} \le \boldsymbol{\rho}_0 < \lambda_2^{\rm typ}$ *if* $\ell = 1$ *or* $\lambda_{n-1}^{\rm typ} < \boldsymbol{\rho}_0 \le \lambda_n^{\rm typ}$ *if* $\ell = n$, *and let the sequences* $\{\boldsymbol{\rho}_i\}, \{\boldsymbol{r}_i\}, \{\boldsymbol{x}_i\}$ *be produced by* Algorithm 7.1. *Suppose* $K_i \succ 0$.

1. *As* $i \to \infty$, $\boldsymbol{\rho}_i$ *monotonically converges to* $\hat\rho = \lambda_\ell^{\rm typ}$, *and* $\boldsymbol{x}_i$ *converges to* $u_\ell^{\rm typ}$ *in direction, i.e.,* $\theta(\boldsymbol{x}_i, u_\ell^{\rm typ}) \to 0$.

2. *The eigenvalues*[19] ω_j *of* $K_i\boldsymbol{Q}(\boldsymbol{\rho}_i)$ *can be ordered as*

$$\omega_1 > 0 > \omega_2 \ge \cdots \ge \omega_n \quad \textit{if}\ ({\rm typ}, \ell) \in \{(+,1), (-,n)\}, \quad \textit{or}, \tag{7.6a}$$
$$\omega_1 < 0 < \omega_2 \le \cdots \le \omega_n \quad \textit{if}\ ({\rm typ}, \ell) \in \{(+,n), (-,1)\}. \tag{7.6b}$$

[19]Their dependency upon i is suppressed for clarity.

If $\boldsymbol{\rho}_i$ *is sufficiently close to* $\lambda_\ell^{\rm typ}$*, then*

$$|\boldsymbol{\rho}_{i+1}-\lambda_\ell^{\rm typ}|\le\varepsilon_m^2|\boldsymbol{\rho}_i-\lambda_\ell^{\rm typ}|+O\Big(\varepsilon_m|\boldsymbol{\rho}_i-\lambda_\ell^{\rm typ}|^{3/2}+|\boldsymbol{\rho}_i-\lambda_\ell^{\rm typ}|^2\Big),\tag{7.7}$$

where ε_m *is defined as in* (6.24)*.*

3. *Denote*[20] *by* γ *and* Γ *the smallest and largest positive eigenvalue of*

$$\begin{cases}-K_i\boldsymbol{Q}(\lambda_\ell^{\rm typ}) & \textit{for } ({\rm typ},\ell)\in\{(+,1),(-,n)\},\\ \ \ K_i\boldsymbol{Q}(\lambda_\ell^{\rm typ}) & \textit{for } ({\rm typ},\ell)\in\{(+,n),(-,1)\}.\end{cases}$$

If $\boldsymbol{\rho}_i$ *is sufficiently close to* $\lambda_\ell^{\rm typ}$*, then*

$$|\boldsymbol{\rho}_{i+1}-\lambda_\ell^{\rm typ}|\le\varepsilon^2|\boldsymbol{\rho}_i-\lambda_\ell^{\rm typ}|+O\Big(\varepsilon|\boldsymbol{\rho}_i-\lambda_\ell^{\rm typ}|^{3/2}+|\boldsymbol{\rho}_i-\lambda_\ell^{\rm typ}|^2\Big),\tag{7.8}$$

where ε *is defined as in* (6.28)*.*

There is a convergence rate estimate, essentially due to Samokish [52, 1958], for the preconditioned steepest descent/ascent method in the case of the standard Hermitian eigenvalue problem. The reader is referred to [29, 46] for detail. Theorem 7.2 below is an extension of Samokish's result for our case.

Theorem 7.2. *Suppose* $K\succ 0$*. Let* $\ell\in\{1,n\}$ *and* ${\rm typ},{\rm typ}'\in\{+,-\}$ *such that* typ *and* typ′ *are opposite, and denote by* γ *and* Γ *the smallest and largest positive eigenvalue of*

$$\begin{cases}-K\boldsymbol{Q}(\lambda_\ell^{\rm typ}) & \textit{for } ({\rm typ},\ell)\in\{(+,1),(-,n)\},\\ \ \ K\boldsymbol{Q}(\lambda_\ell^{\rm typ}) & \textit{for } ({\rm typ},\ell)\in\{(+,n),(-,1)\},\end{cases}$$

and

$$\tau=\frac{2}{\gamma+\Gamma},\quad \kappa=\frac{\Gamma}{\gamma},\quad \varepsilon=\frac{\kappa-1}{\kappa+1}.$$

Let argopt *be as given in* (6.6)*, and*

$$\begin{aligned}t_{\rm opt}&=\operatorname*{argopt}_{t\in\mathbb{C}}\rho_{\rm typ}(x+tKr_{\rm typ}(x)),\quad y=x+t_{\rm opt}Kr_{\rm typ}(x),\\ z&=\begin{cases}x+\tau Kr_\pm(x) & \textit{for } ({\rm typ},\ell)\in\{(+,1),(-,n)\},\\ x-\tau Kr_\pm(x) & \textit{for } ({\rm typ},\ell)\in\{(+,n),(-,1)\}.\end{cases}\end{aligned}$$

We have

$$\begin{aligned}&|\rho_{\rm typ}(y)-\lambda_\ell^{\rm typ}|\le|\rho_{\rm typ}(z)-\lambda_\ell^{\rm typ}|\\ &\qquad\le\frac{1}{|\lambda_\ell^{\rm typ}-\rho_{\rm typ'}(z)|}\left[\frac{\varepsilon\sqrt{|\lambda_\ell^{\rm typ}-\rho_{\rm typ'}(x)|}+\tau\sqrt{\Gamma}\,\delta_1}{1-\tau\left(\sqrt{\Gamma}\delta_2+\delta_3^2\right)}\right]^2|\rho_{\rm typ}(x)-\lambda_\ell^{\rm typ}|,\end{aligned}\tag{7.9}$$

[20] It is worth emphasizing that $K_i\boldsymbol{Q}(\lambda_\ell^{\rm typ})$ is singular and, by Theorem 2.1, $K_i^{1/2}\boldsymbol{Q}(\lambda_\ell^{\rm typ})K_i^{1/2}$ is negative semidefinite if $({\rm typ},\ell)\in\{(+,1),(-,n)\}$ and positive semidefinite if $({\rm typ},\ell)\in\{(+,n),(-,1)\}$.

provided $\tau\left(\sqrt{\Gamma}\delta_2+\delta_3^2\right)<1$, *where*

$$\begin{aligned}
\delta_1 &= \sqrt{|\rho_{\rm typ}(x)-\lambda_\ell^{\rm typ}|}\,\|K^{1/2}\{A[\rho_{\rm typ}(x)+\lambda_\ell^{\rm typ}]+B\}A^{-1/2}\|_2,\\
\delta_2 &= \sqrt{\|K^{1/2}AK^{1/2}\|_2\,|\rho_{\rm typ}(x)-\lambda_\ell^{\rm typ}|\cdot|\lambda_\ell^{\rm typ}-\rho_{\rm typ'}(x)|},\\
\delta_3 &= \sqrt{\|A^{1/2}K\{A[\rho_{\rm typ}(x)+\lambda_\ell^{\rm typ}]+B\}A^{-1/2}\|_2\,|\rho_{\rm typ}(x)-\lambda_\ell^{\rm typ}|}.
\end{aligned}$$

Proof. We will prove the case: $(\text{typ},\ell)=(+,1)$ only. The other cases can be handled in the same way.

Note $z=x+\tau Kr_+(x)=x+\tau K\boldsymbol{Q}(\rho_+(x))\,x$. We have $\lambda_1^+\le\rho_+(y)\le\rho_+(z)$ and thus $\rho_+(y)-\lambda_1^+\le\rho_+(z)-\lambda_1^+$. So it remains to show that $\rho_+(z)-\lambda_1^+$ is no bigger than the right-hand side of (7.9).

Let $M=-\boldsymbol{Q}(\lambda_1^+)\succeq 0$. For any vector w, we have

$$\begin{aligned}
\|w\|_M^2 &= -w^{\rm H}\boldsymbol{Q}(\lambda_1^+)w\\
&= [\rho_+(w)-\lambda_1^+][\lambda_1^+-\rho_-(w)]\|w\|_A^2, \qquad (7.10)\\
\|[I+\tau K\boldsymbol{Q}(\lambda_1^+)]w\|_M &= \|[I-\tau KM]w\|_M\\
&\le \varepsilon\|w\|_M. \qquad (7.11)
\end{aligned}$$

Write

$$\begin{aligned}
z &= [I+\tau K\boldsymbol{Q}(\lambda_1^+)]x-\tau K[\boldsymbol{Q}(\lambda_1^+)-\boldsymbol{Q}(\rho_+(x))]x\\
&= [I+\tau K\boldsymbol{Q}(\lambda_1^+)]x+\tau[\rho_+(x)-\lambda_1^+]K[A(\rho_+(x)+\lambda_1^+)+B]x.
\end{aligned}$$

Without loss of generality, we may assume $\|x\|_A=1$. We have

$$\begin{aligned}
\|z\|_M &= \sqrt{[\rho_+(z)-\lambda_1^+][\lambda_1^+-\rho_-(z)]}\,\|z\|_A, \qquad \text{by (7.10)}\\
\|z\|_M &\le \|[I+\tau K\boldsymbol{Q}(\lambda_1^+)]x\|_M+\tau[\rho_+(x)-\lambda_1^+]\|K[A(\rho_+(x)+\lambda_1^+)+B]x\|_M\\
&\le \varepsilon\|x\|_M+\tau[\rho_+(x)-\lambda_1^+]\sqrt{\Gamma}\,\|[A(\rho_+(x)+\lambda_1^+)+B]x\|_K\\
&\le \varepsilon\sqrt{[\rho_+(x)-\lambda_1^+][\lambda_1^+-\rho_-(x)]}\\
&\quad +\tau[\rho_+(x)-\lambda_1^+]\sqrt{\Gamma}\,\|K^{1/2}[A(\rho_+(x)+\lambda_1^+)+B]A^{-1/2}\|_2\\
&= \left[\varepsilon\sqrt{\lambda_1^+-\rho_-(x)}+\tau\sqrt{\Gamma}\,\delta_1\right]\sqrt{\rho_+(x)-\lambda_1^+}, \qquad (7.12)\\
\|z\|_A &\ge \|x\|_A-\tau\|Kr_+(x)\|_A\\
&= 1-\tau\|Kr_+(x)\|_A,\\
\|Kr_+(x)\|_A &= \|K\boldsymbol{Q}(\lambda_1^+)x-K[\boldsymbol{Q}(\lambda_1^+)-\boldsymbol{Q}(\rho_+(x))]x\|_A\\
&\le \|K\boldsymbol{Q}(\lambda_1^+)x\|_A+[\rho_+(x)-\lambda_1^+]\|K[A(\rho_+(x)+\lambda_1^+)+B]x\|_A\\
&\le \sqrt{\|K^{1/2}AK^{1/2}\|_2\Gamma}\,\|x\|_M\\
&\quad +[\rho_+(x)-\lambda_1^+]\|A^{1/2}K[A(\rho_+(x)+\lambda_1^+)+B]A^{-1/2}\|_2\|x\|_A\\
&= \sqrt{\Gamma}\delta_2+\delta_3^2. \qquad (7.13)
\end{aligned}$$

Finally use

$$\rho_+(z) - \lambda_1^+ = \frac{\|z\|_M^2}{[\lambda_1^+ - \rho_-(z)]\|z\|_A^2} \le \frac{\|z\|_M^2}{[\lambda_1^+ - \rho_-(z)] \cdot [1 - \tau\|Kr_+(x)\|_A]^2}$$

and (7.12) and (7.13) to complete the proof. □

Similarly, we have the following result for Algorithm 7.1.

Theorem 7.3. *Suppose* $K \succ 0$. *Let* $\ell \in \{1, n\}$ *and* $\mathrm{typ}, \mathrm{typ}' \in \{+, -\}$ *such that* typ *and* typ′ *are opposite, and let* γ *and* Γ *be the ones in* Theorem 7.2, *and*

$$\varepsilon = 2\left[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^{m-1} + \left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^{-(m-1)}\right]^{-1}, \qquad \kappa = \frac{\Gamma}{\gamma}.$$

Let argopt *be as given in* (6.6), *and*

$$\begin{aligned} g_{\mathrm{opt}} &= \operatorname*{argopt}_{g \in \mathbb{P}_{m-1}, g(0)=1} \rho_{\mathrm{typ}}(g(K\boldsymbol{Q}(\rho_{\mathrm{typ}}(x))x), \\ y &= g_{\mathrm{opt}}(K\boldsymbol{Q}(\rho_{\mathrm{typ}}(x))x, \\ z &= \hat{g}(K\boldsymbol{Q}(\rho_{\mathrm{typ}}(x))x, \end{aligned}$$

where

$$\begin{aligned} \hat{g}(t) &= \mathscr{T}_{m-1}\left(\frac{2t - (\Gamma + \gamma)}{\Gamma - \gamma}\right) \Big/ \mathscr{T}_{m-1}\left(-\frac{1+\kappa}{1-\kappa}\right) \\ &= 1 + c_1 t + \cdots + c_{m-1}t^{m-1} \end{aligned}$$

since $\hat{g}(0) = 1$, *and* $\mathscr{T}_{m-1}(t)$ *is the* $(m-1)$*st Chebyshev polynomial of the first kind. We have*

$$\begin{aligned} |\rho_{\mathrm{typ}}(y) - \lambda_\ell^{\mathrm{typ}}| &\le |\rho_{\mathrm{typ}}(z) - \lambda_\ell^{\mathrm{typ}}| \\ &\le \frac{1}{|\lambda_\ell^{\mathrm{typ}} - \rho_{\mathrm{typ}'}(z_{\mathrm{typ}'})|}\left[\frac{\varepsilon\sqrt{|\lambda_\ell^{\mathrm{typ}} - \rho_{\mathrm{typ}'}(x)|} + \eta\sqrt{|\rho_{\mathrm{typ}}(x) - \lambda_\ell^{\mathrm{typ}}|}}{1 - \eta|\rho_{\mathrm{typ}}(x) - \lambda_\ell^{\mathrm{typ}}|}\right]^2 \\ &\quad \times |\rho_{\mathrm{typ}}(x) - \lambda_\ell^{\mathrm{typ}}|, \end{aligned} \tag{7.14}$$

provided

$$\begin{aligned} \eta :&= \sum_{i=1}^{m-1} |c_i| \cdot \|K[A\rho_+(x) + \lambda_1^+] + B\|_2 \sum_{j=0}^{i-1} \|K\boldsymbol{Q}(\lambda_1^+)\|_2^{i-j-1} \|K\boldsymbol{Q}(\rho_+(x))\|^j \\ &< \frac{1}{|\rho_{\mathrm{typ}}(x) - \lambda_\ell^{\mathrm{typ}}|}. \end{aligned}$$

Proof. We will prove the case: $(\mathrm{typ}, \ell) = (+, 1)$ only. The other cases can be handled in the same way.

We have $\lambda_1^+ \le \rho_+(y) \le \rho_+(z)$ and thus $\rho_+(y) - \lambda_1^+ \le \rho_+(z) - \lambda_1^+$. So it suffices to show that $\rho_+(z) - \lambda_1^+$ is no bigger than the right-hand side of (7.14).

Let $M = -\boldsymbol{Q}(\lambda_1^+) \succeq 0$. For any vector w, we have

$$\begin{aligned}
\|w\|_M^2 &= -w^{\mathrm{H}}\boldsymbol{Q}(\lambda_1^+)w \\
&= [\rho_+(w) - \lambda_1^+][\lambda_1^+ - \rho_-(w)]\|w\|_A^2, && (7.15)\\
\|[\hat{g}(-K\boldsymbol{Q}(\lambda_1^+))w\|_M &\le \max_{\gamma\le t\le \Gamma} |\hat{g}(\sigma)|\,\|w\|_M = \varepsilon\|w\|_M. && (7.16)
\end{aligned}$$

Write

$$z = \hat{g}(-K\boldsymbol{Q}(\lambda_1^+))x - \sum_{i=1}^{m-1}(-1)^i c_i\big\{[K\boldsymbol{Q}(\lambda_1^+)]^i - [K\boldsymbol{Q}(\rho_+(x))]^i\big\}x.$$

Note that

$$\begin{aligned}
[K\boldsymbol{Q}(\lambda_1^+)]^i - [K\boldsymbol{Q}(\rho_+(x))]^i &= \sum_{j=0}^{i-1}\Big\{[K\boldsymbol{Q}(\lambda_1^+)]^{i-j}[K\boldsymbol{Q}(\rho_+(x))]^j \\
&\qquad - [K\boldsymbol{Q}(\lambda_1^+)]^{i-j-1}[K\boldsymbol{Q}(\rho_+(x))]^{j+1}\Big\} \\
&= \sum_{j=0}^{i-1}[K\boldsymbol{Q}(\lambda_1^+)]^{i-j-1}\Big[K\boldsymbol{Q}(\lambda_1^+) - K\boldsymbol{Q}(\rho_+(x))\Big][K\boldsymbol{Q}(\rho_+(x))]^j.
\end{aligned}$$

Therefore

$$\begin{aligned}
\|[K\boldsymbol{Q}(\lambda_1^+)]^i - [K\boldsymbol{Q}(\rho_+(x))]^i\|_2 &\le \xi_i\|K\boldsymbol{Q}(\lambda_1^+) - K\boldsymbol{Q}(\rho_+(x))\|_2 \\
&\le \xi_i(\rho_+(x) - \lambda_1^+)\|K[A\rho_+(x) + \lambda_1^+] + B\|_2,
\end{aligned}$$

where $\xi_i = \sum_{j=0}^{i-1}\|K\boldsymbol{Q}(\lambda_1^+)\|_2^{i-j-1}\|K\boldsymbol{Q}(\rho_+(x))\|^j$. Without loss of generality, we may assume $\|x\|_A = 1$. We have

$$\begin{aligned}
\|z\|_M &= \sqrt{[\rho_+(z) - \lambda_1^+][\lambda_1^+ - \rho_-(z)]}\,\|z\|_A, \qquad \text{by (7.15)} \\
\|z\|_M &\le \varepsilon\|x\|_M + \eta[\rho_+(x) - \lambda_1^+] \\
&= \left(\varepsilon\sqrt{\lambda_1^+ - \rho_-(x)} + \eta\sqrt{\rho_+(x) - \lambda_1^+}\right)\sqrt{\rho_+(x) - \lambda_1^+}, && (7.17)\\
\|z\|_A &\ge \|x\|_A - \sum_{i=1}^{m-1} c_i\|[K\boldsymbol{Q}(\lambda_1^+)]^i - [K\boldsymbol{Q}(\rho_+(x))]^i\|_2\|x\|_A \\
&\ge 1 - \eta(\rho_+(x) - \lambda_1^+), && (7.18)
\end{aligned}$$

where $\eta = \sum_{i=1}^{m-1}|c_i|\xi_i\|K[A\rho_+(x) + \lambda_1^+] + B\|_2$. Finally use

$$\rho_+(z) - \lambda_1^+ = \frac{\|z\|_M^2}{[\lambda_1^+ - \rho_-(z)]\|z\|_A^2}$$

and (7.17) and (7.18) to complete the proof. $\square$

8 Block preconditioned steepest descent/ascent method

The convergence of any of the previous steepest descent/ascent methods can be very slow if $\lambda_1^{\pm} \approx \lambda_2^{\pm}$ or $\lambda_{n-1}^{\pm} \approx \lambda_n^{\pm}$. This is reflected by $\omega_1 \approx \omega_2$ in Theorem 6.2 and 7.1. Often in practice, there are needs to compute the first few extreme quadratic eigenpairs, not just the first one. For that purpose, block variations of the methods become particularly attractive for at least the following reasons:

1. they can simultaneously compute the first k extreme quadratic eigenpairs $(\lambda_j^{\pm}, u_j^{\pm})$;
2. they run more efficiently on modern computer architecture because more computations can be organized into the matrix-matrix multiplication type;
3. they have better rates of convergence to the desired eigenpairs and save overall cost by using a block size that is slightly bigger than the number of asked eigenpairs.

In summary, the benefits of using a block variation are similar to those of using the simultaneous subspace iteration *vs.* the power method [55].

In what follows, we will explain a block steepest descent/ascent method for computing the first few (λ_j^+, u_j^+). The same reasoning applies to other extreme quadratic eigenpairs.

Any block variation starts with a given $X_0 \in \mathbb{C}^{n\times n_b}$ with $\operatorname{rank}(X_0) = n_b$, instead of just one vector $\boldsymbol{x}_0 \in \mathbb{C}^n$ previously for the single-vector steepest descent type methods. Here either the jth column of X_0 is already an approximation to u_j^+ or the subspace $\mathcal{R}(X_0)$ is a good approximation to the subspace spanned by u_j^+ for $1 \le j \le n_b$ or the canonical angles from $\mathcal{R}([u_1^+, \dots, u_k^+])$ to $\mathcal{R}(X_0)$ are nontrivial, where $k \le n_b$ is the number of desired eigenpairs. In the latter two cases, a preprocessing is needed to turn the case into the first case:

1. solve the HQEP $X_0^{\mathrm{H}}\boldsymbol{Q}(\lambda)X_0$ to get its pos-type quadratic eigenpairs $(\boldsymbol{\rho}_{0;j}^+, y_j^+)$;
2. reset $X_0 := X_0[y_1^+, \dots, y_{n_b}^+]$.

So we will assume henceforth the jth column of the given X_0 is an approximation to u_j^+. Now consider generalizing the steepest descent method to a block version. Its typical ith iterative step may well look like the following. Suppose we have already computed

$$X_i = [x_{i;1}, x_{i;2}, \dots, x_{i;n_b}] \in \mathbb{C}^{n\times n_b}$$

whose jth column $x_{i;j}$ approximates u_j^+ and

$$\varOmega_i = \operatorname{diag}(\boldsymbol{\rho}_{i;1}^+, \boldsymbol{\rho}_{i;2}^+, \dots, \boldsymbol{\rho}_{i;n_b}^+)$$

whose jth diagonal entry $\boldsymbol{\rho}_{i,j}^+ = \rho_+(x_{i;j})$ approximates λ_j^+. Define the residual matrix

$$R_i = [r_+(x_{i;1}), r_+(x_{i;2}), \dots, r_+(x_{i;n_b})] = AX_i\varOmega_i^2 + BX_i\varOmega_i + CX_i.$$

The next set of approximations are computed as follows:

1. compute a basis matrix Z of $\mathcal{R}([X_i, R_i])$ by, e.g., MGS;

2. solve the QEP $Z^{\mathrm{H}}\boldsymbol{Q}(\lambda)Z$ to get its pos-type quadratic eigenpairs $(\boldsymbol{\rho}^{+}_{i+1;j}, y^{+}_{j})$, and let $\Omega_{i+1} = \mathrm{diag}(\boldsymbol{\rho}^{+}_{i+1;1}, \boldsymbol{\rho}^{+}_{i+1;2}, \ldots, \boldsymbol{\rho}^{+}_{i+1;n_b})$;

3. set $X_{i+1} = Z[y^{+}_1, \ldots, y^{+}_{n_b}]$.

In the same way as we explained before, this block steepest descent method can be improved in two directions – extending the search space is one and incorporating preconditioners is the other.

Note that $r_{+}(x_{i;j}) = \boldsymbol{Q}(\boldsymbol{\rho}^{+}_{i;j})x_{i;j}$ and thus

$$\begin{aligned}\mathcal{R}([X_i, R_i]) &= \sum_{j=1}^{n_b} \mathcal{R}([x_{i;j}, \boldsymbol{Q}(\boldsymbol{\rho}^{+}_{i;j})x_{i;j}]) \\ &= \sum_{j=1}^{n_b} \mathcal{K}_2(\boldsymbol{Q}(\boldsymbol{\rho}^{+}_{i;j}), x_{i;j}).\end{aligned}$$

So it is natural to extend the search space, $\mathcal{R}([X_\ell, R_\ell])$ through extending each Krylov subspace $\mathcal{K}_2(\boldsymbol{Q}(\boldsymbol{\rho}^{+}_{i;j}), x_{\ell;j})$ to a high order one, and of course different Krylov subspaces can be extended to different orders. For simplicity, we will extend each to the mth order. The new extended search subspace now is

$$\sum_{j=1}^{n_b} \mathcal{K}_m(\boldsymbol{Q}(\boldsymbol{\rho}^{+}_{i;j}), x_{i;j}). \tag{8.1}$$

Define the linear operator

$$\mathscr{R}_i \,:\, X \in \mathbb{C}^{n\times n_b} \;\to\; \mathscr{R}_i(X) = AX\Omega_i^2 + BX\Omega_i + CX \in \mathbb{C}^{n\times n_b}.$$

Then the subspace in (8.1) can be compactly written as

$$\mathcal{K}_m(\mathscr{R}_i, X_i) = \mathrm{span}\{X_i, \mathscr{R}_i(X_i), \ldots, \mathscr{R}_i^{m-1}(X_i)\}, \tag{8.2}$$

where $\mathscr{R}_i^{j}(\,\cdot\,)$ is understood as successively applying the operator $\mathscr{R}_i$ j times, e.g., $\mathscr{R}_i^{2}(X) = \mathscr{R}_i(\mathscr{R}_i(X))$.

As to incorporate suitable preconditioners, in light of our extensive discussions in subsection 7.1, the search subspace should be modified to

$$\sum_{j=1}^{n_b} \mathcal{K}_m(K_{i;j}\boldsymbol{Q}(\boldsymbol{\rho}^{+}_{i;j}), x_{i;j}), \tag{8.3}$$

where $K_{i;j}$ are the preconditioners, one for each approximate eigenpair $(\boldsymbol{\rho}^{+}_{i;j}, x_{i;j})$ for $1 \le j \le n_b$ in the ith iterative step. As before, $K_{i;j}$ can be constructed in one of the following two ways:

- $K_{i;j}$ is an approximate inverse of $\boldsymbol{Q}(\tilde{\boldsymbol{\rho}}^{+}_{i;j})$ for some $\tilde{\boldsymbol{\rho}}^{+}_{i;j}$ different from $\boldsymbol{\rho}^{+}_{i;j}$, ideally closer to λ^{+}_{j} than to any other quadratic eigenvalue of $\boldsymbol{Q}(\lambda)$. But this requirement on $\tilde{\boldsymbol{\rho}}^{+}_{i;j}$ is impractical because the quadratic eigenvalue λ^{+}_{j} of $\boldsymbol{Q}(\lambda)$ is unknown. A compromise would be to make $\tilde{\boldsymbol{\rho}}^{+}_{i;j}$ closer but not equal to $\boldsymbol{\rho}^{+}_{i;j}$ than to any other $\boldsymbol{\rho}^{+}_{i;j}$.

Algorithm 8.1 Block preconditioned extended steepest descent/ascent method

Given an initial approximation $X_0 \in \mathbb{C}^{n \times n_b}$ with $\operatorname{rank}(X_0) = n_b$, and an integer $m \geq 2$, the algorithm computes approximate quadratic eigenpairs to $(\lambda_j^{\mathrm{typ}}, u_j^{\mathrm{typ}})$ for $j \in \mathbb{J}$, where $\mathbb{J} = \{1 \leq j \leq n_b\}$ for computing the few smallest quadratic eigenpairs of the given type or $\{n - n_b + 1 \leq j \leq n\}$ for computing the few largest quadratic eigenpairs of the given type.

solve the HQEP $X_0^{\mathrm{H}} \boldsymbol{Q}(\lambda) X_0$ to get its quadratic eigenpairs $(\boldsymbol{\rho}_{0;j}^{\mathrm{typ}}, y_j^{\mathrm{typ}})$;
$X_0 = X_0[y_1^{\mathrm{typ}}, \ldots, y_{n_b}^{\mathrm{typ}}]$, $\hat{\mathbb{J}} = \{1 \leq j \leq n_b\}$;
for $i = 0, 1, \ldots$ **do**
  construct preconditioners $K_{i;j}$ for $j \in \hat{\mathbb{J}}$;
  compute a basis matrix Z of the subspace

$$\sum_{j \in \hat{\mathbb{J}}} \mathcal{K}_m(K_{i;j} \boldsymbol{Q}(\boldsymbol{\rho}_{i;j}^{\mathrm{typ}}), x_{i;j}), \tag{8.4}$$

  and let n_Z be its dimension and $\hat{\mathbb{J}} = \{1 \leq j \leq n_b\}$ for computing the few smallest quadratic eigenpairs of the given type or $\{n_Z - n_b + 1 \leq j \leq n_Z\}$ for computing the few largest quadratic eigenpairs of the given type;
  compute the n_b quadratic eigenpairs of $Z^{\mathrm{H}} \boldsymbol{Q}(\lambda) Z$: $(\boldsymbol{\rho}_{i+1;j}^{\mathrm{typ}}, y_j^{\mathrm{typ}})$ for $j \in \hat{\mathbb{J}}$ and let $\Omega_{i+1} = \operatorname{diag}(\ldots, \boldsymbol{\rho}_{i+1;j}^{\mathrm{typ}}, \ldots)$ whose diagonal entries are those for $j \in \hat{\mathbb{J}}$;
  $X_{i+1} = ZW$, where $W = [\ldots, y_j^{\mathrm{typ}}, \ldots]$ whose columns are those for $j \in \hat{\mathbb{J}}$;
end for
return approximate quadratic eigenpairs to $(\lambda_j^{\mathrm{typ}}, u_j^{\mathrm{typ}})$ for $j \in \mathbb{J}$.

- Perform an incomplete LDL^{H} factorization (see [51, Chapter 10]) $\boldsymbol{Q}(\tilde{\boldsymbol{\rho}}_{i;j}^{+}) \approx L_{i;j} D_{i;j} L_{i;j}^{\mathrm{H}}$, where "$\approx$" includes not only the usual "aproximately equal", but also the case when $\boldsymbol{Q}(\tilde{\boldsymbol{\rho}}_{i;j}^{+}) - L_{i;j} D_{i;j} L_{i;j}^{\mathrm{H}}$ is approximately a low rank matrix, and $D_{i;j} = \operatorname{diag}(\pm 1)$. Finally set $K_{i:j} = L_{i;j} L_{i;j}^{\mathrm{H}}$.

Algorithm 8.1 is the general framework of a Block Preconditioned Extended Steepest Descent method (BPeSD) which embeds four methods into one:

1. Block Steepest Descent method: $m = 2$ and all preconditioners $K_{i;j} = I$;
2. Block Preconditioned Steepest Descent method: $m = 2$ and nontrivial $K_{i;j}$;
3. Block Extended Steepest Descent method: $m > 2$ and all preconditioners $K_{i;j} = I$;
4. Block Preconditioned Extended Steepest Descent method: $m > 2$ and nontrivial $K_{i;j}$.

There are three important implementation issues to worry about in turning this general framework into a piece of working code.

1. In (8.3), a different preconditioner is used for each and every approximate eigenpair $(\boldsymbol{\rho}_{i;j}^{+}, x_{i;j})$ for $1 \leq j \leq n_b$. While, conceivably, doing so will speed up convergence for each approximate eigenpair because each preconditioner can be constructed to make that approximate eigenpair converge faster, but the cost in constructing these preconditioners may likely be too heavy to bear. A more practical approach would be to use one preconditioner K_i for all $K_{i;j}$ aiming at speeding up the convergence of $(\rho_{i;1}^{+}, x_{i;1})$ (or the first few approximate quadratic eigenpairs for tightly clustered quadratic eigenvalues). Once it (or the first few in the case of a tightly cluster) is determined to be sufficiently accurate, the converged eigenpairs are locked up and deflated and a new preconditioner is computed to aim at the next non-converged eigenpairs, and the process continues.

2. Consider implementing Line 5, i.e., generating a basis matrix for the subspace (8.4). In the most general case, Z can be gotten by packing the basis matrices of all

$$\mathcal{K}_m(K_{i;j}\boldsymbol{Q}(\boldsymbol{\rho}_{i;j}^{+}), x_{\ell;j}) \quad \text{for } 1 \le j \le n_b$$

together. There could be two problems with this: 1) such Z could be ill-conditioned, i.e., the columns of Z may not be sufficiently numerically linearly independent, and 2) the arithmetic operations in building a basis for each $\mathcal{K}_m(K_{i;j}\boldsymbol{Q}(\boldsymbol{\rho}_{i;j}^{+}), x_{i;j})$ are mostly matrix-vector multiplications, straying from one of the purposes: performing most arithmetic operations through matrix-matrix multiplications in order to achieve high performance on modern computers. To address these two problems, we may do a tradeoff by using $K_{i;j} \equiv K_i$ for all j. This may likely degrade the effectiveness of the preconditioner per step in terms of rates of convergence for all approximate eigenpairs $(\boldsymbol{\rho}_{i;j}^{+}, x_{i;j})$ but may achieve overall gain in using less time because then the code will run much faster in matrix-matrix operations, not to mention the saving in constructing just one preconditioner K_i instead of n_b different preconditioners $K_{i;j}$. To simplify our discussion below, we will drop the subscript i for readability. Since $K_{i;j} \equiv K$ for all j, (8.4) is the same as

$$\mathcal{K}_m(K\mathscr{R}, X) = \operatorname{span}\{X, K\mathscr{R}(X), \dots, [K\mathscr{R}]^{m-1}(X)\}, \tag{8.5}$$

where $[K\mathscr{R}]^j(\,\cdot\,)$ is understood as successively applying the operator $K\mathscr{R}$ j times, e.g., $[K\mathscr{R}]^2(X) = K\mathscr{R}_\ell(K\mathscr{R}(X))$. A basis matrix

$$Z = [Z_1, Z_2, \dots, Z_m]$$

can be computed by the following block Arnoldi-like process.

1: $Z_1T = X$ (MGS);
2: **for** $i = 2$ to m **do**
3: $\quad Y = K(AZ_{i-1}\varOmega^2 + BZ_{i-1}\varOmega + CZ_{i-1})$;
4: $\quad$ **for** $j = 1$ to $i-1$ **do**
5: $\quad\quad T = Z_j^{\mathrm{H}}Y$; $Y = Y - Z_jT$;
6: $\quad$ **end for**
7: $\quad Z_iT = Y$ (MGS);
8: **end for**

There is a possibility that at Line 7 Y is numerically not of full column rank. If it happens, it poses no difficulty at all. In running MGS on Y's columns, anytime if a column is deemed linearly dependent on previous columns, that column should be deleted, along with the corresponding $\boldsymbol{\rho}_j^{+}$ from $\varOmega$ to shrink its size by 1 as well. At the completion of MGS, Z_i will have fewer columns than Y and the size of $\varOmega$ is shrunk accordingly. Finally, at the end, the columns of Z are orthonormal, i.e., $Z^{\mathrm{H}}Z = I$ (of apt size) which may fail to an unacceptably level due to roundoff; so some form of re-orthogonalization should be incorporated.

Algorithm 9.1 Preconditioned conjugate gradient method

Given an initial approximation $\boldsymbol{x}_0$ to $u_\ell^{\rm typ}$, a (positive definite) preconditioner K, and a relative tolerance `rtol`, the algorithm computes an approximate pair to $(\lambda_\ell^{\rm typ}, u_\ell^{\rm typ})$ with the prescribed `rtol`.

$\boldsymbol{x}_0 = \boldsymbol{x}_0/\|\boldsymbol{x}_0\|_2$, , $\boldsymbol{\rho}_0 = \rho_{\rm typ}(\boldsymbol{x}_0)$, $\boldsymbol{r}_0 = r_{\rm typ}(\boldsymbol{x}_0)$, $\boldsymbol{p}_0 = -K\boldsymbol{r}_0$;
for $i = 0, 1, \ldots$ **do**
  if $\|\boldsymbol{r}_i\|_2/(|\boldsymbol{\rho}_i|^2\|A\boldsymbol{x}_i\| + |\boldsymbol{\rho}_i|\,\|B\boldsymbol{x}_i\| + \|C\boldsymbol{x}_i\|) \le$ `rtol` **then**
    BREAK;
  else
    solve the HQEP for $Y_i^{\rm H}\boldsymbol{Q}(\lambda)Y_i$, where $Y_i = [\boldsymbol{x}_i, \boldsymbol{p}_i]$ to get its quadratic eigenvalues $\mu_j^{\pm}$ as in (6.8) and quadratic eigenvectors $y_j^{\pm}$;
    select the next approximate quadratic eigenpair $(\mu, Y_i v)$ according to the table (6.9);
    compute $\alpha_i = t_{\rm opt}$ as in (9.2) and then y as in (6.7) with $x = \boldsymbol{x}_i$ and $p = \boldsymbol{p}_i$;
    $\boldsymbol{x}_{i+1} = y/\|y\|_2$;
    set $\boldsymbol{\rho}_{i+1} = \rho_{\rm typ}(\boldsymbol{x}_{i+1})$, $\boldsymbol{r}_{i+1} = r_{\rm typ}(\boldsymbol{x}_{i+1})$, $\boldsymbol{p}_{i+1} = -K\boldsymbol{r}_{i+1} + \beta_i\boldsymbol{p}_i$, where β_i is commonly given by either one of

$$\text{either}\quad \beta_i = \frac{\boldsymbol{r}_{i+1}^{\rm H}K\boldsymbol{r}_{i+1}}{\boldsymbol{r}_i^{\rm H}K\boldsymbol{r}_i} \quad \text{or} \quad \beta_i = \frac{\boldsymbol{r}_{i+1}^{\rm H}K(\boldsymbol{r}_{i+1} - \boldsymbol{r}_i)}{\boldsymbol{r}_i^{\rm H}K\boldsymbol{r}_i}; \tag{9.1}$$

  end if
end for
return $(\boldsymbol{\rho}_i, \boldsymbol{x}_i)$ as an approximate eigenpair to $(\lambda_\ell^{\rm typ}, u_\ell^{\rm typ})$.

9 Conjugate gradient method

Again because of the equations in (3.8), the nonlinear CG type method [45, 59] and its variations are natural candidates for computing the first or last quadratic eigenpair $(\lambda_j^{\pm}, u_j^{\pm})$, and their block variations can also be devised to simultaneously compute the first or last few quadratic eigenpairs $(\lambda_j^{\pm}, u_j^{\pm})$. Since much of the machinery including gradients and preconditioning has already been built up, what remain are more or less simple adaptations of CG type methods [35] for the generalized Hermitian eigenvalue problem to the current case.

9.1 Preconditioned conjugate gradient method

The single-vector CG type methods heavily rely on the line-search problem (6.5) – (6.7) which was solved by projecting the original $n\times n$ HQEP for $\boldsymbol{Q}(\lambda)$ to a 2×2 HQEP $Y^{\rm H}\boldsymbol{Q}(\lambda)Y$ without actually computing the optimal parameter $t_{\rm opt}$ and thus the next approximation y as in (6.7) for the steepest descent/ascent method and its variations. The outcome of it is that the computed next approximation is a (complex) scalar multiply of y in (6.7). This is good enough for the steepest descent/ascent method but not for the CG method for which y in (6.7) needs to be computed. We now show how this y can be recovered from the approximation given in the table (6.9). Let (μ, Yv) is selected according to the

Algorithm 9.2 Locally optimal block preconditioned extended conjugate gradient method

Given an initial approximation $X_0 \in \mathbb{C}^{n\times n_b}$ with $\operatorname{rank}(X_0) = n_b$, and an integer $m \geq 2$, the algorithm computes approximate eigenpairs to $(\lambda_j^{\text{typ}}, u_j^{\text{typ}})$ for $j \in \mathbb{J}$, where $\mathbb{J} = \{1 \leq j \leq n_b\}$ for computing the few smallest quadratic eigenpairs of the given type or $\{n - n_b + 1 \leq j \leq n\}$ for computing the few largest quadratic eigenpairs of the given type.

1: solve the HQEP $X_0^{\mathrm{H}}\boldsymbol{Q}(\lambda)X_0$ to get its quadratic eigenpairs $(\boldsymbol{\rho}_{0;j}^{\text{typ}}, y_j^{\text{typ}})$;
2: $X_0 = X_0[y_1^{\text{typ}}, \ldots, y_{n_b}^{\text{typ}}]$, $X_{-1} = 0$, $\hat{\mathbb{J}} = \{1 \leq j \leq n_b\}$;
3: **for** $i = 0, 1, \ldots$ **do**
4: construct preconditioners $K_{i;j}$ for $j \in \hat{\mathbb{J}}$;
5: compute a basis matrix Z of the subspace

$$\sum_{j\in\hat{\mathbb{J}}} \mathcal{K}_m(K_{i;j}\boldsymbol{Q}(\boldsymbol{\rho}_{i;j}), x_{i;j}) + \mathcal{R}(X_{i-1}), \tag{9.3}$$

and let n_Z be its dimension and $\hat{\mathbb{J}} = \{1 \leq j \leq n_b\}$ for computing the few smallest quadratic eigenpairs of the given type or $\{n_Z - n_b + 1 \leq j \leq n_Z\}$ for computing the few largest quadratic eigenpairs of the given type;
6: compute the n_b quadratic eigenpairs of $Z^{\mathrm{H}}\boldsymbol{Q}(\lambda)Z$: $(\boldsymbol{\rho}_{i+1;j}^{\text{typ}}, y_j^{\text{typ}})$ for $j \in \hat{\mathbb{J}}$ and let $\Omega_{i+1} = \operatorname{diag}(\ldots, \boldsymbol{\rho}_{i+1;j}^{\text{typ}}, \ldots)$ whose diagonal entries are those for $j \in \hat{\mathbb{J}}$;
7: $X_{i+1} = ZW$, where $W = [\ldots, y_j^{\text{typ}}, \ldots]$ whose columns are those for $j \in \hat{\mathbb{J}}$;
8: **end for**
9: **return** approximate quadratic eigenpairs to $(\lambda_j^{\text{typ}}, u_j^{\text{typ}})$ for $j \in \mathbb{J}$.

table, and write $v = [\nu_1, \nu_2]^{\mathrm{T}}$ and $\hat{y} = Yv = \nu_1 x + \nu_2 p$. Thus

$$t_{\text{opt}} = \nu_2/\nu_1 \text{ if } \nu_1 \neq 0, \text{ and } \infty \text{ otherwise.} \tag{9.2}$$

With this, set y as in (6.7).

Our discussions on selecting a good preconditioner in subsection 7.1 should be followed. Algorithm 9.1 presents the framework for the single-vector preconditioned conjugate gradient method for $\boldsymbol{Q}(\lambda)$.

9.2 Locally optimal block preconditioned extended conjugate gradient method

When it comes to eigenvalue computations by CG type methods, CG's locally optimal variations [48, 60] combined with preconditioning and blocking are more preferable than the usual single-vector CG method as in Algorithm 9.1 [3, 28, 35]. In Algorithm 9.2, we present a framework of the so-called *Locally Optimal Block Preconditioned Extended Conjugate Gradient Method* (LOBPeCG) whose different implementation choice gives rise to a list of CG-type methods which we will elaborate.

The three important implementation issues we discussed for Algorithm 8.1 (Block Preconditioned Extended Steepest Descent method) after its introduction essentially apply here, except some changes are needed in the computation of Z at Line 5 here.

First X_{i-1} can be replaced by something else. Specifically, we modify Lines 2, 6, and 8 of Algorithm 9.2 to

---

2: $X_0 = X_0 W$, and $Y_0 = 0$, $\hat{\mathbb{J}} = \{1 \le j \le n_b\}$;

5: compute a basis matrix Z of the subspace

$$\sum_{j \in \hat{\mathbb{J}}} \mathcal{K}_m(K_{i;j}\boldsymbol{Q}(\boldsymbol{\rho}_{i;j}), x_{i;j}) + \mathcal{R}(Y_i), \tag{9.4}$$

such that $\mathcal{R}(Z_{(:,1:n_b)}) = \mathcal{R}(X_i)$. Let n_Z be its dimension and $\hat{\mathbb{J}} = \{1 \le j \le n_b\}$ for computing the few smallest quadratic eigenpairs of the given type or $\{n_Z - n_b + 1 \le j \le n_Z\}$ for computing the few largest quadratic eigenpairs of the given type;

7: $X_{i+1} = ZW$, where $W = [\ldots, y_j^{\text{typ}}, \ldots]$ whose columns are those for $j \in \hat{\mathbb{J}}$, $Y_{i+1} = Z_{(:,n_b+1:(m+1)n_b)} W_{(n_b+1:(m+1)n_b,:)}$;

Next we will compute a basis matrix for the subspace (9.3) or (9.4). For better performance (by using more matrix-matrix multiplications), we will assume $K_{i;j} \equiv K_i$ for all j for simplification. Dropping the subscript i for readability, we see (9.4) is the same as

$$\mathcal{K}_m(K\mathscr{R}, X) + \mathcal{R}(Y) = \operatorname{span}\{X, K\mathscr{R}(X), \ldots, [K\mathscr{R}]^{m-1}(X)\} + \mathcal{R}(Y). \tag{9.5}$$

We will first compute a basis matrix $[Z_1, Z_2, \ldots, Z_m]$ for $\mathcal{K}_m(K\mathscr{R}, X)$ by the Block Arnoldi-like process outlined at the end of section 8. In particular, $\mathcal{R}(Z_1) = \mathcal{R}(X)$. Then orthogonalize Y against $[Z_1, Z_2, \ldots, Z_m]$ to get Z_{m+1} satisfying $Z_{m+1}^{\mathrm{H}} Z_{m+1} = I$. Finally take $Z = [Z_1, Z_2, \ldots, Z_{m+1}]$.

So far, we have not mentioned any convergence properties of these CG type methods.

10 Numerical examples

In this section, we will present a couple of examples to demonstrate the numerical behavior of Algorithm 9.2 which often performs much better than the steepest descent/ascent type methods. In presenting numerical results, we will use the normalized residuals

$$\frac{\|\boldsymbol{Q}(\mu_i)x_i\|_2}{(\|A\|_1\mu_i^2+\|B\|_1|\mu_i|+\|C\|_1)\|x_i\|_2}$$

to show the convergent progress for approximations (μ_i, x_i) to a particular quadratic eigenpair *vs.* the iteration index, where using the matrix ℓ_1 operator norms $\|A\|_1$, $\|B\|_1$, and $\|C\|_1$ is more for computational convenience than anything else as any other norm would serve the same purpose just as well.

Example 10.1. This is the problem **Wiresaw1** in the collection [5]. It is actually a gyroscopic QEP arising in the vibration analysis of a wiresaw [68], but leads to an HQEP. Here

$$A=\frac{1}{2}I_n, \quad C=\frac{(\nu^2-1)\pi^2}{2}\,\mathrm{diag}(1^2,2^2,\dots,n^2),$$
$$B=\iota\,(b_{ij}) \quad \text{with} \quad b_{ij}=\begin{cases}\nu\dfrac{4ij}{i^2-j^2}, & \text{if } i+j \text{ is odd,}\\ 0, & \text{otherwise,}\end{cases}$$

where $\iota=\sqrt{-1}$ is the imaginary unit, ν is a real nonnegative parameter corresponding to the speed of the wire. For $0<\nu<1$, $\boldsymbol{Q}(0)=C$ is negative definite, and thus $\boldsymbol{Q}(\lambda)=\lambda^2A+\lambda B+C$ is hyperbolic by Theorem 2.1. Moreover

$$\lambda_i^-<0<\lambda_j^+ \quad \text{for all } i,j.$$

Therefore it is rather natural to use $K=-C^{-1}$ as a preconditioner when it comes to compute the few smallest λ_j^+ or largest λ_i^-, or for testing purpose some approximations to C^{-1} such as those corresponding to the linear conjugate gradient methods.

We ran Algorithm 9.2 with $n_b=10$, $m=2$ and random $X_0=\texttt{randn}(n,n_b)$ on this example for $n=1,000$ and $\nu=0.8$ without or with preconditioners

$$K\approx\begin{cases}\ \ [\boldsymbol{Q}(\pm 6.0\cdot 10^3)]^{-1}, & \text{for largest } \lambda_j^+ \text{ or smallest } \lambda_j^-,\\ -[\boldsymbol{Q}(0)]^{-1}=-C^{-1}, & \text{for smallest } \lambda_j^+ \text{ or largest } \lambda_j^-,\end{cases} \tag{10.1}$$

implemented through the linear conjugate gradient method with stopping criteria of normalized residuals for the involved linear systems being no bigger than 10^{-1} or reaching the maximum number CG steps which is 10. We have already explained the use of $-C^{-1}$ or its approximations as possible precondtioners. After running Algorithm 9.2 without any preconditioner, we noticed that all $\lambda_j^\pm$ lie in $(-6.0\cdot 10^3, 6.0\cdot 10^3)$ which leads to the use of $[\boldsymbol{Q}(\pm 6.0\cdot 10^3)]^{-1}$ in (10.1).

Figure 10.1 plots the residual history for computing the largest or smallest few $\lambda_i^\pm$, where the left column is for without any preconditioner while the right column is for with the preconditioners as given in (10.1). We notice without using any preconditioner

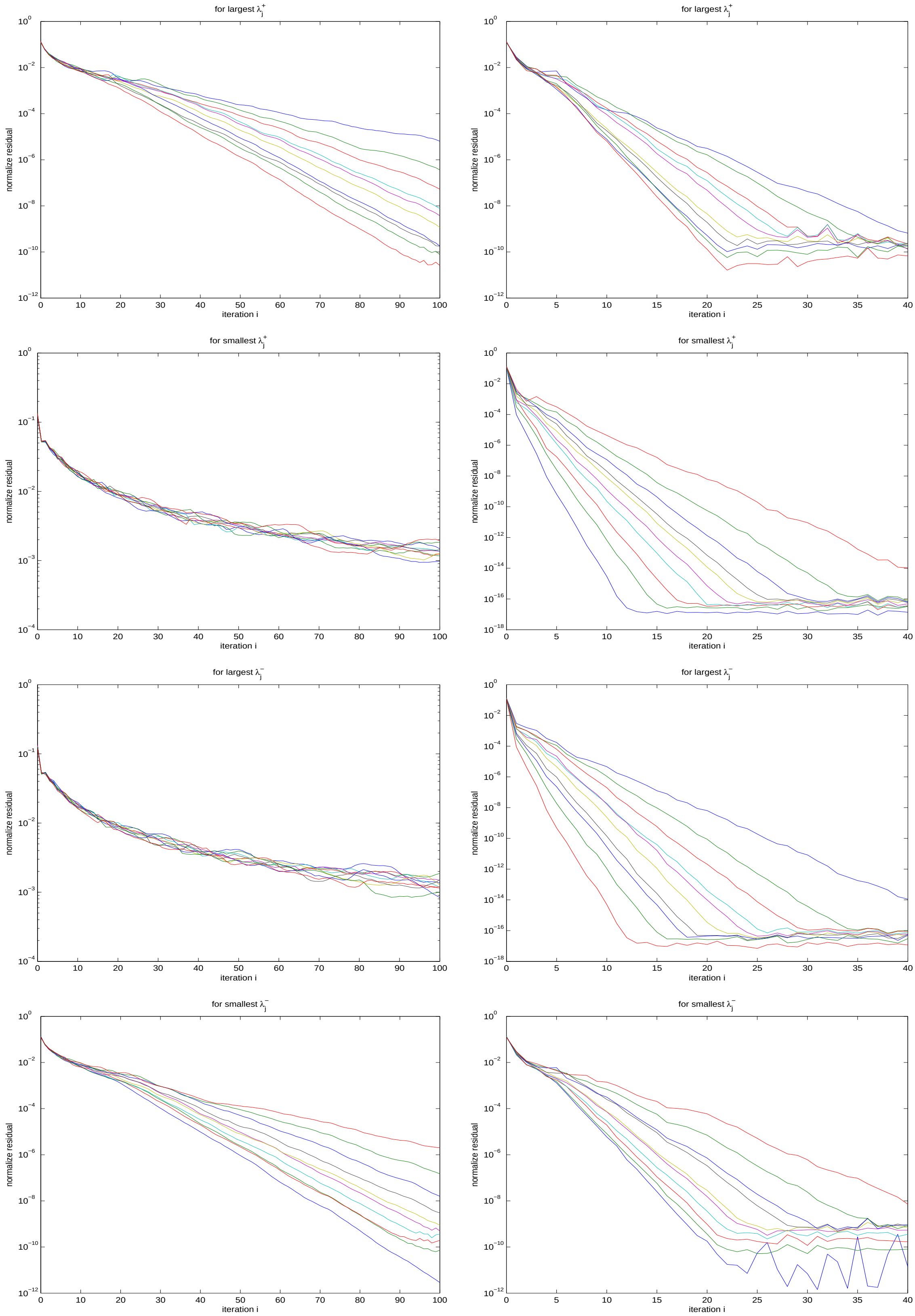


Figure 10.1: Residual history for running Algorithm 9.2 on Example 10.1

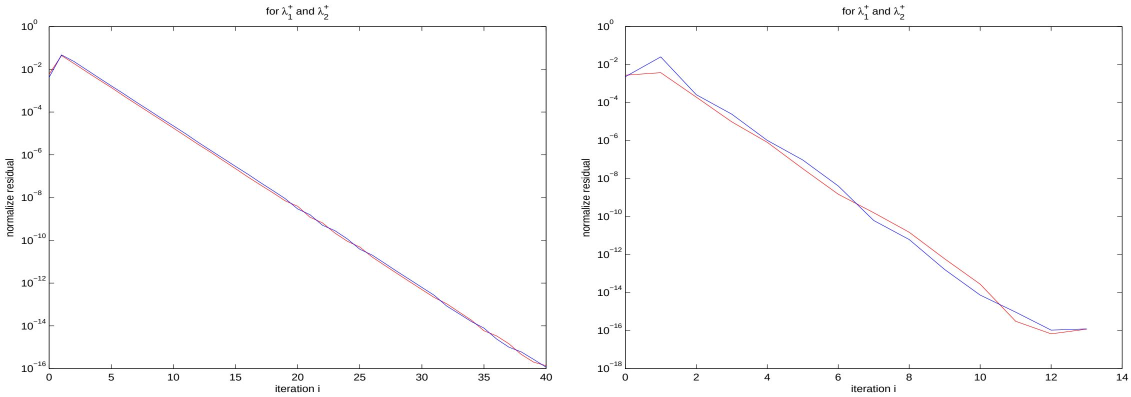


Figure 10.2: Residual history for running Algorithm 9.2 on Example 10.2 for computing λ_1^+ and λ_2^+

Algorithm 9.2 performed poorly for computing smallest λ_j^+ or largest λ_j^- but reasonably well for largest λ_j^+ or smallest λ_j^-. The effectiveness of the preconditioners as in (10.1) is rather evident by comparing the plots in the two columns.

Example 10.2. This is [20, Example 5], where $A = I_n$,

$$B = \xi \begin{bmatrix} 20 & -10 & & & \\ -10 & 30 & -10 & & \\ & \ddots & \ddots & \ddots & \\ & & -10 & 30 & -10 \\ & & & -10 & 20 \end{bmatrix}, C = \begin{bmatrix} 15 & -5 & & & \\ -5 & 15 & -5 & & \\ & \ddots & \ddots & \ddots & \\ & & -5 & 15 & -5 \\ & & & -5 & 15 \end{bmatrix},$$

and ξ is a parameter. We take $n = 1000$ and $\xi = 1.1$. This is a pathological example in the sense that most quadratic eigenvalues are close to one another – share about 3 significant decimal digits with their neighbors, except λ_1^+ and λ_2^+ which has a gap from the rest. When running Algorithm 9.2 with $m = 2$ and various different n_b, we noticed the algorithm really had hard time computing all extreme $\lambda_j^\pm$ even with some preconditioner $K = \pm[\boldsymbol{Q}(\mu)]^{-1}$ with $\mu \in (\lambda_n^-, \lambda_1^+)$ or $\mu > \lambda_n^+$ or $\mu > \lambda_1^-$ purposely selected, except for λ_1^+ and λ_2^+ which are rather easy to compute actually. Figure 10.2 plots the residual history for computing λ_1^+ and λ_2^+, where the left plot is for without any preconditioner while the right plot is for with a preconditioner $K \approx [\boldsymbol{Q}(-8.0)]^{-1}$ implemented through the linear conjugate gradient method with the same stopping criteria as in the previous example.

11 Concluding remarks

We have perform a systematic study of the hyperbolic quadratic eigenvalue problem $\boldsymbol{Q}(\lambda) = \lambda^2 A + \lambda B + C$. Such a problem usually arises from dynamical systems with heavy friction. Such a system appears, for example, in in an elevator or car braking system. It shares many characteristics with the standard Hermitian eigenvalue problem in the category of usual standard linear eigenvalue problems, and had attracted quite some attention in the past. Most of the results were collected in [16, 43, 65].

Our contributions in this paper lie in two fronts. Theoretically, we have established Amir-Moéz/Wielandt-Lidskii type min-max principles for the sums of selected quadratic eigenvalues and, as corollaries, trace min/max type principles, and also perturbation results in the spectral and Frobenius norm, as well as general unitarily invariant norms on how the quadratic eigenvalues will change if A, B, C are perturbed. Numerically, we have justified a naturally extended Rayleigh-Ritz type procedure, with the existing and newly established min-max principles, why the procedure will produce the best approximations to quadratic eigenvalues/eigenvectors, proposed steepest descent/ascent and CG type methods for computing extreme quadratic eigenpairs, and established convergence results, including the rate of convergence for the steepest descent/ascent methods, which shed light on preconditioning in what constitutes a good preconditioner and how to construct one.

Block steepest descent/ascent type methods often perform much better in practice than their single-vector counterparts, as they should be. But their exact rates of convergence are hard to establish. Experience shows that their corresponding locally optimal CG type methods perform even better, but then again we do not know the exact rates of convergence locally optimal CG type methods, either. It is recommended that locally optimal CG type methods should be preferred to their corresponding steepest descent/ascent type methods.

Despite many successes we have so far in this paper in extending, as many as we can, the important results (both theoretically and numerically) for the standard Hermitian eigenvalue problem, there are more to be done. We list a few here.

- We established perturbation bounds for quadratic eigenvalues, but didn't do so for quadratic eigenvectors/eigenspaces. The latter is worth investigating, too. We expect that $\min_x \varsigma_0(x)$ will play a role.

- Higham, Mackey, and Tisseur [23] expanded hyperbolic quadratic matrix polynomials to include the case when A is positive semidefinite. Conceivably, many results in this paper may be extensible to quadratic definite matrix polynomials in the sense of [23], but care must be taken to deal with infinite quadratic eigenvalues.

- Many results in this paper should be extensible to hyperbolic matrix polynomials of degrees higher than 2 [43]. We are working on it and results will be detailed in a separate paper.

A Digression: positive semidefinite matrix pencil

Let $\boldsymbol{A} - \lambda \boldsymbol{B}$ be a matrix pencil of order n, i.e., $\boldsymbol{A}$, $\boldsymbol{B} \in \mathbb{C}^{n\times n}$.

Definition A.1 ([38])**.** $\boldsymbol{A} - \lambda\boldsymbol{B}$ is said *Hermitian* if both $\boldsymbol{A}, \boldsymbol{B}$ are Hermitian, *positive (semi)definite* if it is Hermitian and there exists $\lambda_0 \in \mathbb{R}$ such that $\boldsymbol{A} - \lambda_0\boldsymbol{B} \succ 0$ ($\boldsymbol{A} - \lambda_0\boldsymbol{B} \succeq 0$).

The concept of positive semidefinite pencil is closely related to that of the so-called *definite pencil* in the past literature [54, 57, 58]. The latter only requires that some linear combination (with possibly complex coefficients) is positive definite and thus is necessarily a regular pencil, i.e., $\det(\boldsymbol{A} - \lambda\boldsymbol{B}) \not\equiv 0$. Definition A.1 uses more restrictive linear combinations, and also a positive semidefinite pencil of this definition may possibly be singular, i.e., $\det(\boldsymbol{A} - \lambda\boldsymbol{B}) \equiv 0$.

To include, possibly, the case in which $\boldsymbol{A} - \lambda\boldsymbol{B}$ is a singular pencil, we say $\mu \neq \infty$ is a *finite eigenvalue* of $\boldsymbol{A} - \lambda\boldsymbol{B}$ if

$$\operatorname{rank}(\boldsymbol{A} - \mu\boldsymbol{B}) < \max_{\lambda\in\mathbb{C}} \operatorname{rank}(\boldsymbol{A} - \lambda\boldsymbol{B}), \tag{A.1}$$

and $x \in \mathbb{C}^n$ is a corresponding *eigenvector* if $0 \neq x \notin \mathcal{N}(\boldsymbol{A}) \cap \mathcal{N}(\boldsymbol{B})$ satisfies

$$\boldsymbol{A}x = \mu\boldsymbol{B}x, \tag{A.2}$$

or equivalently, $0 \neq x \in \mathcal{N}(\boldsymbol{A} - \mu\boldsymbol{B})\backslash(\mathcal{N}(\boldsymbol{A}) \cap \mathcal{N}(\boldsymbol{B}))$, where $\mathcal{N}(\cdot)$ is the null space of a matrix.

In the rest of this subsection, $\boldsymbol{A} - \lambda\boldsymbol{B}$ is assumed to be a positive semidefinite pencil. Let the inertia of $\boldsymbol{B}$ be $(i_-(\boldsymbol{B}), i_0(\boldsymbol{B}), i_+(\boldsymbol{B}))$, meaning that $\boldsymbol{B}$ has $i_-(\boldsymbol{B})$ negative, $i_0(\boldsymbol{B})$ zero, and $i_+(\boldsymbol{B})$ positive eigenvalues, respectively, and set

$$n_- := i_-(\boldsymbol{B}), \quad n_+ := i_+(\boldsymbol{B}), \quad r := \operatorname{rank}(\boldsymbol{B}) = n_+ + n_-.$$

Given $0 \leq k_+ \leq n_+$ and $0 \leq k_- \leq n_-$, set

$$J_k = \begin{bmatrix} I_{k_+} & \\ & -I_{k_-} \end{bmatrix}.$$

We proved the following theorem in [38, Lemma 3.8], but later found out that it had been obtained in [13, Theorem 4.1] for the regular pencil case. This theorem play a major role in this paper.

Theorem A.1 ([13, 38])**.** *Let $\boldsymbol{A} - \lambda\boldsymbol{B}$ be a positive semidefinite Hermitian pencil of order n, and suppose that $\lambda_0 \in \mathbb{R}$ such that $\boldsymbol{A} - \lambda_0\boldsymbol{B} \succeq 0$.*

1. *There exists a nonsingular $W \in \mathbb{C}^{n\times n}$ such that*

$$W^{\mathrm{H}}\boldsymbol{A}W = \begin{array}{c} \\ {\scriptstyle n_1} \\ {\scriptstyle r-n_1} \\ {\scriptstyle n-r} \end{array}\!\!\begin{array}{c} \begin{array}{ccc} {\scriptstyle n_1} & {\scriptstyle r-n_1} & {\scriptstyle n-r} \end{array} \\ \begin{bmatrix} \Lambda_1 & & \\ & \Lambda_0 & \\ & & \Lambda_\infty \end{bmatrix} \end{array}, \quad W^{\mathrm{H}}\boldsymbol{B}W = \begin{array}{c} \\ {\scriptstyle n_1} \\ {\scriptstyle r-n_1} \\ {\scriptstyle n-r} \end{array}\!\!\begin{array}{c} \begin{array}{ccc} {\scriptstyle n_1} & {\scriptstyle r-n_1} & {\scriptstyle n-r} \end{array} \\ \begin{bmatrix} \Omega_1 & & \\ & \Omega_0 & \\ & & 0 \end{bmatrix} \end{array}, \tag{A.3}$$

where

(a) $\Lambda_1 = \mathrm{diag}(s_1\alpha_1, \dots, s_{n_1}\alpha_{n_1})$, $\Omega_1 = \mathrm{diag}(s_1, \dots, s_{n_1})$, $s_i = \pm 1$, *and* $\Lambda_1 - \lambda_0\Omega_1 \succ 0$*;*

(b) $\Lambda_0 = \mathrm{diag}(\Lambda_{0,1}, \dots, \Lambda_{0,m+m_0})$ *and* $\Omega_0 = \mathrm{diag}(\Omega_{0,1}, \dots, \Omega_{0,m+m_0})$ *with*

$$\Lambda_{0,i} = t_i\lambda_0, \qquad \Omega_{0,i} = t_i = \pm 1, \quad \text{for } 1 \le i \le m,$$
$$\Lambda_{0,i} = \begin{bmatrix} 0 & \lambda_0 \\ \lambda_0 & 1 \end{bmatrix}, \ \Omega_{0,i} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \text{for } m+1 \le i \le m+m_0.$$

There is no such pair (Λ_0, Ω_0) *if* $\boldsymbol{A} - \lambda_0\boldsymbol{B} \succ 0$*. Evidently* $m + 2m_0 = r - n_1$*.*

(c) $\Lambda_\infty = \mathrm{diag}(\alpha_{r+1}, \dots, \alpha_n) \succeq 0$ *with* $\alpha_i \in \{1, 0\}$ *for* $r+1 \le i \le n$*.*

The representations in (A.3) *are uniquely determined by* $\boldsymbol{A} - \lambda\boldsymbol{B}$*, up to a simultaneous permutation of the corresponding* 1×1 *and* 2×2 *diagonal block pairs* $(s_i\alpha_i, s_i)$ *for* $1 \le i \le n_1$*,* $(\Lambda_{0,i}, \Omega_{0,i})$ *for* $1 \le i \le m+m_0$*, and* $(\alpha_i, 0)$ *for* $r+1 \le i \le n$*.*

2. $\boldsymbol{A} - \lambda\boldsymbol{B}$ *has* $n_+ + n_-$ *finite eigenvalues all of which are real. Denote these finite eigenvalues by* $\lambda_i^\pm$ *and arrange them as*[21]
$$\lambda_1^- \le \cdots \le \lambda_{n_-}^- \le \lambda_1^+ \le \cdots \le \lambda_{n_+}^+. \tag{A.4}$$

3. $\{\gamma \in \mathbb{R} \,|\, \boldsymbol{A} - \gamma\boldsymbol{B} \succeq 0\} = [\lambda_{n_-}^-, \lambda_1^+]$*. Moreover, if* $\boldsymbol{A} - \lambda\boldsymbol{B}$ *is regular, then* $\boldsymbol{A} - \lambda\boldsymbol{B}$ *is a positive definite pencil if and only if* $\lambda_{n_-}^- < \lambda_1^+$*, in which case*
$$\{\gamma \in \mathbb{R} \,|\, \boldsymbol{A} - \gamma\boldsymbol{B} \succ 0\} = (\lambda_{n_-}^-, \lambda_1^+).$$

The next perturbation theorem for positive definite pencils seem to be new. It resembles various perturbation bounds in [8, 32, 33, 54, 57]. For the definition and properties of such unitarily invariant norms, the reader is referred to [6, 56] 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
$$\|X\|_2 \le \|X\|_{\mathrm{ui}}, \quad \|XYZ\|_{\mathrm{ui}} \le \|X\|_2 \cdot \|Y\|_{\mathrm{ui}} \cdot \|Z\|_2 \tag{A.5}$$
for any matrices X, Y, and Z of compatible sizes.

Theorem A.2. *Let* $\boldsymbol{A} - \lambda\boldsymbol{B}$ *and* $\widetilde{\boldsymbol{A}} - \lambda\widetilde{\boldsymbol{B}}$ *be two positive definite Hermitian pencils of order* n*, admitting the following eigen-decompositions*[22]*:*
$$W^{\mathrm{H}}\boldsymbol{A}W = J\Lambda, \quad W^{\mathrm{H}}\boldsymbol{B}W = J, \tag{A.6a}$$
$$\widetilde{W}^{\mathrm{H}}\widetilde{\boldsymbol{A}}\widetilde{W} = \widetilde{J}\widetilde{\Lambda}, \quad \widetilde{W}^{\mathrm{H}}\widetilde{\boldsymbol{B}}\widetilde{W} = \widetilde{J}, \tag{A.6b}$$
where Λ *is diagonal with diagonal entries consisting eigenvalues of* $\boldsymbol{A} - \lambda\boldsymbol{B}$ *in ascending order,* $J = \mathrm{diag}(-I_{i_-(\boldsymbol{B})}, I_{i_+(\boldsymbol{B})})$*, and similarly for* $\widetilde{\Lambda}$ *and* $\widetilde{J}$*. Then for any unitarily invariant norm* $\|\cdot\|_{\mathrm{ui}}$*,*
$$\|\widetilde{\Lambda} - \Lambda\|_{\mathrm{ui}} \le \|W\|_2\|\widetilde{W}\|_2\Big(\|\widetilde{\boldsymbol{A}} - \boldsymbol{A}\|_{\mathrm{ui}} + \xi\|\widetilde{\boldsymbol{B}} - \boldsymbol{B}\|_{\mathrm{ui}}\Big), \tag{A.7}$$
where $\xi = \max\{\|\Lambda\|_2, \|\widetilde{\Lambda}\|_2\}$*.*

[21] This ordering is different from the one we used in [38, 37] for the neg-type eigenvalues, in order to be consistent with what we will be using later for hyperbolic matrix polynomials. See Theorem 2.1.

[22] Such decompositions are guaranteed by Theorem A.1

Proof. We have

$$\begin{aligned}
&\boldsymbol{A}WW^{\mathrm{H}}\boldsymbol{B} - \boldsymbol{B}WW^{\mathrm{H}}\boldsymbol{A} = 0, \\
&\widetilde{\boldsymbol{A}}WW^{\mathrm{H}}\boldsymbol{B} - \widetilde{\boldsymbol{B}}WW^{\mathrm{H}}\boldsymbol{A} = \widetilde{\boldsymbol{A}}WW^{\mathrm{H}}\boldsymbol{B} - \widetilde{\boldsymbol{B}}WW^{\mathrm{H}}\boldsymbol{A} - (\boldsymbol{A}WW^{\mathrm{H}}\boldsymbol{B} - \boldsymbol{B}WW^{\mathrm{H}}\boldsymbol{A}) \\
&\qquad = (\widetilde{\boldsymbol{A}} - \boldsymbol{A})WW^{\mathrm{H}}\boldsymbol{B} - (\widetilde{\boldsymbol{B}} - \boldsymbol{B})WW^{\mathrm{H}}\boldsymbol{A}. \qquad \text{(A.8)}
\end{aligned}$$

Pre- and post-multiply (A.8) by $\widetilde{J}\widetilde{W}^{\mathrm{H}}$ and WJ, and plug the decompositions in (A.6) into (A.8) to get

$$\widetilde{\Lambda}\widetilde{W}^{-1}W - \widetilde{W}^{-1}W\Lambda = \widetilde{J}\widetilde{W}^{\mathrm{H}}(\widetilde{\boldsymbol{A}} - \boldsymbol{A})W - \widetilde{J}\widetilde{W}^{\mathrm{H}}(\widetilde{\boldsymbol{B}} - \boldsymbol{B})W\Lambda. \qquad \text{(A.9)}$$

Switching the roles of $\boldsymbol{A} - \lambda\boldsymbol{B}$ and $\widetilde{\boldsymbol{A}} - \lambda\widetilde{\boldsymbol{B}}$, we conclude from (A.9) that

$$\Lambda W^{-1}\widetilde{W} - W^{-1}\widetilde{W}\widetilde{\Lambda} = JW^{\mathrm{H}}(\boldsymbol{A} - \widetilde{\boldsymbol{A}})\widetilde{W} - JW^{\mathrm{H}}(\boldsymbol{B} - \widetilde{\boldsymbol{B}})\widetilde{W}\widetilde{\Lambda}. \qquad \text{(A.10)}$$

It follows from (A.9) and (A.10) that

$$\|\widetilde{\Lambda}\widetilde{W}^{-1}W - \widetilde{W}^{-1}W\Lambda\|_{\mathrm{ui}} \le \|W\|_2\|\widetilde{W}\|_2\Big(\|\widetilde{\boldsymbol{A}} - \boldsymbol{A}\|_{\mathrm{ui}} + \xi\|\widetilde{\boldsymbol{B}} - \boldsymbol{B}\|_{\mathrm{ui}}\Big), \qquad \text{(A.11a)}$$

$$\|\Lambda W^{-1}\widetilde{W} - W^{-1}\widetilde{W}\widetilde{\Lambda}\|_{\mathrm{ui}} \le \|W\|_2\|\widetilde{W}\|_2\Big(\|\widetilde{\boldsymbol{A}} - \boldsymbol{A}\|_{\mathrm{ui}} + \xi\|\widetilde{\boldsymbol{B}} - \boldsymbol{B}\|_{\mathrm{ui}}\Big). \qquad \text{(A.11b)}$$

Let $\widetilde{W}^{-1}W = U\Sigma V^{\mathrm{H}}$ be the SVD of $\widetilde{W}^{-1}W$ and set $\boldsymbol{C} = V^{\mathrm{H}}\Lambda V$ and $\widetilde{\boldsymbol{C}} = U^{\mathrm{H}}\widetilde{\Lambda}U$, both of which are Hermitian. It can be verified that

$$\begin{aligned}
\widetilde{\Lambda}\widetilde{W}^{-1}W - \widetilde{W}^{-1}W\Lambda &= U(\widetilde{\boldsymbol{C}}\Sigma - \Sigma\boldsymbol{C})V^{\mathrm{H}}, \\
\Lambda W^{-1}\widetilde{W} - W^{-1}\widetilde{W}\widetilde{\Lambda} &= V(\boldsymbol{C}\Sigma^{-1} - \Sigma^{-1}\widetilde{\boldsymbol{C}})U.
\end{aligned}$$

Theorem 2.1 of [7] yields

$$\|\widetilde{\boldsymbol{C}} - \boldsymbol{C}\|_{\mathrm{ui}}^2 \le \|\widetilde{\boldsymbol{C}}\Sigma - \Sigma\boldsymbol{C}\|_{\mathrm{ui}}\|\boldsymbol{C}\Sigma^{-1} - \Sigma^{-1}\widetilde{\boldsymbol{C}}\|_{\mathrm{ui}}. \qquad \text{(A.12)}$$

Mirsky's theorem [56, p.204] says

$$\|\widetilde{\Lambda} - \Lambda\|_{\mathrm{ui}} \le \|\widetilde{\boldsymbol{C}} - \boldsymbol{C}\|_{\mathrm{ui}}. \qquad \text{(A.13)}$$

The main result (A.7) is now a consequence of (A.11) – (A.13). □

In Theorem A.2, the upper bound by (A.7) contains $\|W\|_2$ and $\|\widetilde{W}\|_2$. They can be bounded, too, in terms of extreme pos- and neg-type eigenvalues.

Theorem A.3. *Let $\boldsymbol{A} - \lambda\boldsymbol{B}$ be a positive definite Hermitian pencil of order n, with eigenvalues given by and ordered as in* (A.4)*, and let its eigen-decomposition be given by* (A.6a)*. Then for any $\lambda_0 \in (\lambda_{n_-}^-, \lambda_1^+)$*

$$\|W\|_2 \le \sqrt{\max\{\lambda_{n_+}^+ - \lambda_0, \lambda_0 - \lambda_1^-\}\|(\boldsymbol{A} - \lambda_0\boldsymbol{B})^{-1}\|_2}, \qquad \text{(A.14a)}$$

$$\|W^{-1}\|_2 \le \sqrt{\frac{1}{\min\{\lambda_1^+ - \lambda_0, \lambda_0 - \lambda_{n_-}^-\}}\|\boldsymbol{A} - \lambda_0\boldsymbol{B}\|_2}. \qquad \text{(A.14b)}$$

In particular, taking $\lambda_0 = (\lambda_{n_-}^- + \lambda_1^+)/2$ *gives*

$$\|W\|_2 \leq \sqrt{(\lambda_{n_+}^+ - \lambda_1^-)\|(\boldsymbol{A} - \lambda_0 \boldsymbol{B})^{-1}\|_2}, \tag{A.15a}$$

$$\|W^{-1}\|_2 \leq \sqrt{\frac{2}{\lambda_1^+ - \lambda_{n_-}^-}\|\boldsymbol{A} - \lambda_0 \boldsymbol{B}\|_2}. \tag{A.15b}$$

Proof. For $\lambda_0 \in (\lambda_{n_-}^-, \lambda_1^+)$, $\boldsymbol{A} - \lambda_0 \boldsymbol{B} \succ 0$. We have $\boldsymbol{A} - \lambda_0 \boldsymbol{B} \succeq \lambda_{\min}(\boldsymbol{A} - \lambda_0 \boldsymbol{B}) I_n$ and thus

$$\lambda_{\min}(\boldsymbol{A} - \lambda_0 \boldsymbol{B})\, W^{\mathrm{H}} W \preceq W^{\mathrm{H}}(\boldsymbol{A} - \lambda_0 \boldsymbol{B}) W = J(\Lambda - \lambda_0 I) \preceq \max\{\lambda_{n_+}^+ - \lambda_0, \lambda_0 - \lambda_1^-\}\, I$$

which gives (A.14a). We also have

$$W^{\mathrm{H}}(\boldsymbol{A} - \lambda_0 \boldsymbol{B}) W = J(\Lambda - \lambda_0 I) \succeq \min\{\lambda_1^+ - \lambda_0, \lambda_0 - \lambda_{n_-}^-\} I$$

to give

$$W^{-\mathrm{H}} W^{-1} \preceq \frac{1}{\min\{\lambda_1^+ - \lambda_0, \lambda_0 - \lambda_{n_-}^-\}} (\boldsymbol{A} - \lambda_0 \boldsymbol{B})$$

which yields (A.14b). $\square$

References

[1] A.R. Amir-Moéz. Extreme properties of eigenvalues of a Hermitian transformation and singular values of the sum and product of linear transformations. *Duke Math. J.*, 23:463–476, 1956.

[2] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen. *LAPACK Users' Guide*. SIAM, Philadelphia, 3rd edition, 1999.

[3] Zhaojun Bai, Ren-Cang Li, and Yangfeng Su. Lecture notes on matrix eigenvalue computations. Prepared for 2009 Summer School on Numerical Linear Algebra, Chinese Academy of Science, July 2009.

[4] Lawrence Barkwell and Peter Lancaster. Overdamped and gyroscopic vibrating systems. *J. Appl. Mech.*, 59(1):176–181, 1992.

[5] T. Betcke, N. J. Higham, V. Mehrmann, C. Schröder, and F. Tisseur. NLEVP: A collection of nonlinear eigenvalue problems. *ACM Trans. Math. Software*, 39(2):7:1–7:28, 2013.

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

[7] R. Bhatia, F. Kittaneh, and Ren-Cang Li. Some inequalities for commutators and an application to spectral variation. II. *Lin. Multilin. Alg.*, 43(1-3):207–220, 1997.

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

[9] Moody T. Chu and Shufang Xu. Spectral decomposition of real symmetric quadratic λ-matrices and its applications. *Math. Comp.*, 78:293–313, 2009.

[10] G. Davis. Numerical solution of a quadratic matrix equation. *SIAM J. Sci. Statist. Comput.*, 2(2):164–175, 1981.

[11] J. Demmel. *Applied Numerical Linear Algebra.* SIAM, Philadelphia, PA, 1997.

[12] R. Duffin. A minimax theory for overdamped networks. *Indiana Univ. Math. J.*, 4:221–233, 1955.

[13] 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.

[14] D. K. Faddeev and V. N. Faddeeva. *Computational Methods of Linear Algebra.* Undergraduate Mathematics Books. W.H.Freeman & Co Ltd, San Francisco, 1963. Translated by R. C. Williams.

[15] Ky Fan. On a theorem of Weyl concerning eigenvalues of linear transformations. I. *Proc. Natl. Acad. Sci. USA*, 35(11):pp. 652–655, 1949.

[16] I. Gohberg, P. Lancaster, and L. Rodman. *Matrix Polynomials*, volume 58 of *Classics in Applied Mathematics*. SIAM, Philadelphia, 2009.

[17] G. H. Golub and C. F. Van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, Maryland, 3rd edition, 1996.

[18] Gene H. Golub and Qiang Ye. An inverse free preconditioned Krylov subspace method for symmetric generalized eigenvalue problems. *SIAM J. Sci. Comput.*, 24(1):312–334, 2002.

[19] Anne Greenbaum. *Iterative Methods for Solving Linear Systems.* SIAM, Philadelphia, 1997.

[20] C.-H. Guo and P. Lancaster. Algorithms for hyperbolic quadratic eigenvalue problems. *Math. Comp.*, 74:1777–1791, 2005.

[21] Chun-Hua Guo. Numerical solution of a quadratic eigenvalue problem. *Linear Algebra Appl.*, 385(0):391–406, 2004.

[22] Chun-Hua Guo, Nicholas J. Higham, and Françoise Tisseur. Detecting and solving hyperbolic quadratic eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 30(4):1593–1613, 2009.

[23] N. J. Higham, D. Mackey, and F. Tisseur. Definite matrix polynomials and their linearization by definite pencils. *SIAM J. Matrix Anal. Appl.*, 31(2):478–502, 2009.

[24] Nicholas J. Higham and Hyun-Min Kim. Numerical analysis of a quadratic matrix equation. *IMA J. Numer. Anal.*, 20(4):499–519, 2000.

[25] Nicholas J. Higham, Ren-Cang Li, and Françoise Tisseur. Backward error of polynomial eigenproblems solved by linearization. *SIAM J. Matrix Anal. Appl.*, 29(4):1218–1241, 2007.

[26] Nicholas J. Higham, Françoise Tisseur, and Paul M. Van Dooren. Detecting a definite Hermitian pair and a hyperbolic or elliptic quadratic eigenvalue problem, and associated nearness problems. *Linear Algebra Appl.*, 351-352:455–474, 2002.

[27] A. J. Hoffman and H. W. Wielandt. The variation of the spectrum of a normal matrix. *Duke Math. J.*, 20:37–39, 1953.

[28] Andrew V. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM J. Sci. Comput.*, 23(2):517–541, 2001.

[29] Andrew V. Knyazev and Klaus Neymeyr. A geometric theory for preconditioned inverse iteration III: A short and sharp convergence estimate for generalized eigenvalue problems. *Linear Algebra Appl.*, 358(1-3):95–114, 2003.

[30] J. Kovač-Striko and K. Veselić. Trace minimization and definiteness of symmetric pencils. *Linear Algebra Appl.*, 216:139–158, 1995.

[31] Peter Lancaster. Inverse spectral problems for semisimple damped vibrating systems. *SIAM J. Matrix Anal. Appl.*, 29(1):279–301, 2007.

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

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

[34] Ren-Cang Li. On Meinardus' examples for the conjugate gradient method. *Math. Comp.*, 77(261):335–352, 2008. Electronically published on September 17, 2007.

[35] Ren-Cang Li. Rayleigh quotient based numerical methods for eigenvalue problems. Gene Golub SIAM Summer School 2013, Fudan University, Shanghai, July 22-August 9 2013. SIAM. `http://math.fudan.edu.cn/g2s3/`.

[36] Ren-Cang Li, Wen-Wei Lin, and Chern-Shuh Wang. Structured backward error for palindromic polynomial eigenvalue problems. *Numer. Math.*, 116(1):95–122, 2010.

[37] Xin Liang and Ren-Cang Li. Extensions of Wielandt's min-max principles for positive semi-definite pencils. *Lin. Multilin. Alg.*, 2013. `DOI:10.1080/03081087.2013.803242`.

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

[39] V. B. Lidskii. The proper values of the sum and product of symmetric matrices. *Dokl. Akad. Nauk SSSR*, 75:769–772, 1950. In Russian. Translation by C. Benster available from the National Translation Center of the Library of Congress.

[40] D. E. Longsine and S. F. McCormick. Simultaneous Rayleigh-quotient minimization methods for $Ax = \lambda Bx$. *Linear Algebra Appl.*, 34:195–234, 1980.

[41] D. Steven Mackey, Niloufer Mackey, Christian Mehl, and Volker Mehrmann. Structured polynomial eigenvalue problems: Good vibrations from good linearizations. *SIAM J. Matrix Anal. Appl.*, 28(4):1029–1051, 2006.

[42] D. Steven Mackey, Niloufer Mackey, Christian Mehl, and Volker Mehrmann. Vector spaces of linearizations for matrix polynomials. *SIAM J. Matrix Anal. Appl.*, 28(4):971–1004, 2006.

[43] A.S. Markus. *Introduction to the Spectral Theory of Polynomial Operator Pencils.* Translations of mathematical monographs, vol. 71. AMS, Providence, RI, 1988.

[44] C. B. Moler and G. W. Stewart. An algorithm for generalized matrix eigenvalue problems. *SIAM J. Numer. Anal.*, 10(2):241–256, 1973.

[45] Jorge Nocedal and Stephen Wright. *Numerical Optimization.* Springer, 2nd edition, 2006.

[46] E. E. Ovtchinnikov. Sharp convergence estimates for the preconditioned steepest descent method for Hermitian eigenvalue problems. *SIAM J. Numer. Anal.*, 43(6):2668–2689, 2006.

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

[48] Boris Teodorovich Polyak. *Introduction to Optimization.* Optimization Software, New York, 1987.

[49] E. H. Rogers. A mimmax theory for overdamped systems. *Arch. Rational Mech. Anal.*, 16:89–96, 1964.

[50] E. H. Rogers. Variational properties of nonlinear spectra. *Indiana Univ. Math. J.*, 18:479–490, 1969.

[51] Yousef Saad. *Iterative Methods for Sparse Linear Systems.* SIAM, Philadelphia, 2nd edition, 2003.

[52] B. Samokish. The steepest descent method for an eigenvalue problem with semi-bounded operators. *Izv. Vyssh. Uchebn. Zaved. Mat.*, 5:105–114, 1958. in Russian.

[53] Bernhard A. Schmitt. Perturbation bounds for matrix square roots and Pythagorean sums. *Linear Algebra Appl.*, 174:215–227, 1992.

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

[55] G. W. Stewart. *Matrix Algorithms, Vol. II: Eigensystems.* SIAM, Philadelphia, 2001.

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

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

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

[59] Wenyu Sun and Ya-Xiang Yuan. *Optimization Theory and Methods – Nonlinear Programming.* Springer, New York, 2006.

[60] I. Takahashi. A note on the conjugate gradient method. *Inform. Process. Japan*, 5:45–49, 1965.

[61] F. Tisseur and K. Meerbergen. The quadratic eigenvalue problem. *SIAM Rev.*, 43(2):235–386, 2001.

[62] Françoise Tisseur. Backward error and condition of polynomial eigenvalue problems. *Linear Algebra Appl.*, 309(1-3):339–361, April 2000.

[63] K. Veselić. A Jacobi eigenreduction algorithm for definite matrix pairs. *Numer. Math.*, 64:241–269, 1993.

[64] K. Veselić. Note on interlacing for hyperbolic quadratic pencils. In Jussi Behrndt, Karl-Heinz Förster, and Carsten Trunk, editors, *Recent Advances in Operator Theory in Hilbert and Krein Spaces*, volume 198 of *Oper. Theory: Adv. Appl.*, pages 305–307. Birkhäuser, Boston, 2010.

[65] K. Veselić. *Damped Oscillations of Linear Systems*, volume 2023 of *Lecture Notes in Mathematics*. Springer, Berlin, 2011.

[66] H. Voss and B. Werner. A minimax principle for nonlinear eigenvalue problems with applications to nonoverdamped systems. *Math. Meth. Appl. Sci.*, 4:415–424, 1982.

[67] Heinrich Voss. A minmax principle for nonlinear eigenproblems depending continuously on the eigenparameter. *Numer. Linear Algebra Appl.*, 16(11-12):899–913, 2009.

[68] Songbin Wei and Imin Kao. Vibration analysis of wire and frequency response in the modern wiresaw manufacturing process. *Journal of Sound and Vibration*, 231(5):1383–1395, 2000.

[69] H. Wielandt. An extremum property of sums of eigenvalues. *Proc. Amer. Math. Soc.*, 6:106–110, 1955.

[70] H. Yang. Conjugate gradient methods for the Rayleigh quotient minimization of generalized eigenvalue problems. *Computing*, 51:79–94, 1993.