Perturbation analysis of an eigenvector-dependent nonlinear eigenvalue problem with applications

Yunfeng Cai · Zhigang Jia · Zheng-Jian Bai

Received: 20 June 2018 / Accepted: 21 June 2019 / Published online: 3 July 2019
© Springer Nature B.V. 2019

Abstract
The eigenvector-dependent nonlinear eigenvalue problem arises in many important applications, such as the discretized Kohn–Sham equation in electronic structure calculations and the trace ratio problem in linear discriminant analysis. In this paper, we perform a perturbation analysis for the eigenvector-dependent nonlinear eigenvalue problem, which gives upper bounds for the distance between the solution to the original nonlinear eigenvalue problem and the solution to the perturbed nonlinear eigenvalue problem. A condition number for the nonlinear eigenvalue problem is introduced, which reveals the factors that affect the sensitivity of the solution. Furthermore, two computable error bounds are given for the nonlinear eigenvalue problem, which can be used to measure the quality of an approximate solution. Numerical results on practical problems, such as the Kohn–Sham equation and the trace ratio optimization, indicate that the proposed upper bounds are sharper than the state-of-the-art bounds.

Keywords
Nonlinear eigenvalue problem · Perturbation analysis · Kohn–Sham equation · Trace ratio optimization

Mathematics Subject Classification 65F15 · 65F30 · 15A18 · 47J10

Communicated by Daniel Kressner.

This work was supported by NSFC Nos. 11671023, 11671337, 11771188. The research of Zheng-Jian Bai was partially supported by the Fundamental Research Funds for the Central Universities (No. 20720180008). The research of Zhigang Jia is partially supported by the Priority Academic Program Development Project (PAPD) and the Top-notch Academic Programs Project (No. PPZY2015A013) of Jiangsu Higher Education Institutions, and the Natural Science Foundation of the Jiangsu Higher Educations of China (No.18KJA110001).

Zheng-Jian Bai
zjbai@xmu.edu.cn

Extended author information available on the last page of the article
1 Introduction

In this paper, we study the perturbation theory of the following eigenvector-dependent nonlinear eigenvalue problem (NEPv)

\[ A(P)V = V \Lambda, \]  

(1.1)

where \( V \in \mathbb{C}^{n \times k} \) has orthonormal column vectors, \( A(P) \) is a continuous Hermitian matrix-valued function of \( P = V V^H \), and \( \Lambda = V^H A(P) V \in \mathbb{C}^{k \times k} \) is Hermitian, the eigenvalues of \( \Lambda \) are also eigenvalues of \( A(P) \). Here, \( \mathbb{C}^{n \times m} \) stands for the set of all \( n \times m \) matrices with complex entries, the superscript \(.H\) stands for the complex conjugate transpose of a matrix or vector. Usually, in practical applications, \( k \ll n \), and the eigenvalues of \( \Lambda \) are the \( k \) smallest or largest eigenvalues of \( A(P) \). In this paper, we restrict our discussions to the case of the \( k \) smallest eigenvalues, where \( k \leq n/2 \). Furthermore, we consider \( A(P) \) in the following form

\[ A(P) = A_0 + A_1(P), \]  

(1.2)

where \( A_0 \) and \( A_1(P) \) are all Hermitian, \( A_0 \in \mathbb{C}^{n \times n} \) is a constant matrix, and \( A_1(P) \) is a function of \( P \) with \( A_1(0) = 0 \).

Notice that if \( V \) is a solution to (1.1), then so is \( VQ \) for any \( k \times k \) unitary matrix \( Q \). Therefore, two solutions \( V, \tilde{V} \) are essentially the same if \( \mathcal{R}(V) = \mathcal{R}(\tilde{V}) \), where \( \mathcal{R}(V) \) and \( \mathcal{R}(\tilde{V}) \) are the subspaces spanned by the column vectors of \( V \) and \( \tilde{V} \), respectively. Throughout the rest of this paper, when we say that \( V \) is a solution to (1.1), we mean that the class \( \{VQ \mid Q^H Q = I_k\} \) solves (1.1).

Perhaps, the most well-known NEPv of the form (1.1) is the discretized Kohn–Sham (KS) equation arising from density function theory in electronic structure calculations (see [3,11,14] and references therein). NEPv (1.1) also arises from the trace ratio optimization in the linear discriminant analysis (LDA) for dimension reduction [12,20,21], and the Gross–Pitaevskii equation for modeling particles in the state of matter called the Bose–Einstein condensate [1,5,6]. We believe that more potential applications will emerge.

The most widely used method for solving NEPv (1.1) is the self-consistent field (SCF) iteration [11,14]. Starting with orthonormal \( V_0 \in \mathbb{C}^{n \times k} \), at the \( l \)th SCF iteration, one computes an orthonormal eigenvector matrix \( V_l \) associated with the \( k \) smallest eigenvalues of \( A(V_{l-1} V_{l-1}^H) \), and then \( V_l \) is used as the approximation in the next iteration. Convergence analysis of SCF iteration for the KS equation is studied in [9,10,19], for the trace ratio problem in [21]. Quite recently, in [2], a sufficient condition for the existence and uniqueness of the solution to NEPv (1.1) is given, and the convergence of the SCF iteration is also studied.

In practical applications, \( A(P) \) is usually obtained from the discretization of operators or constructed from empirical data, thus, contaminated by errors and noises. As a result, the NEPv (1.1) to be solved is, in fact, a perturbed NEPv. So, it is natural to ask whether we can trust the approximate solution obtained by solving the perturbed NEPv via certain numerical methods, say the SCF iteration. To be specific, let the perturbed NEPv be of the form

\[ A(P) = A_0 + A_1(P), \]  

(1.2)
where $\tilde{V}$ has orthonormal column vectors, $\tilde{P} = \tilde{V}\tilde{V}^H$, $\tilde{A} = \tilde{V}^H\tilde{A}(\tilde{P})\tilde{V} \in \mathbb{C}^{k \times k}$, and

$$\tilde{A}(\tilde{P}) = \tilde{A}_0 + \tilde{A}_1(\tilde{P})$$  \hspace{1cm} (1.4)

is a continuous Hermitian matrix-valued function of $\tilde{P}$, $\tilde{A}_0$ is a constant Hermitian matrix, $\tilde{A}_1$ is the perturbed function of $A_1$, and $\tilde{A}_1(\tilde{P})$ is still Hermitian with $\tilde{A}_1(0) = 0$.

Assume that the original NEPv (1.1) has a solution $V_*$. Then we need to answer the following two fundamental questions:

**Q1.** Under what conditions the perturbed NEPv (1.3) has a solution $\tilde{V}_*$ nearby $V_*$?

**Q2.** What’s the distance between $\mathcal{R}(V_*)$ and $\mathcal{R}(\tilde{V}_*)$?

In this paper, we will focus on Q1 and Q2. The results are established via two approaches. One is based on the well-known sin $\Theta$ theorem in the perturbation theory of Hermitian matrices [4] and Brouwer’s fixed-point theorem [7]; The other is inspired by J.-G. Sun’s technique (e.g., [8,16–18])—finding the radius of the perturbation by constructing an equation of the radius via the fixed-point theorem. Two perturbation bounds can be obtained from these two approaches, and each of them has its own merits. Based on the perturbation bounds, a condition number for the NEPv (1.1) is introduced, which quantitatively reveals the factors that affect the sensitivity of the solution. As corollaries, two computable error bounds are provided to measure the quality of the computed solution. Theoretical results are validated by numerical experiments for the KS equation and the trace ratio optimization.

The rest of this paper is organized as follows. In Sect. 2, we use two approaches to answer Q1 and Q2, followed by some discussions on the condition number and error bounds for NEPv (1.1). In Sect. 3, we apply our theoretical results to the KS equation and the trace ratio optimization problem, respectively. Finally, we give our concluding remarks in Sect. 4.

## 2 Main results

In this section, we provide two approaches to answer Q1 and Q2. A condition number and error bounds for NEPv will also be discussed. Before we proceed, we introduce the following notation, which will be used throughout the rest of this paper.

The symbol $\| \cdot \|_2$ denotes the 2-norm of a matrix or vector. Unless otherwise specified, we denote by $\lambda_j(H)$ for $1 \leq j \leq n$ the eigenvalues of a Hermitian matrix $H \in \mathbb{C}^{n \times n}$ and they are always arranged in nondecreasing order: $\lambda_1(H) \leq \lambda_2(H) \leq \cdots \leq \lambda_n(H)$.

Let $\mathcal{X}$ and $\mathcal{Y}$ be two $k$-dimensional subspaces of $\mathbb{C}^n$. Let the columns of $X$ form an orthonormal basis for $\mathcal{X}$ and the columns of $Y$ form an orthonormal basis for $\mathcal{Y}$. We use $\| \sin \Theta(\mathcal{X}, \mathcal{Y}) \|_2$ as in [15] to measure the distance between $\mathcal{X}$ and $\mathcal{Y}$, where

$$\Theta(\mathcal{X}, \mathcal{Y}) = \text{diag}(\theta_1(\mathcal{X}, \mathcal{Y}), \ldots, \theta_k(\mathcal{X}, \mathcal{Y})).$$  \hspace{1cm} (2.1)
Here, $\theta_j(\mathcal{X}, \mathcal{Y})$'s denote the $k$ canonical angles between $\mathcal{X}$ and $\mathcal{Y}$ [15, p. 43], which can be defined as

$$0 \leq \theta_j(\mathcal{X}, \mathcal{Y}) := \arccos \sigma_j \leq \frac{\pi}{2} \text{ for } 1 \leq j \leq k, \quad (2.2)$$

where $\sigma_j$'s are the singular values of $X^H Y$.

Define

$$\mathbb{V}_k := \{ V \in \mathbb{C}^{n \times k} \mid V^H V = I_k \}, \quad (2.3a)$$

$$\mathbb{P}_k := \{ P \in \mathbb{C}^{n \times n} \mid P = VV^H, V \in \mathbb{V}_k \}. \quad (2.3b)$$

Let $V_*$ and $\tilde{V}_*$ be the solutions to (1.1) and (1.3), respectively, and $P_* = V_*V_*^H$, $\tilde{P}_* = \tilde{V}_*\tilde{V}_*^H$. For any $0 \leq \xi < \frac{1}{2}$, define

$$\mathbb{V}_{\xi,*} := \{ V \in \mathbb{C}^{n \times k} \mid V^H V = I_k, \| \sin \Theta(\mathcal{R}(V), \mathcal{R}(V_*)) \|_2 \leq \xi \}, \quad (2.4)$$

$$\mathbb{P}_{\xi,*} := \{ P \in \mathbb{C}^{n \times n} \mid P = P^H, \| P - P_* \|_2 \leq \xi \}. \quad (2.5)$$

For any $P \in \mathbb{P}_{\xi,*}$, the $k$th largest eigenvalue of $P$ is strictly larger than its $(k + 1)$th largest eigenvalue since $\xi < \frac{1}{2}$. Then we can extend the definition of the function $A$ defined in (1.1) to

$$A(P) := A(V_P V_P^H), \quad \text{for any } P \in \mathbb{P}_{\xi,*},$$

where $V_P \in \mathbb{C}^{n \times k}$ is such that $V_P^H V_P = I_k$ and its columns are the eigenvectors of $P$ corresponding to the $k$ largest eigenvalues $\lambda_{n-k+1}(P), \ldots, \lambda_n(P)$. Similarly, we can extend the definition of the function $\tilde{A}$ defined in (1.3) to every element of $\mathbb{P}_{\xi,*}$.

Denote $\Delta A_0 = \tilde{A}_0 - A_0$, and also

$$\delta_0 = \| \tilde{A}_0 - A_0 \|_2, \quad \delta_1 = \sup_{P \in \mathbb{P}_{\xi,*}} \| \tilde{A}_1(P) - A_1(P) \|_2, \quad \delta = \delta_0 + \delta_1, \quad (2.6a)$$

$$d = \sup_{P \neq P_* \in \mathbb{P}_{\xi,*}} \frac{\| A_1(P) - A_1(P_*) \|_2}{\| P - P_* \|_2}. \quad (2.6b)$$

Note here that $\delta$ can be used to measure the magnitude of the perturbation, and $d$ is a “local Lipschitz constant” such that

$$\| A(P) - A(P_*) \|_2 \leq d \| P - P_* \|_2 \quad (2.7)$$

for all $P \in \mathbb{P}_{\xi,*}$. Thus, we may use $d$ to measure the sensitivity of $A(P)$ at $P_* \in \mathbb{P}_{\xi,*}$.

### 2.1 Approach I

In this section, we use the famous Weyl’s Theorem [15, p. 203], Davis-Kahan sin $\Theta$ theorem [4], and Brouwer’s fixed-point theorem [7] to answer questions $Q1$ and $Q2$. 

\[ \square \] Springer
Let $V_\ast \in \mathbb{V}_k$ be a solution to (1.1), $P_\ast = V_\ast V_\ast^H$, and

$$g = \lambda_{k+1}(A(P_\ast)) - \lambda_k(A(P_\ast)) > 0.$$  \hfill (2.8)

If

$$\delta < \frac{1}{2} (g - d) \quad \text{and} \quad \xi_\ast = \frac{2\delta}{g - d - \delta + \sqrt{(g - d - \delta)^2 - 4d\delta}} < \frac{1}{2},$$

then the perturbed NEPv (1.3) has a solution $\tilde{V}_\ast \in \mathbb{V}_{\xi_\ast, \ast}$.

**Proof** Using (2.9), we know that $\xi_\ast$ is a nonnegative constant. Then it is easy to see that $\mathbb{P}_{\xi_\ast, \ast}$ is a nonempty bounded closed convex set in $\mathbb{C}^{n \times n}$. For any $\tilde{P} \in \mathbb{P}_{\xi_\ast, \ast}$, we define

$$\phi(\tilde{P}) = \tilde{P}_\phi = \tilde{V}_\phi V_\phi^H$$

for $\tilde{V}_\phi = [\tilde{v}_{\phi 1}, \ldots, \tilde{v}_{\phi k}] \in \mathbb{V}_k$, where $\tilde{v}_{\phi j}$ is an eigenvector of $\tilde{A}(\tilde{P})$ corresponding to $\lambda_j(\tilde{A}(\tilde{P}))$ for $j = 1, \ldots, k$. If we can show that

(a) $\lambda_{k+1}(\tilde{A}(\tilde{P})) - \lambda_k(\tilde{A}(\tilde{P})) > 0$ (which implies that the mapping $\phi(\cdot)$ is well-defined in the sense that $\phi(\tilde{P})$ is unique);
(b) $\phi(\cdot)$ is a continuous mapping within $\mathbb{P}_{\xi_\ast, \ast}$;
(c) $\phi(\tilde{P}) \in \mathbb{P}_{\xi_\ast, \ast}$,

then by Brouwer’s fixed-point theorem [7], $\phi(\tilde{P})$ has a fixed point in $\mathbb{P}_{\xi_\ast, \ast}$. Let $\tilde{P}_\ast = \tilde{V}_\ast \tilde{V}_\ast^H$ be the fixed point, where $\tilde{V}_\ast \in \mathbb{V}_{\xi_\ast, \ast}$. Then $\tilde{V}_\ast$ is a solution to (1.3). Hence the conclusion follows immediately. Next, we show (a), (b) and (c) in order.

**Proof of (a)** First, direct calculations give rise to

$$\|\tilde{A}(\tilde{P}) - A(P_\ast)\|_2 \leq \|\tilde{A}_0 - A_0\|_2 + \|\tilde{A}_1(\tilde{P}) - A_1(P_\ast)\|_2$$

$$\leq \delta_0 + \|\tilde{A}_1(\tilde{P}) - A_1(\tilde{P})\|_2 + \|A_1(\tilde{P}) - A_1(P_\ast)\|_2$$

$$\leq \delta + d\|\tilde{P} - P_\ast\|_2$$

$$\leq \delta + d\xi_\ast,$$

where (2.10b) uses (2.6), (2.10c) uses $\tilde{P} \in \mathbb{P}_{\xi_\ast, \ast}$.

Second, by the famous Weyl’s Theorem [15, p. 203], we have

$$|\lambda_j(\tilde{A}(\tilde{P})) - \lambda_j(A(P_\ast))| \leq \|\tilde{A}(\tilde{P}) - A(P_\ast)\|_2, \quad \text{for} \ j = 1, 2, \ldots, n.$$  \hfill (2.11)

Then it follows that

$$\lambda_{k+1}(\tilde{A}(\tilde{P})) - \lambda_k(\tilde{A}(\tilde{P}))$$

$$= g + [\lambda_{k+1}(\tilde{A}(\tilde{P})) - \lambda_{k+1}(A(P_\ast))] + [\lambda_k(A(P_\ast)) - \lambda_k(\tilde{A}(\tilde{P}))]$$

$$\geq g - 2\|\tilde{A}(\tilde{P}) - A(P_\ast)\|_2$$

$$\geq g - 2\delta - 2d\xi_\ast$$

$$> 0,$$

where (2.12a) uses (2.11), (2.12b) uses (2.10), (2.12c) uses (2.9).
Proof of (b) We verify that $\phi(\cdot)$ is a continuous mapping within $\mathbb{P}_\xi,\ast$ by showing that for any $\tilde{P}_1, \tilde{P}_2 \in \mathbb{P}_\xi,\ast$, $\|\phi(\tilde{P}_1) - \phi(\tilde{P}_2)\|_2 \to 0$ as $\|\tilde{P}_1 - \tilde{P}_2\|_2 \to 0$.

Let $\phi(\tilde{P}_1) = \tilde{V}_{1\phi} \tilde{V}_{1\phi}^H$, $\phi(\tilde{P}_2) = \tilde{V}_{2\phi} \tilde{V}_{2\phi}^H$, and

$$\tilde{R} = \tilde{A}(\tilde{P}_1) \tilde{V}_{2\phi} - \tilde{V}_{2\phi} \text{diag}(\lambda_1(\tilde{A}(\tilde{P}_2)), \ldots, \lambda_k(\tilde{A}(\tilde{P}_2))).$$

Then

$$\tilde{R} = [\tilde{A}(\tilde{P}_1) - \tilde{A}(\tilde{P}_2)] \tilde{V}_{2\phi}.$$

Hence

$$\|\tilde{R}\|_2 = \|\tilde{A}(\tilde{P}_1) - \tilde{A}(\tilde{P}_2)\| \tilde{V}_{2\phi}\|_2 \leq \|\tilde{A}(\tilde{P}_1) - \tilde{A}(\tilde{P}_2)\|_2.$$

Using (2.9)–(2.11), we have

$$\begin{align*}
\lambda_{k+1}(\tilde{A}(\tilde{P}_2)) - \lambda_k(\tilde{A}(\tilde{P}_1)) \\
= g + [\lambda_{k+1}(\tilde{A}(\tilde{P}_2)) - \lambda_{k+1}(A(P_\ast))] - [\lambda_{k}(\tilde{A}(\tilde{P}_1)) - \lambda_{k}(A(P_\ast))] \\
\geq g - 2(\delta + d\xi_\ast) \geq g - 2\delta - d > 0. 
\end{align*}$$

By Davis–Kahan sin $\Theta$ theorem [4], we have

$$\|\sin \Theta(\mathcal{R}(\tilde{V}_{1\phi}), \mathcal{R}(\tilde{V}_{2\phi}))\|_2 \leq \frac{\|\tilde{R}\|_2}{\lambda_{k+1}(\tilde{A}(\tilde{P}_2)) - \lambda_{k}(A(P_\ast))}.$$

(2.13)

Letting $\|\tilde{P}_1 - \tilde{P}_2\|_2 \to 0$, we know that $\|\tilde{R}\|_2 \to 0$ since $\tilde{A}(\cdot)$ is continuous. Then it follows from (2.13) and (2.14) that

$$\|\phi(\tilde{P}_1) - \phi(\tilde{P}_2)\|_2 = \|\sin \Theta(\mathcal{R}(\tilde{V}_{1\phi}), \mathcal{R}(\tilde{V}_{2\phi}))\|_2 \leq \frac{\|\tilde{R}\|_2}{g - 2\delta - d} \to 0.$$

Therefore, $\|\phi(\tilde{P}_1) - \phi(\tilde{P}_2)\|_2 \to 0$.

Proof of (c) Define

$$R = \tilde{A}(\tilde{P}) V_\ast - V_\ast A_\ast,$$

where $A_\ast = V_\ast^H A(P_\ast) V_\ast$. Then

$$R = [\tilde{A}(\tilde{P}) - A(P_\ast)] V_\ast.$$

(2.15)

Using (2.10) and (2.11), we have

$$\begin{align*}
\lambda_{k+1}(A(P_\ast)) - \lambda_k(\tilde{A}(\tilde{P})) \\
= \lambda_{k+1}(A(P_\ast)) - \lambda_k(A(P_\ast)) + \lambda_k(A(P_\ast)) - \lambda_k(\tilde{A}(\tilde{P})) \\
\geq g - \delta - d\xi_\ast > 0.
\end{align*}$$

(2.16)
Then it follows that

\[ \| P_\ast - \phi(\tilde{P}) \|_2 = \| \sin \Theta(\mathcal{R}(V_\ast), \mathcal{R}(\tilde{V}_\phi)) \|_2 \]

\[ \leq \frac{\| R \|_2}{\lambda_{k+1}(A(P_\ast)) - \lambda_k(\tilde{A}(\tilde{P}))} \]

\[ \leq \frac{\| \tilde{A}(\tilde{P}) - A(P_\ast) \|_2}{g - \delta - d\xi_\ast} \]

\[ \leq \frac{\delta + d\xi_\ast}{g - \delta - d\xi_\ast} \]

\[ = \xi_\ast \]  

(2.17a) uses Davis–Kahan sin \Theta theorem [4], (2.17b) uses (2.15) and (2.16), (2.17c) uses (2.10), (2.17d) uses (2.9). Therefore, \( \phi(\tilde{P}) \in \mathbb{P}_{\xi_\ast, \ast} \). This completes the proof.

Remark 1 Zhang and Yang [22] considered the perturbation analysis of the trace ratio problem and established an upper perturbation bound, which may be useful in the practical applications in LDA. The perturbation theory in this paper is for more general eigenvector-dependent nonlinear eigenvalue problems, and the newly derived upper perturbation bounds are sharper. See Sect. 3.3 for details.

Remark 2 Cai et al. [2] presented the existence and uniqueness of the solution to (1.1) and studied the convergence of the SCF iteration. They provided fundamental results of building the perturbation theory of the unique or converged solution. The main difference between the perturbation analysis and the solvability analysis is that the former focuses on how much the original coefficient matrices can be perturbed such that the perturbed problem still has a unique solution, and how far such solution is to the solution of the original problem. The perturbation analysis of NEPv (1.1) also provides an upper bound of the distance of the computed solution to the exact one under the finite precision calculation (see Theorems 1 and 2).

2.2 Approach II

In this section, we use another approach to answer questions Q1 and Q2, which is inspired by Sun’s technique, see e.g., [8,16–18].

Theorem 2 Let \( V_\ast \in \mathbb{V}_k \) be a solution to (1.1), \( P_\ast = V_\ast V_\ast^H \), \( g \) be given by (2.8), and

\[ h = \max_{1 \leq j \leq k} [\lambda_{k+j}(A(P_\ast)) - \lambda_j(A(P_\ast))] \]

\[ \zeta = \frac{\sqrt{g}}{\sqrt{g} + \sqrt{2h}} \]

(2.18)

Assume that \( \delta \) is sufficiently small such that

\[ f(\eta) \equiv g\eta - d\eta \sqrt{1 + \eta^2} - (1 + \eta^2)\delta = 0 \]

(2.19)
has positive roots, and its smallest positive root, denoted by \( \eta_* \), is smaller than \( \xi \). Then the perturbed NEPv (1.3) has a solution \( \tilde{V}_* \in \mathcal{V}_{\tau_*,*} \) with

\[
\tau_* = \frac{\eta_*}{\sqrt{1 + \eta_*^2}}.
\] (2.20)

**Proof** Let \( [V_*, V_c] \) be a unitary matrix such that

\[
[V_*, V_c]^H A(P_*) [V_*, V_c] = \begin{bmatrix} \Lambda_* & 0 \\ 0 & \Lambda_c \end{bmatrix},
\] (2.21)

where \( \Lambda_* \) is Hermitian, and its eigenvalues are the \( k \) smallest eigenvalues of \( A(P_*) \). Then that the perturbed NEPv (1.3) has a solution \( \tilde{V}_* \) is equivalent to that there exists a unitary matrix \( [	ilde{V}_*, \tilde{V}_c] \) such that

\[
[	ilde{V}_*, \tilde{V}_c]^H A(\tilde{P}_*) [\tilde{V}_*, \tilde{V}_c] = \begin{bmatrix} \tilde{\Lambda}_* & 0 \\ 0 & \tilde{\Lambda}_c \end{bmatrix},
\] (2.22)

where \( \tilde{\Lambda}_* \) is Hermitian and its eigenvalues are the \( k \) smallest eigenvalues of \( \tilde{A}(\tilde{P}_*) \).

Without loss of generality,\(^1\) we let

\[
[V_*, V_c] = [V_*, V_c] \begin{bmatrix} Q_* & 0 \\ 0 & Q_c \end{bmatrix} \begin{bmatrix} I_k - Z^H \\ Z & I_{n-k} \end{bmatrix}
\]

\[
(I_k + Z^H Z)^{-\frac{1}{2}} \begin{bmatrix} 0 & (I_{n-k} + ZZ^H)^{-\frac{1}{2}} \\ (I_{n-k} + ZZ^H)^{-\frac{1}{2}} & 0 \end{bmatrix} \begin{bmatrix} \tilde{Q}_*^H & 0 \\ 0 & \tilde{Q}_c^H \end{bmatrix},
\] (2.23)

where \( Z \in \mathbb{C}^{(n-k)\times k} \) is a parameter matrix, \( Q_*, \tilde{Q}_* \in \mathbb{C}^{k\times k} \) and \( Q_c, \tilde{Q}_c \in \mathbb{C}^{(n-k)\times (n-k)} \) are arbitrary unitary matrices. Substituting (2.23) into (2.22), we get

\[
\tilde{Q}_*(I_k + Z^H Z)^{-\frac{1}{2}} [I_k, Z^H] D \begin{bmatrix} I_k \\ Z \end{bmatrix} (I_k + Z^H Z)^{-\frac{1}{2}} \tilde{Q}_*^H = \tilde{\Lambda}_*,
\] (2.24a)

\[
\tilde{Q}_c(I_{n-k} + ZZ^H)^{-\frac{1}{2}} [-Z, I_{n-k}] D \begin{bmatrix} -Z^H \\ I_{n-k} \end{bmatrix} (I_{n-k} + ZZ^H)^{-\frac{1}{2}} \tilde{Q}_c^H = \tilde{\Lambda}_c,
\] (2.24b)

\[
[-Z, I_{n-k}] D \begin{bmatrix} I_k \\ Z \end{bmatrix} = 0,
\] (2.24c)

\(^1\) Note that \( 2k \leq n \). By the CS decomposition [15, Chapter 1, Theorem 5.1], we know that there exist unitary matrices \( \text{diag}(U_1, U_2) \) and \( \text{diag}(U_3, U_4) \) with \( U_1, U_3 \in \mathbb{C}^{k\times k} \) and \( U_2, U_4 \in \mathbb{C}^{(n-k)\times (n-k)} \), such that \( [V_*, \tilde{V}_c] = [V_*, V_c] \text{diag}(U_1, U_2) \Sigma \Gamma \text{diag}(U_3, U_4)^H \), where \( \Sigma \) and \( \Gamma \) are diagonal matrices and \( \Sigma^2 + \Gamma^2 = I_k \). Rewrite \( [V_*, \tilde{V}_c] = [V_*, \tilde{V}_c] \text{diag}(U_3 U_1^H \tilde{Q}_* \tilde{Q}_c^H, U_4 U_2^H Q_c \tilde{Q}_c^H), (2.22) \) still holds. Then (2.23) follows immediately by setting \( Z = \tilde{Q}_c U_2 \begin{bmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{bmatrix} U_1^H \tilde{Q}_* \).
where
\[ D = [\tilde{V}_*, \tilde{V}_c]^H \tilde{A}(\tilde{P}_*) [\tilde{V}_*, \tilde{V}_c]. \quad [\tilde{V}_*, \tilde{V}_c] = [V_* Q_*, V_c Q_c]. \quad (2.25) \]

Then the perturbed NEPv (1.3) has a solution \( \tilde{V}_* \) is equivalent to

(a) there exists \( Z \) such that (2.24c) holds;
(b) \( \lambda_1(\tilde{A}_c) - \lambda_k(\tilde{A}_*) > 0. \)

Next, we first prove (a) then (b).

Proof of (a) It follows from (2.21), (2.24c) and (2.25) that
\[
0 = [ -Z, I_{n-k}][\tilde{V}_*, \tilde{V}_c]^H \tilde{A}(\tilde{P}_*) [\tilde{V}_*, \tilde{V}_c] \begin{bmatrix} I_k \\ Z \end{bmatrix} \\
= Q_c^H A_c Q_c Z - Z Q_c^H A_* Q_* + (-Z \tilde{V}_c^H + \tilde{V}_c^H) [\tilde{A}(\tilde{P}_*) - A(P_*)] (\tilde{V}_* + \tilde{V}c Z) \\
= L(Z) + \Phi(Z),
\]

where
\[
L(Z) = Q_c^H A_c Q_c Z - Z Q_c^H A_* Q_*, \quad \Phi(Z) = (-Z \tilde{V}_c^H + \tilde{V}_c^H) [\tilde{A}(\tilde{P}_*) - A(P_*)] (\tilde{V}_* + \tilde{V}c Z). \quad (2.26)
\]

Since \( g \) defined in (2.8) is positive, \( L(\cdot) \) is an invertible linear operator with
\[
\|L^{-1}\|_2^{-1} = \min_{\lambda \in \lambda(A_*), \tilde{\lambda} \in \lambda(A_c)} |\lambda - \tilde{\lambda}| = \lambda_{k+1}(A(P_*)) - \lambda_k(A(P_*)) = g > 0. \quad (2.27)
\]

Therefore, we may define a mapping \( \mu : \mathbb{C}^{(n-k) \times k} \to \mathbb{C}^{(n-k) \times k} \) as
\[
\mu(Z) \equiv -L^{-1}(\Phi(Z)). \quad (2.28)
\]

By (2.23), we have
\[
\|\tilde{P}_* - P_*\|_2 = \|\tilde{V}_* \tilde{V}_c^H - V_* V_c^H\|_2 \\
= \left\| [\tilde{V}_*, \tilde{V}_c] \begin{bmatrix} I_k \\ Z \end{bmatrix} (I_k + Z^H Z)^{-1} [I_k, Z^H] [\tilde{V}_*, \tilde{V}_c]^H - V_* V_c^H \right\|_2 \\
= \left\| (I_k + Z^H Z)^{-1} - I_k \\
\left\| Z \right\|_2 \sqrt{1 + \left\| Z \right\|_2^2}, \quad (2.29)
\]

where the last equality is obtained using the SVD of \( Z \). Then it follows from (2.26), (2.10) and (2.29) that
\[
\|L^{-1}\Phi(Z)\|_2 \leq \frac{1}{g}(1 + \|Z\|_2^2)(\delta + d\|\tilde{P}_* - P_*\|_2).
\]
\[ = \frac{1}{g}((1 + \|Z\|_2^2)\delta + d\|Z\|_2\sqrt{1 + \|Z\|_2^2}). \quad (2.30) \]

Denote
\[ \mathcal{B}_{\eta_*} = \{ Z \mid \| Z \|_2 \leq \eta_* \}. \]

Note that \( \mathcal{B}_{\eta_*} \) is a nonempty bounded closed convex set, \( \mu(\cdot) \) defined in (2.28) is a continuous mapping, and for any \( Z \in \mathcal{B}_{\eta_*} \), by (2.30) and (2.19), it holds
\[ \| \mu(Z) \|_2 \leq \frac{1}{g}((1 + \eta_*^2)\delta + d\eta_*\sqrt{1 + \eta_*^2}) = \eta_*, \]
i.e., \( \mu(Z) \) maps \( \mathcal{B}_{\eta_*} \) into itself. So by Brouwer’s fixed-point theorem [7], there is a fixed point \( Z_* \in \mathcal{B}_{\eta_*} \) such that \( \mu(Z_*) = Z_* \). In other words, (2.24c) has a solution \( Z_* \in \mathcal{B}_{\eta_*} \). This completes the proof of (a).

**Proof of (b)** By (2.24a), (2.24b) and (2.25), we know that \( \tilde{A}_* \) is a function of \( Q_*, \tilde{Q}_*, \) and \( Q_c \), and \( \tilde{A}_c \) is a function of \( Q_*, \tilde{Q}_*, \) and \( Q_c \). If
\[ \min_{Q^H Q_*=I_k, \tilde{Q}^H \tilde{Q}_*=I_k} \min_{Q^H Q_*(-I_{n-k}, \tilde{Q}^H \tilde{Q}_*(-I_{n-k}} (\| \tilde{A}_* - A_* \|_2 + \| \tilde{A}_c - A_c \|_2) < g, \quad (2.31) \]

then by Weyl Theorem [15], we have \( |\lambda_k(\tilde{A}_*) - \lambda_k(A_*)| + |\lambda_1(\tilde{A}_c) - \lambda_1(A_c)| < g \) with the arguments \( Q_*, \tilde{Q}_*, Q_c \) and \( \tilde{Q}_c \) being chosen as minimizers. Consequently,
\[ \lambda_1(\tilde{A}_c) - \lambda_k(\tilde{A}_*) = g + [\lambda_1(\tilde{A}_c) - \lambda_1(A_c)] - [\lambda_k(\tilde{A}_*) - \lambda_k(A_*)] > g - g = 0. \]

Therefore, we only need to show (2.31), under the assumption \( Z \in \mathcal{B}_{\eta_*} \).

We get by (2.10), (2.20), and (2.25) that
\[
D = [\hat{V}_*, \hat{V}_c]^H A(P_*)[\hat{V}_*, \hat{V}_c] + [\hat{V}_*, \hat{V}_c]^H [\tilde{A}(\tilde{P}_*) - A(P_*)][\hat{V}_*, \hat{V}_c]
= \begin{bmatrix}
Q^H A_* Q_* & 0 \\
0 & Q^H A_c Q_c
\end{bmatrix}
+ \Delta D. \quad (2.32)
\]

Here, \( \Delta D = [\hat{V}_*, \hat{V}_c]^H [\tilde{A}(\tilde{P}_*) - A(P_*)][\hat{V}_*, \hat{V}_c] \) satisfies
\[ \| \Delta D \|_2 = \| \tilde{A}(\tilde{P}_*) - A(P_*) \|_2 \leq \delta + d\| \tilde{P}_* - P_* \|_2 \leq \delta + d\tau_*, \quad (2.33) \]

where the last inequality uses (2.29) and \( Z \in \mathcal{B}_{\eta_*} \).

Let the singular value decomposition (SVD) of \( Z \) be \( Z = U_Z \Sigma_Z V_Z^H \), where \( U_Z \in \mathbb{C}^{(n-k)\times(n-k)}, V_Z \in \mathbb{C}^{k\times k} \) are unitary, \( \Sigma_Z = \begin{bmatrix}
\hat{\Sigma} & 0 \\
0 & \bar{\Sigma}
\end{bmatrix} \in \mathbb{C}^{(n-k)\times k}, \hat{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_k), \sigma_1 \geq \cdots \geq \sigma_k \geq 0. \) Let \( \hat{C} = \text{diag}(\cos \theta_1, \ldots, \cos \theta_k), \hat{S} = \) Springer.
$$\text{diag}(\sin \theta_1, \ldots, \sin \theta_k), \text{ where } \theta_i = \arctan \sigma_i \text{ for } i = 1, \ldots, k. \text{ Then using (2.24a), (2.32), (2.33), we have}$$

$$\min_{Q^H Q_e = I_k, \tilde{Q}^H \tilde{Q}_e = I_k} \left\| \tilde{\Lambda}_* - \Lambda_* \right\|_2$$

$$= \min_{Q^H Q_e = I_k, \tilde{Q}^H \tilde{Q}_e = I_k} \left\| \tilde{\Lambda}_* \right\|_2$$

$$\leq \| \Delta D \|_2 + \min_{Q^H Q_e = I_k, \tilde{Q}^H \tilde{Q}_e = I_k} \left\| \tilde{\Lambda}_* \right\|_2$$

$$\leq \delta + d \tau_* + h \sin^2 \theta_1$$

$$\leq \delta + d \tau_* + h \tau_*^2,$$  \hspace{1cm} (2.34)

where $\tilde{\Lambda}_* = \text{diag}(\lambda_1(A(P_e)), \ldots, \lambda_k(A(P_e)))$, $\hat{\Lambda}_c = \text{diag}(\lambda_{k+1}(A(P_e)), \ldots, \lambda_n(A(P_e)))$, and the upper bound holds for $Q_*, Q_c$, and $\tilde{Q}_e$ satisfying $V^H Q^H \Lambda_* Q_* V Z = \hat{\Lambda}_e, U^H Q^H \Lambda_c Q_c V Z = \hat{\Lambda}_c$, and $V^H \hat{Q}^H \hat{\Lambda}_* \hat{Q}^* V Z = \hat{\Lambda}_*$, accordingly.

Similarly,

$$\min_{Q^H Q_e = I_k, \tilde{Q}^H \tilde{Q}_e = I_k} \left\| \tilde{\Lambda}_c - \Lambda_c \right\|_2 \leq \delta + d \tau_* + h \tau_*^2,$$  \hspace{1cm} (2.35)

where the upper bound holds for $Q_*, Q_c$, and $\tilde{Q}_e$ satisfying $V^H Q^H \Lambda_* Q_* V Z = \hat{\Lambda}_e, U^H Q^H \Lambda_c Q_c V Z = \hat{\Lambda}_c$, and $U^H \hat{Q}^H \Lambda_c \hat{Q}^* U Z = \hat{\Lambda}_c$, accordingly.

Direct calculations give rise to

$$2[\delta + d \tau_* + h \tau_*^2] - g = 2 \left( \delta + d \frac{\eta_*}{1 + \eta_*^2} \right) + 2h \frac{\eta_*^2}{1 + \eta_*^2} - g$$

$$= 2g \frac{\eta_*}{1 + \eta_*^2} + 2h \frac{\eta_*^2}{1 + \eta_*^2} - g$$  \hspace{1cm} (2.36a)

$$< 2g \frac{\xi}{1 + \xi^2} + 2h \frac{\xi^2}{1 + \xi^2} - g$$  \hspace{1cm} (2.36b)

$$= 2h \xi^2 - g(1 - \xi)^2$$

$$= 0,$$  \hspace{1cm} (2.36c)
where (2.36a) uses the fact $\eta^*$ is a root of (2.19), (2.36b) uses $\eta^* < \zeta$, (2.36c) uses (2.18). Notice that (2.34) and (2.35) hold for $Q^*$ and $Q_c$ satisfying the common conditions. This, together with (2.36), yields (b).

Finally, we show that $\tilde{V}^* \in \mathcal{V}_{\tau^*,*}$ under the assumption $Z \in \mathbb{B}_{\eta^*}$. From (2.23) we have

$$\tilde{V}^* = [V^*, V_c] \begin{bmatrix} Q_s & 0 \\ 0 & Q_c \end{bmatrix} \begin{bmatrix} I_k \\ Z \end{bmatrix} (I_k + Z^HZ)^{-\frac{1}{2}} \tilde{Q}^H.$$  

It is easy to see that $\tilde{V}^H \tilde{V}^* = I_k$. On the other hand, using (2.29) and the assumption that $Z \in \mathbb{B}_{\eta^*}$ we have

$$\| \sin \Theta(\mathcal{A}(V^*), \mathcal{A}(\tilde{V}^*)) \|_2 = \| P^* - \tilde{P}^* \|_2 = \frac{\| Z \|_2}{\sqrt{1 + \| Z \|_2^2}} \leq \frac{\eta^*}{\sqrt{1 + \eta^*_s}} = \tau^*.$$  

Therefore, $\tilde{V}^* \in \mathcal{V}_{\tau^*,*}$. The proof is completed.

Note that $g > d$ is a necessary condition for that $f(\eta) = 0$ has positive roots. Otherwise, $f(\eta)$ is always negative, and hence, $f(\eta) = 0$ has no roots. Next, we have several remarks in order.

**Remark 3** When the perturbation is sufficiently small, i.e., $\delta \ll 1$, we have the following two claims:

1. The assumption of Theorem 2 is weaker than that of Theorem 1.
2. The perturbation bound of Theorem 2 is sharper than that of Theorem 1.

Claim (1) can be verified as follows. Let the perturbation $\delta$ be sufficiently small and less than $\frac{1}{2}(g - d)\zeta$, we have

$$f \left( \frac{2\delta}{g - d} \right) = \frac{2g\delta}{g - d} - \frac{2d\delta}{g - d} - \delta + O(\delta^2) = \delta + O(\delta^2) > 0. \quad (2.37)$$

Note that $f(0) = -\delta < 0$. Therefore, $f(\eta) = 0$ has at least one positive root within interval $(0, \frac{2\delta}{g - d}) \subset (0, \zeta)$. In other words, the assumption of Theorem 2, which requires $f(\eta) = 0$ has a positive root within $(0, \zeta)$, is satisfied if $g > d$, provided that the perturbation is sufficiently small. For the assumption of Theorem 1, no matter how small the perturbation $\delta$ is, it requires $g > 2d$. Claim (2) can be verified as follows. Using the second order Taylor’s expansion of $\sqrt{1 + x} = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + O(x^3)$, we have by calculations,

$$f \left( \frac{\frac{\xi^*}{\sqrt{1 - \xi^*_2}}}{g - d} \right) = \frac{g\delta^2}{(g - d)^2} + O(\delta^3).$$
Thus, \( f\left(\frac{\xi_*}{\sqrt{1-\xi_*^2}}\right) > 0 \) since \( \delta \ll 1 \). Also note that \( f(0) < 0 \), we know \( \eta_* < \frac{\xi_*}{\sqrt{1-\xi_*^2}} \), which leads to \( \frac{\eta_*}{\sqrt{1+\eta_*^2}} < \xi_* \).

**Remark 4** Note that \( h > g \), then \( \xi \) defined in Theorem 2 is less than \( \frac{1}{1+\sqrt{2}} \), and \( \tau_* \) is less than \( \frac{1}{\sqrt{1+(1+\sqrt{2})^2}} \approx 0.3827 \). Therefore, when \( \delta \) is not sufficiently small, Theorem 2 may not be applicable since \( \| \sin \Theta(V_*, \widetilde{V}_*) \|_2 \) can be larger than 0.3827, meanwhile Theorem 1 can be still applicable as long as \( g > 2d \). If \( h \gg 1 \) or even \( h \to \infty \), then \( \eta_* \to 0 \) and again Theorem 2 has no meaning, while Theorem 1 still holds. For example, let

\[
A(P) = \text{diag}(0, 0, 2, 16), \quad k = 2, \quad \widetilde{A}(P) = A(P) + \text{diag}\left(0, \begin{bmatrix} 0 & \frac{1}{2} & 0 \end{bmatrix}, 0\right).
\]

Then \( g = 2, d = 0, \delta = \frac{1}{2}, h = 16, \xi = \frac{1}{2} \). And by simple calculations, we have \( \text{span}(V_*) = \text{span}\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right), \text{span}(\widetilde{V}_*) = \text{span}\left(\begin{bmatrix} 1 & 0 & 2+\sqrt{5} \\ 0 & 0 & -1 \end{bmatrix}\right) \), \( \| \sin \Theta(V_*, \widetilde{V}_*) \|_2 \approx 0.2298 \). On one hand, since \( \delta < \frac{1}{2} g - d \), we can apply Theorem 1, and direct calculation gives the bound \( \xi_* = \frac{1}{3} \). On the other hand, the roots of (2.19) are \( 2 \pm \sqrt{3} \), the smaller one \( \eta_* = 2 - \sqrt{3} \approx 0.2679 \), is larger than \( \zeta = 0.2 \). Thus, we can not apply Theorem 2. However, \( \tau_* = \frac{\eta_*}{\sqrt{1+\eta_*^2}} \approx 0.2588 \) is still a valid upper bound, even sharper than \( \xi_* \).

This fact tells us that the assumptions in Theorem 2 perhaps can be relaxed, which deserves further study.

**Remark 5** Consider the following perturbation problem of a Hermitian matrix: Given a Hermitian matrix \( A_0 \), a perturbation matrix \( \Delta A_0 \), which is also Hermitian. Let the eigenvalues of \( A_0 \) be \( \lambda_1 \leq \cdots \leq \lambda_n \), the column vectors of \( V_* \) and \( \widetilde{V}_* \) be the eigenvectors of \( A_0 \) and \( A_0 + \Delta A_0 \) associated with their \( k \) smallest eigenvalues, respectively. Assume \( g = \lambda_{k+1} - \lambda_k > 0 \). What’s the upper bound for \( \| \sin \Theta(V_*, \widetilde{V}_*) \|_2 \)?

Note that since \( d = 0 \), (2.19) becomes a quadratic equation of \( \eta \). It is easy to see that it has positive roots if and only if \( g \geq 2 \delta \). And when \( g \geq 2 \delta \), it has two positive roots, and the smaller one is \( \frac{2 \delta}{g + \sqrt{g^2 - 4 \delta^2}} \). Then Theorem 2 can be rewritten as:

If \( \delta \leq \frac{1}{2} g \) and \( \frac{2 \delta}{g + \sqrt{g^2 - 4 \delta^2}} < \zeta \), then \( \| \tan \Theta(V_*, \widetilde{V}_*) \|_2 \leq \frac{2 \delta}{g + \sqrt{g^2 - 4 \delta^2}} \).

This conclusion is similar to the perturbation theorems in [15, Chapter V, Sect. 2.2].
2.3 Condition number

In this section, we provide a condition number for NEPv (1.1). Recall the theory of condition developed by Rice [13], also note that

\[
\frac{\| P_* - \tilde{P}_* \|_2}{\| P_* \|_2} = \| \sin(\mathcal{R}(V_*), \mathcal{R}(\tilde{V}_*)) \|_2.
\]

We may define a condition number as

\[
\kappa = \lim_{\varepsilon \to 0} \left\{ \frac{\| \sin(\mathcal{R}(V_*), \mathcal{R}(\tilde{V}_*)) \|_2}{\varepsilon} \right\},
\]

where \( V_* \) and \( \tilde{V}_* \) are the solutions to (1.1) and (1.3), respectively, \( \delta \) is defined in (2.6).

Now using the second-order Taylor’s expansion of \((1 + x)^{1/2}\), by (2.9), we have

\[
\xi_* = \frac{1}{g - d} \delta + O(\delta^2). \tag{2.38}
\]

Combining it with Theorem 1, we can obtain the first order absolute perturbation bound for the eigenvector subspace \( V_* \):

\[
\| \sin(\mathcal{R}(V_*), \mathcal{R}(\tilde{V}_*)) \|_2 \leq \frac{1}{g - d} \delta + O(\delta^2) \leq \frac{1}{g - d} \varepsilon + O(\varepsilon^2). \tag{2.39}
\]

Then it follows that

\[
\lim_{\varepsilon \to 0} \frac{\| \sin(\mathcal{R}(V_*), \mathcal{R}(\tilde{V}_*)) \|_2}{\varepsilon} \leq \frac{1}{g - d}.
\]

The upper bound can be attained for some special NEPv with some special perturbations \( A_0 - \tilde{A}_0 \) and \( \tilde{A}_1(P) - A_1(P) \). For example, consider the following NEPv

\[
A(P)v = \lambda v,
\]

where \( v \in \mathbb{R}^2 \) is a unit vector, \( P = vv^T \), \( \lambda = v^TA(P)v \in \mathbb{R} \) and

\[
A(P) = A_0 + A_1(P), \quad A_0 = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}, \quad A_1(P) = \frac{1}{2} \text{diag}(P).
\]

Here, the superscript “\( \cdot^T \)” means the transpose of a matrix or vector. Then the NEPv has two eigenpairs \((\frac{1}{2}, [1, 0]^T)\) and \((\frac{5}{2}, [0, 1]^T)\) and thus \( g = \frac{5}{2} - \frac{1}{2} = 2 \) and \( v_* = [1, 0]^T \) is such that \( A(P_*)v_* = v_*(v_*^TA(P_*)v_*) \). Suppose that the perturbed NEPv is given by

\[
\tilde{A}(P)\tilde{v} = \tilde{\lambda}\tilde{v}, \tag{2.40}
\]
where $\tilde{v}$ is a unit vector, $\tilde{P} = \tilde{v}\tilde{v}^T, \tilde{A} = \tilde{v}^T\tilde{A}(\tilde{P})\tilde{v} \in \mathbb{R}$, and

$$\tilde{A}(\tilde{P}) = \tilde{A}_0 + \tilde{A}_1(\tilde{P}), \quad \tilde{A}_0 = A_0 + \begin{bmatrix} 0 & -v \\ -v & 0 \end{bmatrix}^T, \quad \tilde{A}_1(\tilde{P}) = A_1(\tilde{P}), \quad \forall v \in \mathbb{R}.$$  

It is easy to check that $\delta = |v|$ and $d = \frac{1}{2}$. Let $\tilde{v}_* = [\cos(\theta), \sin(\theta)]^T$. Thus,

$$\|\sin \Theta(\tilde{A}(v_*), \tilde{A}(\tilde{v}_*))\|_2 = |\sin(\theta)|.$$  

Also, substituting $\tilde{v}_*$ into (2.40) yields

$$\nu = \tan(2\theta) - \frac{1}{4} \sin(2\theta).$$  

Let $\varepsilon = \delta = |v|$. Then we have

$$\lim_{\varepsilon \to 0} \frac{\|\sin \Theta(\tilde{A}(V_*), \tilde{A}(\tilde{V}_*))\|_2}{\varepsilon} = \lim_{\theta \to 0} \frac{|\sin(\theta)|}{|\tan(2\theta) - \frac{1}{4} \sin(2\theta)|} = \frac{2}{3} = \frac{1}{g - d}.$$  

The upper bound cannot be improved in general since it is attainable for $n = 2$. However, we may use the following quantity

$$\kappa_* \equiv \frac{1}{g - d}$$  

as the condition number for NEPv (1.1).

On the other hand, the form of (2.41) can also be derived from Theorem 2. In fact, letting $\delta \to 0$ in (2.19), by (2.37), we know that $\eta_*$ is less than $\frac{2\delta}{g - d}$, thus, $\eta_* \to 0$. Then (2.19) can be rewritten as

$$g \eta_* - d \eta_* \leq \delta + \mathcal{O}(\delta^2).$$  

Therefore, $\frac{\eta_*}{\sqrt{1 + \eta_*^2}} \leq \eta_* \leq \frac{\delta}{g - d} + \mathcal{O}(\delta^2)$. Thus, by Theorem 2, we have

$$\|\sin \Theta(\tilde{A}(V_*), \tilde{A}(\tilde{V}_*))\|_2 \leq \frac{1}{g - d} \delta + \mathcal{O}(\delta^2), \quad \text{as } \delta \to 0.$$  

from which we can use the quantity defined as in (2.41) as the condition number for NEPv (1.1).

Recall that $g$ is the gap between the $k$th and $(k + 1)$th smallest eigenvalues of $A(P_*)$, and $d$ is a local Lipschitz constant for the inequality $\|A(P) - A(P_*)\|_2 \leq d \|P - P_*\|_2$ for all $P \in \mathbb{P}_{\xi_*}$. Thus, the newly defined condition number $\kappa_*$, which may be used to measure the sensitivity of NEPv at $V_*$, depends on the eigenvalue gap as well as the sensitivity of $A(P)$ at $P = P_*$. A large $g$ and a small $d$ will ensure a good conditioned NEPv (1.1).
Remark 6 Notice that $\delta$ can be used to measure the magnitude of the backward error [see (2.43) below]. Then using the rule of thumb – “forward error $\lesssim$ backward error $\times$ condition number”, we may use $\frac{\delta}{g-d}$ as an approximate perturbation bound.

2.4 Error bounds

In this section we give two error bounds for NEPv (1.1), which can be used to measure the quality of approximate solutions to NEPv (1.1).

Let $\widehat{V} \in \mathbb{V}_k$ be an approximate solution to NEPv (1.1), and denote the residual by

$$R = A(\widehat{P})\widehat{V} - \widehat{V}[\widehat{V}^H A(\widehat{P})\widehat{V}], \quad (2.42)$$

where $\widehat{P} = \widehat{V}\widehat{V}^H \in \mathbb{P}_k$. It is easy to verify that (2.42) can be rewritten as

$$\widehat{A}(\widehat{P})\widehat{V} = \widehat{V}[\widehat{V}^H \widehat{A}(\widehat{P})\widehat{V}], \quad (2.43)$$

where $\widehat{A}(\widehat{P}) = A_0 + \Delta A_0 + A_1(\widehat{P})$ and $\Delta A_0 = -R\widehat{V}^H - \widehat{V}R^H$. Now we take (1.1) as a perturbed NEPv of (2.43), where only the constant matrix $A_0$ is perturbed, the matrix function $A_1$ remains unchanged. Noticing that $\delta_0 = \|R\widehat{V}^H + \widehat{V}R^H\| = \|R\|, \delta_1 = 0$ and $\delta = \|R\|$, we can rewrite Theorems 1 and 2 as the following two corollaries.

Corollary 1 Let $\widehat{V}$ be an approximate solution to NEPv (1.1), $\widehat{P} = \widehat{V}\widehat{V}^H$, $R$ be given by (2.42). Define $\hat{d}$ as in (2.6) by replacing $P_*$ by $\widehat{P}$, and assume

$$\hat{g} = \lambda_{k+1}(\widehat{A}(\widehat{P})) - \lambda_k(\widehat{A}(\widehat{P})) > 0. \quad (2.44)$$

If

$$\|R\| < \frac{1}{2}\hat{g} - \hat{d} \quad \text{and} \quad \hat{\xi} = \frac{2\|R\|}{\hat{g} - \hat{d} - \|R\| + \sqrt{(\hat{g} - \hat{d} - \|R\|)^2 - 4\hat{d}\|R\|}} < \frac{1}{2}, \quad (2.45)$$

then NEPv (1.1) has a solution $V_* \in \mathbb{V}_{\hat{\xi}^*,*}$.

Corollary 2 Let $\widehat{V}$ be an approximate solution to NEPv (1.1), $\widehat{P} = \widehat{V}\widehat{V}^H$, $R$ be given by (2.42). Assume (2.44), define $\hat{d}$ as in Corollary 1, and denote

$$\hat{h} = \max_{1 \leq j \leq k} [\lambda_{k+j}(\widehat{A}(\widehat{P})) - \lambda_j(\widehat{A}(\widehat{P}))], \quad \hat{\xi} = \frac{\sqrt{\hat{g}}}{\sqrt{\hat{g}} + \sqrt{2\hat{h}}}. \quad (2.46)$$

Suppose that $\|R\|$ is sufficiently small such that

$$\hat{f}(\eta) = \hat{g}\eta - \hat{d}\eta\sqrt{1 + \eta^2} - (1 + \eta^2)\|R\| = 0 \quad (2.47)$$

Springer
has positive roots, and its smallest positive root, denoted by \( \hat{\eta}_* \), is smaller than \( \hat{\zeta} \). Then the NEPv (1.1) has a solution \( V_* \in \mathbb{V}_{\hat{\xi}_*,*} \) with

\[
\hat{\tau}_* = \frac{\hat{\eta}_*}{\sqrt{1 + \hat{\eta}_*^2}}.
\]

(2.48)

It is worth mentioning here that both \( \hat{\xi}_* \) in (2.45) and \( \hat{\tau}_* \) in (2.48) are computable as long as \( \hat{g} \) and \( \hat{d} \) are available.

**Remark 7** By (2.43), we can use \( \delta = \| \Delta A_0 \|_2 = \| R \|_2 \) to measure the magnitude of the backward error. Using the condition number \( \kappa_* \) defined in (2.41) and the rule of thumb, we may use \( \frac{\| R \|_2}{\hat{g} - \hat{d}} \) as an approximate error bound, where \( \hat{g} \) is given by (2.44).

### 3 Applications

In this section, we apply our theoretical results to two practical problems: the Kohn–Sham equation and the trace ratio optimization. All numerical experiments are carried out using MATLAB R2016b, with machine epsilon \( \varepsilon \approx 2.2 \times 10^{-16} \).

The exact solution \( V_* \) to NEPv (1.1) is approximated by \( \hat{V}_* \), which is obtained by solving NEPv (1.1) via SCF iteration with stopping criterion

\[
\frac{\| A(\hat{V}_* \hat{V}_*^H)\hat{V}_* - \hat{V}_*[\hat{V}_*^H A(\hat{V}_* \hat{V}_*^H)\hat{V}_*]\|_2}{\| A(\hat{V}_* \hat{V}_*^H)\|_2} \leq 10^{-14}.
\]

And the exact solution \( \tilde{V}_* \) to NEPv (1.3) is approximated similarly. At the \( l \)th SCF iteration, an approximate solution \( V_l \) is obtained. Then we can use \( V_l \) to validate our error bounds, which will tell us how far away the approximate solution \( V_l \) from the exact solution \( V_* \).

The following notations will be used to illustrate our results. The solution perturbation \( \| \sin \Theta (\mathcal{R}(V_*), \mathcal{R}(\hat{V}_*)) \|_2 \), the perturbation bounds given by Theorems 1 and 2, and Remark 6 are denoted by \( \chi_* \), \( \xi_* \), \( \tau_* \) and \( \gamma_* \), respectively. For the approximate solution \( V_l \), the solution error \( \| \sin \Theta (\mathcal{R}(V_*), \mathcal{R}(V_l)) \|_2 \) and the error bounds given by Corollaries 1, 2 and Remark 7 are denoted by \( \hat{\chi}_* \), \( \hat{\xi}_* \), \( \hat{\tau}_* \) and \( \hat{\gamma}_* \), respectively.

### 3.1 Application to the Kohn–Sham equation

We consider the perturbation of the discretized KS equation:

\[
H(V)V = V\Lambda,
\]

(3.1)

where \( V \in \mathbb{R}^{n \times k} \) is orthonormal, the discretized Hamiltonian \( H(V) \in \mathbb{R}^{n \times n} \) is a matrix function with respect to \( V \), and \( \Lambda \in \mathbb{R}^{k \times k} \) is a diagonal matrix consisting of the \( k \) smallest eigenvalues of \( H(V) \). In particular, we consider the discretized Hamiltonian in the form of
In Table 1, we list where \( \rho \) sprandsym as follows:

\[
H(V) = \frac{1}{2} L + V_{\text{ion}} + \text{Diag}(L^\dagger \rho) - 2\gamma \text{Diag}(\rho^{\frac{1}{3}}),
\]

where \( L \) is a finite dimensional representation of the Laplacian operator, \( V_{\text{ion}} \) is the ionic pseudopotentials sampled on the suitably chosen Cartesian grid, \( L^\dagger \) denotes the pseudoinverse of \( L \), \( \rho = \text{diag}(VV^T) \) denotes the vector containing the diagonal elements of the matrix \( VV^T \), and \( \text{Diag}(x) \) denotes a diagonal matrix with \( x \) on its diagonal. The last term of (3.2) is derived from \( e_{xc}(\rho) \) defined in [10, Eq. (2.11)], where \( \rho^{\frac{1}{3}} \) means the componentwise cubic root of \( \rho \) and \( \gamma = 2\left(\frac{3}{\pi}\right)^{1/3} \).

Let

\[
A_0 = \frac{1}{2} L + V_{\text{ion}}, \quad A_1(P) = \text{Diag}(L^\dagger \rho(P)) - 2\gamma \text{Diag}(\rho(P)^{\frac{1}{3}}),
\]

where \( P = VV^T \). Then the discretized Hamiltonian \( H(V) \) can be rewritten as

\[
A(P) = A_0 + A_1(P).
\]

Thus, the KS equation (3.1) with \( H(V) \) given by (3.2) can be written in the form of (1.1) with (1.2), indeed.

Next, we set the perturbed KS equation as in the form (1.3) with \( \tilde{A}_0 := \frac{1}{2} L + V_{\text{ion}} + \Delta L + \Delta V_{\text{ion}}, \tilde{A}_1(\tilde{P}) := \text{Diag}((L + \Delta L)^\dagger \rho(\tilde{P})) - 2\gamma \text{Diag}(\rho(\tilde{P})^{\frac{1}{3}} \cdot \Delta \rho(\tilde{P})) \), where \( \tilde{P} = V\tilde{V}^T \), \( \Delta L, \Delta V_{\text{ion}} \) are real symmetric matrices, \( \Delta \rho(\tilde{P}) \) is a vector, and \( \rho(\tilde{P})^{\frac{1}{3}} \cdot \Delta \rho(\tilde{P}) \) means the entrywise product of \( \rho(\tilde{P})^{\frac{1}{3}} \) and \( \Delta \rho(\tilde{P}) \). Then, for our case, let \( \delta_0, \delta_1, \) and \( d \) be defined as in (2.6). In our numerical tests, \( L, V_{\text{ion}}, \Delta L, \) and \( \Delta V_{\text{ion}} \) are generated by using the \text{MATLAB} built-in functions \text{eye}, \text{diag}, \text{ones}, \text{zeros}, \text{or sprandsym} as follows:

\[
L = \text{eye}(n) - \text{diag}(\text{ones}(n-1, 1), 1), \quad L = (L + L')/\alpha^2,
\]

\[
\Delta L \text{ is a random symmetric matrix satisfying } \|\Delta L\|_F \leq \varepsilon_1 \|L\|_F,
\]

\[
V_{\text{ion}} = \text{zeros}(n), \quad \Delta V_{\text{ion}} = \varepsilon_2 \cdot \text{sprandsym}(n, 0.5), \quad \Delta \rho(\tilde{P}) = \rho(\tilde{P})^{\varepsilon_3}.
\]

Here, \( n \) is the matrix size, \( \alpha \) denotes the step size, \( \varepsilon_1, \varepsilon_2 \) and \( \varepsilon_3 \) are three parameters used to control the magnitude of the perturbation.

Set \( n = 50, k = 8, \varepsilon_1 = \varepsilon_2 = \varepsilon_3 = \varepsilon = 10^{-j} \) with \( j = 3, 4, \ldots, 12 \). In Fig. 1, we plot \( \chi_s, \xi_s, \tau_s \) versus \( \varepsilon \) for four different step sizes \( \alpha = 0.05, 0.06, 0.07, 0.08 \). In Table 1, we list \( \frac{\xi_s}{d}, \frac{\chi_s}{g - d}, \xi_s, \eta_s, \tau_s, \) and \( \gamma_s \) for different \( \varepsilon \). We can observe that the perturbation bounds \( \xi_s, \tau_s \) and \( \gamma_s \) are good upper bounds for the solution perturbation \( \chi_s \), and \( \tau_s \) is sharper, especially when \( \frac{\xi_s}{d} \) is close to one. And as \( \alpha \) increases, \( \frac{\xi_s}{d} \) decreases, the condition number \( \frac{1}{g - d} \) increases, and as a result, the perturbation bounds become less sharp. Also note that, when \( \alpha = 0.08 \), the assumption of Theorem 1 does not hold since \( \frac{\xi_s}{d} < 2 \), thus, \( \xi_s \) is no longer available (denoted by “-” in Table 1) and we can only use Theorem 2 in this case.

Set \( n = 50, k = 4, \alpha = 0.04 \). Figure 2 displays \( \hat{\xi}_s, \) the error bounds \( \hat{\xi}_s \) and \( \hat{\tau}_s \). We can see from Fig. 2 that as SCF iterations converge, \( \hat{\xi}_s, \hat{\xi}_s \) and \( \hat{\tau}_s \) decrease linearly.
Fig. 1 Logarithm of \( \| \sin \theta(\mathcal{R}(V_x), \mathcal{R}(\tilde{V}_x)) \|_2 \) vs. Logarithm of perturbation bounds on the y-axis against \( \varepsilon \) on the x-axis for the KS equation

The error bounds \( \hat{\xi}_* \) and \( \hat{\tau}_* \) are good upper bounds for \( \hat{\chi}_* \), and the latter one is sharper. Also, note that \( \hat{\tau}_* \) is applicable from the second iteration, meanwhile \( \hat{\xi}_* \) is applicable from the third, which indicates that Corollary 2 has a weaker assumption than that of Corollary 1 in this case.

### 3.2 Application to the trace ratio optimization

We consider the following maximization problem of the sum of the trace ratio:

\[
\max_{V \in \mathbb{R}^{n \times k}, V^T V = I_k} f(V) := \frac{\text{tr}(V^T AV)}{\text{tr}(V^T BV)} + \text{tr}(V^T CV),
\]

(3.3)

where \( \text{tr}(\cdot) \) means the trace of a square matrix, \( A, B, C \in \mathbb{R}^{n \times n} \) are real symmetric with \( B \) being positive definite, and \( k < n/2 \).

As shown in [20], any critical point \( V \) of (3.3) is a solution to the following nonlinear eigenvalue problem

\[
E(V)V = V(V^T E(V)V), \quad V^T V = I_k,
\]

(3.4)
Table 1  Perturbation bounds for the KS equation

| $\varepsilon$ | $g/d$ | $1/(g-d)$ | $\chi^*$ | $\xi^*$ | $\tau^*$ | $\gamma^*$ |
|---------------|-------|---------|--------|-------|-------|-------|
| $\alpha = 0.05$ |       |         |        |       |       |       |
| $10^{-12}$    | 7.3943e+00 | 9.3538e−02 | 7.1017e−14 | 8.8083e−13 | 6.7096e−13 | 8.8083e−13 |
| $10^{-10}$    | 7.3943e+00 | 9.3538e−02 | 6.7042e−12 | 8.8083e−11 | 6.7096e−11 | 8.8083e−11 |
| $10^{-8}$     | 7.3943e+00 | 9.3538e−02 | 6.7173e−10 | 8.8083e−09 | 6.7096e−09 | 8.8083e−09 |
| $10^{-6}$     | 7.3943e+00 | 9.3538e−02 | 6.7171e−08 | 8.8083e−07 | 6.7096e−07 | 8.8083e−07 |
| $10^{-4}$     | 7.3943e+00 | 9.3538e−02 | 6.7171e−06 | 8.8092e−05 | 6.7096e−05 | 8.8083e−05 |
| $\alpha = 0.06$ |       |         |        |       |       |       |
| $10^{-12}$    | 4.2579e+00 | 1.5199e−01 | 1.1136e−13 | 1.5148e−12 | 9.3858e−12 | 1.5148e−12 |
| $10^{-10}$    | 4.2579e+00 | 1.5199e−01 | 1.2152e−11 | 1.5148e−10 | 9.3858e−10 | 1.5148e−10 |
| $10^{-8}$     | 4.2579e+00 | 1.5199e−01 | 1.1823e−09 | 1.5148e−08 | 9.3858e−08 | 1.5148e−08 |
| $10^{-6}$     | 4.2579e+00 | 1.5199e−01 | 1.1822e−07 | 1.5148e−06 | 9.3858e−07 | 1.5148e−06 |
| $10^{-4}$     | 4.2579e+00 | 1.5199e−01 | 1.1822e−05 | 1.5151e−04 | 9.3858e−05 | 1.5148e−04 |
| $\alpha = 0.07$ |       |         |        |       |       |       |
| $10^{-12}$    | 2.5815e+00 | 2.5787e−01 | 1.2296e−13 | 2.6457e−12 | 1.1683e−12 | 2.6457e−12 |
| $10^{-10}$    | 2.5815e+00 | 2.5787e−01 | 1.1013e−11 | 2.6457e−10 | 1.1683e−10 | 2.6457e−10 |
| $10^{-8}$     | 2.5815e+00 | 2.5787e−01 | 1.1030e−09 | 2.6457e−08 | 1.1683e−08 | 2.6457e−08 |
| $10^{-6}$     | 2.5815e+00 | 2.5787e−01 | 1.1030e−07 | 2.6457e−06 | 1.1683e−06 | 2.6457e−06 |
| $10^{-4}$     | 2.5815e+00 | 2.5787e−01 | 1.1030e−05 | 2.6468e−04 | 1.1683e−04 | 2.6457e−04 |
| $\alpha = 0.08$ |       |         |        |       |       |       |
| $10^{-12}$    | 1.6569e+00 | 5.1928e−01 | 2.1045e−13 | – | 1.3537e−12 | 5.4750e−12 |
| $10^{-10}$    | 1.6569e+00 | 5.1928e−01 | 2.6767e−11 | – | 1.3537e−10 | 5.4750e−10 |
| $10^{-8}$     | 1.6569e+00 | 5.1928e−01 | 2.6712e−09 | – | 1.3537e−08 | 5.4750e−08 |
| $10^{-6}$     | 1.6569e+00 | 5.1928e−01 | 2.6709e−07 | – | 1.3537e−06 | 5.4750e−06 |
| $10^{-4}$     | 1.6569e+00 | 5.1928e−01 | 2.6709e−05 | – | 1.3537e−04 | 5.4750e−04 |

where

$$E(V) = A \frac{1}{\phi_B(V)} - B \frac{\phi_A(V)}{\phi_B^2(V)} + C,$$

and for any symmetric matrix $S$, $\phi_S(V)$ is defined as $\phi_S(V) := \text{tr}(V^T SV)$. Moreover, if $V$ is a global maximizer, then it is an orthonormal eigenbasis of $E(V)$ corresponding to its $k$ largest eigenvalues.

Let $P = V V^T$, and note that $\phi_A(V) = \text{tr}(AP)$ and $\phi_B(V) = \text{tr}(BP)$ are functions of $P$, then by setting

$$A_0 = C, \quad A_1(P) = A \frac{1}{\phi_B(V)} - B \frac{\phi_A(V)}{\phi_B^2(V)},$$

\( \square \) Springer
Problem (3.4) can be rewritten as

\[(A_0 + A_1(P))V = V(V^T(A_0 + A_1(P))V), \quad V^T V = I_k,\] (3.5)

which is of the form (1.1).

Suppose that \(A, B, C\) are perturbed slightly, we have the following perturbed equation of (3.5):

\[(\tilde{A}_0 + \tilde{A}_1(\tilde{P}))\tilde{V} = \tilde{V}(\tilde{V}^T(\tilde{A}_0 + \tilde{A}_1(\tilde{P}))\tilde{V}), \quad \tilde{V}^T \tilde{V} = I_k,\] (3.6)

where

\[
\tilde{P} = \tilde{V}\tilde{V}^T, \quad \tilde{A}_0 = A_0 + \Delta C = C + \Delta C,
\]

\[
\tilde{A}_1(\tilde{P}) = (A + \Delta A) \frac{1}{\phi_{B+\Delta B}(\tilde{V})} - (B + \Delta B) \frac{\phi_{A+\Delta A}(\tilde{V})}{\phi_{B+\Delta B}(\tilde{V})},
\]

and \(\Delta A, \Delta B, \Delta C\) are real symmetric matrices.

Then by calculations, we have

\[
\delta_0 = \|\Delta C\|_2,
\]

\[
\delta_1 = \sup_{P \in \tilde{P}_{\tilde{x},\ast}} \|\tilde{A}_1(P) - A_1(P)\|_2 \leq \|A\|_2 \frac{\Omega_{\Delta B}}{\omega_{B+\Delta B} \omega_B} \|\Delta A\|_2 \frac{1}{\omega_{B+\Delta B}^2} + \|B\|_2 \frac{\Omega_{\Delta A} \Omega_{\Delta B}^2 + \Omega_A \Omega_B + \Omega_{B+\Delta B} \Omega_{\Delta B}}{\omega_{B+\Delta B}^2 \omega_B^2} \|\Delta B\|_2 \frac{\Omega_{A+\Delta A}}{\omega_{B+\Delta B}^2}.
\]
where

$$\Omega_W = \sum_{j=n-k+1}^{n} |\lambda_j(W)|, \quad \omega_W = \sum_{j=1}^{k} |\lambda_j(W)|.$$  

Here, $\{\lambda_j(W)\}_{j=1}^{n}$ are the eigenvalues of a Hermitian matrix $W \in \mathbb{C}^{n \times n}$ with

$$|\lambda_1(W)| \leq |\lambda_2(W)| \leq \cdots \leq |\lambda_n(W)|.$$  

To illustrate our theoretical results, we randomly generate the real symmetric matrices $A, B, C, \Delta A, \Delta B, \Delta C$, by using the MATLAB built-in functions `rand`, `randn`, `orth`, `diag` and `ones`:

$$A = \text{rand}(n, n); \quad A = (A' + A)/2; \quad Q = \text{orth}([\text{rand}(n, n)];$$

$$B = Q \ast \text{diag}(50 + \beta \ast (2 \ast \text{rand}(n, 1) - \text{ones}(n, 1))) \ast Q';$$

$$\Delta A = A - A_0; \quad \Delta B = B - B_0; \quad \Delta C = C - C_0.$$  

Fig. 3 Logarithm of $\| \sin(\theta(\hat{V}_s), \hat{V}_s)) \|_2$ vs. Logarithm of perturbation bounds on the y-axis against $\varepsilon$ on the x-axis for the trace ratio optimization.
For simplicity, we fix \( n = 100, k = 5, \) and \( \beta = 10. \) Figure 3 plots \( \chi_* \), and the perturbation bounds \( \xi_* \) and \( \tau_* \) for varying \( \varepsilon. \) Figure 4 shows \( \hat{\chi}_* \) versus the error bounds \( \hat{\xi}_* \) and \( \hat{\tau}_* \) for different \( \beta \) in terms of the SCF iterations.

We observe from Fig. 3 that when \( \frac{\delta}{\beta} > 2, \) both the assumptions of Theorems 1 and 2 hold. In this case, the perturbation bounds \( \xi_* \) and \( \tau_* \) are good upper bounds for the solution perturbation \( \chi_* \) when \( \delta \) is small, while the perturbation bound \( \tau_* \) is slightly sharper than \( \xi_* \). However, when \( 1 < \frac{\delta}{\beta} < 2, \) only the assumption of Theorem 2 holds under the assumption that \( \delta \) is sufficient small. In this case, the perturbation bound \( \tau_* \) is good upper bounds for the solution perturbation \( \chi_* \). We must point out that when \( \varepsilon \) is close to or equal to \( 10^{-2}, \) \( \delta \) is not sufficiently small and (2.19) has no positive root. In this case, Theorem 2 is not applicable. We have the similar observation for Fig. 4 on \( \hat{\chi}_* \) and the error bounds \( \hat{\xi}_* \) and \( \hat{\tau}_* \) in terms of the SCF iterations.
Table 2  Perturbation bounds for the trace ratio optimization

| $\beta$ | $g/d$     | $1/(g-d)$ | $\chi_*$ | $\xi_*$ | $\tau_*$ | $\gamma_*$ |
|---------|-----------|-----------|----------|---------|---------|-----------|
| $\varepsilon = 10^{-12}$ |           |           |          |         |         |           |
| 5       | 2.7524e+00 | 3.8890e+00 | 1.0404e−12 | 3.2359e−11 | 3.2359e−11 | 3.2359e−11 |
| 8       | 2.1591e+00 | 4.6130e+00 | 1.0399e−12 | 3.8604e−11 | 3.8604e−11 | 3.8604e−11 |
| 10      | 1.7640e+00 | 5.7162e+00 | 1.0397e−12 | 4.8148e−11 | 4.8148e−11 | 4.8148e−11 |
| 12      | 1.4396e+00 | 8.1084e+00 | 1.0412e−12 | 6.8860e−11 | 6.8860e−11 | 6.8860e−11 |
| 15      | 1.0762e+00 | 3.4970e+01 | 1.0429e−12 | 3.0160e−10 | 3.0160e−10 | 3.0160e−10 |
| $\varepsilon = 10^{-6}$ |           |           |          |         |         |           |
| 5       | 2.7524e+00 | 3.8890e+00 | 1.0408e−06 | 3.2359e−05 | 3.2359e−05 | 3.2359e−05 |
| 8       | 2.1591e+00 | 4.6130e+00 | 1.0409e−06 | 3.8604e−05 | 3.8604e−05 | 3.8604e−05 |
| 10      | 1.7640e+00 | 5.7162e+00 | 1.0406e−06 | 4.8148e−05 | 4.8148e−05 | 4.8148e−05 |
| 12      | 1.4396e+00 | 8.1084e+00 | 1.0407e−06 | 6.8860e−05 | 6.8860e−05 | 6.8860e−05 |
| 15      | 1.0762e+00 | 3.4970e+01 | 1.0407e−06 | 3.0160e−04 | 3.0160e−04 | 3.0160e−04 |
| $\varepsilon = 10^{-4}$ |           |           |          |         |         |           |
| 5       | 2.7524e+00 | 3.8890e+00 | 1.0407e−04 | 3.2359e−03 | 3.2359e−03 | 3.2359e−03 |
| 8       | 2.1591e+00 | 4.6130e+00 | 1.0409e−04 | 3.8604e−03 | 3.8604e−03 | 3.8604e−03 |
| 10      | 1.7640e+00 | 5.7162e+00 | 1.0406e−04 | 4.8148e−03 | 4.8148e−03 | 4.8148e−03 |
| 12      | 1.4396e+00 | 8.1084e+00 | 1.0407e−04 | 6.8860e−03 | 6.8860e−03 | 6.8860e−03 |
| 15      | 1.0762e+00 | 3.4970e+01 | 1.0407e−04 | 3.0358e−02 | 3.0160e−02 | 3.0160e−02 |

To further illustrate our theoretical results, in Table 2, we report the estimated values of $\frac{g}{d}$ and $\frac{1}{g-d}$, the solution perturbation $\chi_*$, the perturbation bounds $\xi_*$, $\tau_*$, and $\gamma_*$ for fixed $\varepsilon$ and varying $\beta$, where the symbol “−” means the upper bound $\xi_*$ is not available since the assumption of Theorem 1 does not hold. Also, Table 3 displays the estimated values of $\frac{\hat{g}}{d}$ and $\frac{1}{\hat{g}-d}$, the solution perturbation $\hat{\chi}_*$, the error bounds $\hat{\xi}_*$, $\hat{\tau}_*$, and $\hat{\gamma}_*$ for varying $\beta$ in terms of the SCF iterations, where the symbol “−” means the corresponding error bound is not available since the assumption of Corollary 1 or Corollary 2 does not hold or the perturbation $\|R\|_2$ is not sufficiently small.

We see from Table 2 that, for a fixed $\varepsilon$ and different $\beta$, the estimated values of $\xi_*$, $\tau_*$, and $\gamma_*$ are valid upper bounds for the solution perturbation bound $\chi_*$. We also see that $\tau_*$ is slightly sharper than $\xi_*$ and the assumption of Theorem 2 is weaker than that of Theorem 1. We have the similar observation for Table 3 on $\hat{\chi}_*$ and the error bounds $\hat{\xi}_*$, $\hat{\tau}_*$, and $\hat{\gamma}_*$ in terms of the SCF iterations.

### 3.3 Comparison results for the trace ratio optimization

In [22], a perturbation bound was derived for a special case of the trace ratio optimization (3.3) discussed in Sect. 3.2, i.e.,

$$
\max_{V \in \mathbb{R}^{n \times k}, V^TV = I_k} f(V) := \frac{\text{tr}(V^TAV)}{\text{tr}(V^TBV)}.
$$

(3.7)
where $A, B \in \mathbb{R}^{n \times n}$ are real symmetric with $B$ being positive definite and $k < n/2$.

The first-order necessary optimality condition of (3.7) is given by

$$E(V)V = V(V^T E(V)V), \quad V^TV = I_k, \quad (3.8)$$

where

$$E(V) = A - B \frac{\phi_A(V)}{\phi_B(V)}.$$ 

Let

$$A_0 = A \quad \text{and} \quad A_1(P) = -B \frac{\phi_A(V)}{\phi_B(V)}.$$ 

Then the condition (3.8) takes the form of

$$(A_0 + A_1(P))V = V(V^T (A_0 + A_1(P))V), \quad V^TV = I_k,$$

which is of the form (1.1).
Suppose that $A$ and $B$ are perturbed slightly, we get a perturbed equation of (3.7):

$$
(\tilde{A}_0 + \tilde{A}_1(\tilde{P}))\tilde{V} = \tilde{V}(\tilde{V}^T(\tilde{A}_0 + \tilde{A}_1(\tilde{P})))\tilde{V}, \quad \tilde{V}^T\tilde{V} = I_k,
$$

where

$$
\tilde{P} = \tilde{V}^T\tilde{V}, \quad \tilde{A}_0 = A_0 + \Delta A = A + \Delta A,
\tilde{A}_1(\tilde{P}) = -(B + \Delta B)\frac{\phi_{A+\Delta A}(\tilde{V})}{\phi_{B+\Delta B}(\tilde{V})},
$$

and $\Delta A, \Delta B \in \mathbb{R}^{n \times n}$ are real symmetric matrices.

By simple calculations we have

$$
\delta_0 = \|\Delta A\|_2, \\
\delta_1 = \sup_{P \in \mathbb{R}^{n \times n}} \|\tilde{A}_1(P) - A_1(P)\|_2 \\
\leq \|B\|_2 \frac{\Omega_A \omega_B \Omega_B + \Omega_B \omega_B \Omega_A}{\omega_B + \Delta B \omega_B} + \|\Delta B\|_2 \frac{\Omega_A + \Delta A}{\omega_B + \Delta B}.
$$

Springer
Table 4  Perturbation bounds for the trace ratio optimization

| $\beta$ | $g/d$ | $1/(g-d)$ | $\chi_*$ | $\xi_*$ | $\tau_*$ | $\psi_*$ |
|--------|--------|-----------|-----------|---------|----------|---------|
| $\varepsilon = 10^{-12}$ |        |           |           |         |          |         |
| 5      | 2.0352e+00 | 1.5244e+00 | 5.5900e−13 | 7.0419e−12 | 7.0419e−12 | 1.4615e−11 |
| 8      | 1.7412e+00 | 1.7581e+00 | 5.7535e−13 | –       | 8.8663e−12 | 1.4838e−11 |
| 10     | 1.4450e+00 | 2.8148e+00 | 5.3240e−13 | –       | 1.4503e−11 | 1.7400e−11 |
| 12     | 1.7058e+00 | 1.8362e+00 | 5.2163e−13 | –       | 9.2758e−12 | 1.5017e−11 |
| 15     | 1.7739e+00 | 1.4791e+00 | 5.3977e−13 | –       | 7.8708e−12 | 1.3306e−11 |
| 18     | 1.4108e+00 | 2.2015e+00 | 4.5183e−13 | –       | 1.3650e−11 | 1.4246e−11 |
| 20     | 1.0681e+00 | 1.2859e+01 | 4.2000e−13 | –       | 8.0079e−11 | 1.8176e−11 |
| $\varepsilon = 10^{-6}$ |        |           |           |         |          |         |
| 5      | 2.0352e+00 | 1.5244e+00 | 5.5824e−07 | 7.0419e−06 | 7.0419e−06 | 1.4615e−05 |
| 8      | 1.7412e+00 | 1.7581e+00 | 5.7556e−07 | –       | 8.8663e−06 | 1.4838e−05 |
| 10     | 1.4450e+00 | 2.8148e+00 | 5.3283e−07 | –       | 1.4503e−05 | 1.7400e−05 |
| 12     | 1.7058e+00 | 1.8362e+00 | 5.2203e−07 | –       | 9.2758e−06 | 1.5017e−05 |
| 15     | 1.7739e+00 | 1.4791e+00 | 5.4021e−07 | –       | 7.8708e−06 | 1.3306e−05 |
| 18     | 1.4108e+00 | 2.2015e+00 | 4.5290e−07 | –       | 1.3650e−05 | 1.4246e−05 |
| 20     | 1.0681e+00 | 1.2859e+01 | 4.2604e−07 | –       | 8.0079e−05 | 1.8176e−05 |
| $\varepsilon = 10^{-4}$ |        |           |           |         |          |         |
| 5      | 2.0352e+00 | 1.5244e+00 | 5.5824e−05 | 7.0516e−04 | 7.0419e−04 | 1.4615e−03 |
| 8      | 1.7412e+00 | 1.7581e+00 | 5.7556e−05 | –       | 8.8663e−04 | 1.4838e−03 |
| 10     | 1.4450e+00 | 2.8148e+00 | 5.3283e−05 | –       | 1.4503e−03 | 1.7400e−03 |
| 12     | 1.7058e+00 | 1.8362e+00 | 5.2203e−05 | –       | 9.2758e−04 | 1.5017e−03 |
| 15     | 1.7739e+00 | 1.4791e+00 | 5.4021e−05 | –       | 7.8708e−04 | 1.3306e−03 |
| 18     | 1.4108e+00 | 2.2015e+00 | 4.5290e−05 | –       | 1.3650e−03 | 1.4246e−03 |
| 20     | 1.0681e+00 | 1.2859e+01 | 4.2605e−05 | –       | 8.0119e−03 | 1.8176e−03 |

\[ d = \sup_{P \neq P_*, P \in \mathcal{P}_{\varepsilon_*}} \frac{\|A_1(P) - A_1(P_*)\|_2}{\|P - P_*\|_2} \leq \frac{2\|A\|_2\|B\|_2^2}{\omega_B^2}. \]

We now compare the proposed perturbation bounds in Theorems 1 and 2 with the perturbation bound in [22], which is given by

\[ \psi_* := \frac{4(\|\Delta A\|_2 + f(V_*)\|\Delta B\|_2)}{g} \left(1 + k \frac{\|B + \Delta B\|_2}{\omega_B} \right). \]

For simplicity, we fix $n = 10$, $k = 4$, and $\beta = 8$. We randomly generate the real symmetric matrices $A$, $B$, $\Delta A$, $\Delta B$ as in Sect. 3.2. Figure 5 plots $\chi_*$, and the perturbation bounds $\xi_*$, $\tau_*$, and $\psi_*$ for varying $\varepsilon$. Table 4 displays the estimated values of $\frac{g}{d}$ and $\frac{1}{g-d}$, the solution perturbation $\chi_*$, the perturbation bounds $\xi_*$, $\tau_*$, and $\psi_*$ for fixed $\varepsilon$ and varying $\beta$. 
We see from Fig. 5 and Table 4 that the proposed perturbation bounds are slightly more accurate than the perturbation bound in [22] when $\frac{g}{d}$ is not close to 1 while the perturbation bound in [22] is slightly sharper than the proposed perturbation bound $\tau_*$ when $\frac{g}{d}$ is close to 1. We point out that our method needs more numerical computational efforts including $\Omega_A \Delta A$, $\Omega_B \Delta B$, $\Omega A + \Delta A$, and $\omega_B + \Delta B$.

4 Conclusion

In this paper, we have studied the perturbation theory of NEPv (1.1). Two perturbation bounds are established, based on which the condition number for the NEPv can be introduced. Furthermore, two computable error bounds are also obtained. Theoretical results are applied to the KS equation and the trace ratio problem. Numerical results show that both the perturbation bounds and the error bounds are fairly sharp, especially the perturbation bound in Theorem 2 and the error bound in Corollary 2.

Acknowledgements We are grateful to the referees for their valuable comments and suggestions. We also would like to thank Prof. Ren-Cang Li from the University of Texas at Arlington for helpful comments on a preliminary draft of this paper.

References

1. Bao, W., Du, Q.: Computing the ground state solution of Bose–Einstein condensates by a normalized gradient flow. SIAM J. Sci. Comput. 25, 1674–1697 (2006)
2. Cai, Y., Zhang, L.-H., Bai, Z., Li, R.-C.: On an eigenvector-dependent nonlinear eigenvalue problem. SIAM J. Matrix Anal. Appl. 39, 1360–1382 (2018)
3. Chen, H., Dai, X., Gong, X., He, L., Zhou, A.: Adaptive finite element approximations for Kohn–Sham models. Multiscale Model. Simul. 12, 1828–1869 (2014)
4. Davis, C., Kahan, W.M.: The rotation of eigenvectors by a perturbation. III. SIAM J. Numer. Anal. 7, 1–46 (1970)
5. Jarlebring, E., Kvaal, S., Mechiels, W.: An inverse iteration method for eigenvalue problems with eigenvector nonlinearities. SIAM J. Sci. Comput. 36, A1978–A2001 (2014)
6. Jia, S.-H., Xie, H.-H., Xie, M.-T., Xu, F.: A full multigrid method for nonlinear eigenvalue problems. Sci. China Math. 59, 2037–2048 (2016)
7. Khamsi, M.A.: Introduction to metric fixed point theory. In: Almezel, S., Ansari, Q.H., Khamsi, M.A. (eds.) Topics in Fixed Point Theory, pp. 1–32. Springer, Cham (2014)
8. Lin, W.-W., Sun, J.-G.: Perturbation analysis of the periodic discrete-time algebraic Riccati equation. SIAM J. Matrix Anal. Appl. 24, 411–438 (2002)
9. Liu, X., Wang, X., Wen, Z., Yuan, Y.: On the convergence of the self-consistent field iteration in Kohn–Sham density functional theory. SIAM J. Matrix Anal. Appl. 35, 546–558 (2014)
10. Liu, X., Wen, Z., Wang, X., Ulbrich, M., Yuan, Y.: On the analysis of the discretized Kohn–Sham density functional theory. SIAM J. Matrix Anal. Appl. 53, 1758–1785 (2015)
11. Martin, R.M.: Electronic Structure: Basic Theory and Practical Methods. Cambridge University Press, Cambridge (2004)
12. Ngo, T., Bellalij, M., Saad, Y.: The trace ratio optimization problem for dimensionality reduction. SIAM J. Matrix Anal. Appl. 31, 2950–2971 (2010)
13. Rice, J.: A theory of condition. SIAM. J. Numer. Anal. 3, 287–310 (1966)
14. Saad, Y., Chelikowsky, J.R., Shontz, S.M.: Numerical methods for electronic structure calculations of materials. SIAM Rev. 52, 3–54 (2010)
15. Stewart, G.W., Sun, J.g: Matrix Perturbation Theory. Academic Press, Boston (1990)
16. Sun, J.-G.: Perturbation theory for algebraic Riccati equations. SIAM J. Matrix Anal. Appl. 19, 39–65 (1998)
17. Sun, J.-G.: Perturbation analysis of the matrix equation $X = Q + A^H (\hat{X} - C)^{-1} A$. Linear Algebra Appl. 372, 33–51 (2003)
18. Sun, J.-G., Xu, S.-F.: Perturbation analysis of the maximal solution of the matrix equation $X + A^* X^{-1} A = P$. II. Linear Algebra Appl. 362, 211–228 (2003)
19. Yang, C., Meza, J.C., Wang, L.-W.: A trust region direct constrained minimization algorithm for the Kohn–Sham equation. SIAM J. Sci. Comput. 29, 1854–1875 (2007)
20. Zhang, L.-H., Li, R.-C.: Maximization of the sum of the trace ratio on the Stiefel manifold, I: theory. Sci. China Math. 57, 2495–2508 (2014)
21. Zhang, L.-H., Li, R.-C.: Maximization of the sum of the trace ratio on the Stiefel manifold, II: computation. Sci. China Math. 58, 1549–1566 (2015)
22. Zhang, L.-H., Yang, W.H.: Perturbation analysis for the trace quotient problem. Linear Multilinear Algebra 61, 1629–1640 (2013)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Yunfeng Cai1 · Zhigang Jia2 · Zheng-Jian Bai3

Yunfeng Cai
yfcai1116@gmail.com

Zhigang Jia
zhgjia@jsnu.edu.cn

1 Cognitive Computing Lab, Baidu Research, Beijing 100193, China
2 School of Mathematics and Statistics and Jiangsu Key Laboratory of Education Big Data Science and Engineering, Jiangsu Normal University, Xuzhou 221116, China
3 School of Mathematical Sciences and Fujian Provincial Key Laboratory on Mathematical Modeling and High Performance Scientific Computing, Xiamen University, Xiamen 361005, China