On an Eigenvector-Dependent Nonlinear Eigenvalue Problem

Yunfeng Cai[*], Lei-Hong Zhang[†], Zhaojun Bai[‡], and Ren-Cang Li[§]

November 30, 2017

Abstract

We first provide existence and uniqueness conditions for the solvability of an algebraic eigenvalue problem with eigenvector nonlinearity. We then present a local and global convergence analysis for a self-consistent field (SCF) iteration for solving the problem. The well-known $\sin\Theta$ theorem in the perturbation theory of Hermitian matrices plays a central role. The near-optimality of the local convergence rate of the SCF iteration revealed in this paper are demonstrated by examples from the discrete Kohn-Sham eigenvalue problem in electronic structure calculations and the maximization of the trace ratio in the linear discriminant analysis for dimension reduction.

Keywords. Nonlinear eigenvalue problem, Self-consistent-field iteration, convergence analysis

AMS subject claasifications. 15A18, 65F15, 65H17, 47J10

1 Introduction

The eigenvector-dependent nonlinear eigenvalue problem (NEPv) is to find $V \in \mathbb{C}^{n\times k}$ with orthonormal columns and $\Lambda \in \mathbb{C}^{k\times k}$ such that

$$H(V)V = V\Lambda, \tag{1.1}$$

where $H(V) \in \mathbb{C}^{n\times n}$ is an Hermitian matrix-valued function of $V \in \mathbb{C}^{n\times k}$ with orthonormal columns, i.e., $V^{\mathrm{H}}V = I_k$, $k \le n$ (usually $k \ll n$). Immediately, we infer from (1.1) that $\Lambda = V^{\mathrm{H}}H(V)V$, necessarily Hermitian, and the eigenvalues of Λ are k of the n eigenvalues of $H(V)$. For the problem of practical interests, they are usually either the k smallest or the k largest eigenvalues of $H(V)$. We will state all our results for the case of the smallest eigenvalues. But they are equally valid if the word "smallest" is replaced by "largest".

Often the dependency on V of $H(V)$ satisfies

$$H(V) \equiv H(VQ) \quad \text{for any unitary } Q \in \mathbb{C}^{k\times k}. \tag{1.2}$$

The condition (1.2) implies that $H(V)$ is a function of k-dimensional subspaces of $\mathbb{C}^n$, or equivalently, a function on the complex Grassmann manifold $\mathscr{G}_k(\mathbb{C}^n)$. In particular, if V is a solution, then so is

[*]LMAM & School of Mathematical Sciences, Peking University, Beijing, 100871, China, yfcai@math.pku.edu.cn

[†]School of Mathematics and Research School for Interdisciplinary Sciences, Shanghai University of Finance and Economics, 777 Guoding Road, Shanghai 200433, China, zhang.leihong@mail.shufe.edu.cn

[‡]Department of Computer Science and Department of Mathematics, University of California, Davis, CA 95616, USA, zbai@ucdavis.edu

[§]Department of Mathematics, University of Texas at Arlington, Arlington, TX 76019-0408, USA, rcli@uta.edu

VQ for any $k \times k$ unitary matrix Q. Therefore, any solution V to (1.1) essentially represents a class $\{VQ : Q \in \mathbb{C}^{k\times k}, Q^{\mathrm{H}}Q = I_k\}$ each of which solves (1.1). In light of this, we say that the solution to (1.1) is unique if V, $\widetilde{V}$ are two solutions to (1.1), then $\mathcal{R}(V) = \mathcal{R}(\widetilde{V})$, where $\mathcal{R}(V)$ and $\mathcal{R}(\widetilde{V})$ are the column subspaces of V and $\widetilde{V}$, respectively. While we will assume (1.2) throughout this paper, all our results can be extended with minor modifications to work for NEPv (1.1) without (1.2).

The most well-known origin of NEPv (1.1) is from Kohn–Sham density functional theory in electronic structure calculations, see [14, 19, 5] and references therein. NEPv (1.1) also arises from the discretized Gross-Pitaevskii equation for modeling particles in the state of matter called the Bose-Einstein condensate [1, 7, 8], optimization of the trace ratio in the linear discriminant analysis for dimension reduction [15], and balanced graph cuts [9].

In the first part of this paper, we present two sufficient conditions for the existence and uniqueness of the solution of NEPv (1.1). One is a Lipschitz-like condition on the matrix-value function $H(V)$. The other is a uniform gap between the kth and $(k+1)$st smallest eigenvalues of $H(V)$, known as the "uniform well-posedness" property for the Hatree-Fock equation in electronic structure calculations [4]. To the best of our knowledge, it is the first such kind of results on the existence and uniqueness of the solution of NEPv (1.1) from the linear algebraic point of view.

Self-consistent field (SCF) iteration is the most widely used algorithm to solve NEPv (1.1), see [14, 19] and references therein. It is conceivably a natural one to try. At the ith SCF iteration, one computes an approximation to the eigenvector matrix V_i associated with the k smallest eigenvalues of $H(V_{i-1})$ evaluated at the previous approximation V_{i-1}, and then V_i is used as the next approximation to the solution of NEPv (1.1). When the iterative process converges, the computed eigenvectors are said to be self-consistent. In the second part of this paper, we provide a local and global convergence analysis of a plain SCF iteration for solving NEPv (1.1). We use two examples to show the near-optimality of the newly established local convergence rate. We closely examine applications of derived convergence results to electronic structure calculations and linear discriminant analysis for dimension reduction. In particular, with weaker assumptions, we can significantly improve previous convergence results in [13, 25] on the SCF iteration for solving the discrete Kohn-Sham NEPv.

We will begin the presentation in section 2 with a review of matrix norms, angles between subspaces and perturbation bounds to be used in this paper. In section 3, we establish the existence and uniqueness of NEPv (1.1) under a Lipschitz-like condition and uniform well-posedness of the eigenvalue gap of $H(V)$. In section 4, we start by stating a plain SCF iteration for solving NEPv (1.1), and then establish local and global convergence results for the SCF iteration. In section 5, we discuss two applications. Concluding remarks are in section 6.

Notation. $\mathbb{C}^{n\times m}$ is the set of all $n \times m$ matrices with complex entries, $\mathbb{C}^n = \mathbb{C}^{n\times 1}$, and $\mathbb{C} = \mathbb{C}^1$. Correspondingly, we will use $\mathbb{R}^{n\times m}$, $\mathbb{R}^n$, and $\mathbb{R}$ for their counterparts for the real number case. The superscripts "$\cdot^{\mathrm{T}}$" and "$\cdot^{\mathrm{H}}$" take the transpose and the complex conjugate transpose of a matrix or vector, respectively. $\mathbb{U}^{n\times k} = \{V \mid V \in \mathbb{C}^{n\times k}, V^{\mathrm{H}}V = I_k\}$, i.e., the set of all $n \times k$ complex matrices with orthonormal columns, and $\mathscr{G}_k(\mathbb{C}^n)$ denotes the complex Grassmann manifold of all k-dimensional subspaces of $\mathbb{C}^n$. 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. $\mathcal{R}(X)$ is the column space of matrix X. Denote by $\lambda_j(H)$ for $1 \le j \le n$ the eigenvalues of a Hermitian matrix $H \in \mathbb{C}^{n\times n}$ and they are always arranged in nondecreasing order: $\lambda_1(H) \le \lambda_2(H) \le \cdots \le \lambda_n(H)$. $\mathrm{Diag}(x)$ denotes the diagonal matrix with the vector x on its diagonal. $\mathrm{diag}(A)$ stands for the column vector containing the diagonal elements of the the matrix A.

2 Preliminaries

For completeness, in this section, we review matrix norms, angles between subspaces, and perturbation bounds to be used later in this paper.

Unitarily invariant norm. A matrix norm $\|\cdot\|_{\rm ui}$ is called a *unitarily invariant norm* on $\mathbb{C}^{m\times n}$ if it is a matrix norm and has the following two properties:

1. $\|X^{\rm H}AY\|_{\rm ui} = \|A\|_{\rm ui}$ for all unitary matrices X and Y.
2. $\|A\|_{\rm ui} = \|A\|_2$ whenever A is of rank one, where $\|\cdot\|_2$ is the spectral norm.

Two commonly used unitarily invariant norms are

$$\text{the spectral norm: } \|A\|_2 = \max_j \sigma_j,$$
$$\text{the Frobenius norm: } \|A\|_{\rm F} = (\sum_j \sigma_j^2)^{1/2},$$

where $\sigma_1, \sigma_2, \ldots, \sigma_{\min\{m,n\}}$ are the singular values of A, see, e.g., [2, 20].

In this article, for convenience, any $\|\cdot\|_{\rm ui}$ we use is generic to matrix sizes in the sense that it applies to matrices of all sizes [20, p.79]. Examples include the matrix spectral norm $\|\cdot\|_2$ and the Frobenius norm $\|\cdot\|_{\rm F}$. One important property of unitarily invariant norms is

$$\|ABC\|_{\rm ui} \le \|A\|_2 \cdot \|B\|_{\rm ui} \cdot \|C\|_2 \tag{2.1}$$

for any matrices A, B, and C of compatible sizes. Comparing $\|\cdot\|_2$ with any $\|\cdot\|_{\rm ui}$, we have

$$\|A\|_2 \le \|A\|_{\rm ui} \le \min\{m,n\}\|A\|_2 \tag{2.2}$$

for any $A \in \mathbb{C}^{m\times n}$. Sharper bounds than this are possible for a particular unitarily invariant norm. For example,

$$\|A\|_2 \le \|A\|_{\rm F} \le \sqrt{\min\{m,n\}}\,\|A\|_2. \tag{2.3}$$

Angles between subspaces. Consider the complex Grassmann manifold $\mathscr{G}_k(\mathbb{C}^n)$ consisting of all k-dimensional subspaces of $\mathbb{C}^n$, and let $\mathcal{X}$, $\mathcal{Y} \in \mathscr{G}_k(\mathbb{C}^n)$. Let $X, Y \in \mathbb{C}^{n\times k}$ be the orthonormal basis matrices of $\mathcal{X}$ and $\mathcal{Y}$, respectively, i.e.,

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

and let σ_j for $1 \le j \le k$ be the singular values of $Y^{\rm H}X$ in ascending order, i.e., $\sigma_1 \le \cdots \le \sigma_k$, then the k *canonical angles* $\theta_j(\mathcal{X},\mathcal{Y})$ *between* $\mathcal{X}$ *to* $\mathcal{Y}$ are defined by

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

They are in descending order, i.e., $\theta_1(\mathcal{X},\mathcal{Y}) \ge \cdots \ge \theta_k(\mathcal{X},\mathcal{Y})$. Set

$$\Theta(\mathcal{X},\mathcal{Y}) = \mathrm{Diag}(\theta_1(\mathcal{X},\mathcal{Y}), \ldots, \theta_k(\mathcal{X},\mathcal{Y})). \tag{2.5}$$

It can be seen that angles so defined are independent of the orthonormal basis matrices X and Y. A different way to define these angles is through the orthogonal projections onto $\mathcal{X}$ and $\mathcal{Y}$ [24]. Note

that when $k = 1$, i.e., X and Y are vectors, there is only one canonical angle between $\mathcal{X}$ and $\mathcal{Y}$ and so we will simply write $\theta(\mathcal{X}, \mathcal{Y})$. With the definition of canonical angles, Sun [21, p.95] proved that for any unitarily invariant norm $\|\cdot\|_{\mathrm{ui}}$ on $\mathbb{C}^{k\times k}$, $\|\sin\Theta(\mathcal{X}, \mathcal{Y})\|_{\mathrm{ui}}$ defines a unitarily invariant metric on $\mathscr{G}_k(\mathbb{C}^n)$. A different proof can be found in [18].

In what follows, we sometimes place a vector or matrix in one or both arguments of $\theta_j(\,\cdot\,,\,\cdot\,)$, $\theta(\,\cdot\,,\,\cdot\,)$, and $\Theta(\,\cdot\,,\,\cdot\,)$ with the understanding that it is about the subspace spanned by the vector or the columns of the matrix argument. The following lemma provides a convenient way to compute $\|\sin\Theta(\mathcal{X}, \mathcal{Y})\|_{\mathrm{ui}}$.

Lemma 1. *Let $[X,\, X_{\mathrm{c}}]$ and $[Y,\, Y_{\mathrm{c}}]$ be two unitary matrices with $X, Y \in \mathbb{C}^{n\times k}$. Then*

$$\|\sin\Theta(X, Y)\|_{\mathrm{ui}} = \|X_{\mathrm{c}}^{\mathrm{H}} Y\|_{\mathrm{ui}} = \|X^{\mathrm{H}} Y_{\mathrm{c}}\|_{\mathrm{ui}}$$

for any unitarily invariant norm $\|\cdot\|_{\mathrm{ui}}$ on $\mathbb{C}^{k\times k}$.

Because orthonormal bases for subspaces are not unique, two subspaces $\mathcal{X}$ and $\mathcal{Y}$ of dimension k are close in terms of their canonical angles, or equivalently some norm $\|\sin\Theta(\mathcal{X}, \mathcal{Y})\|_{\mathrm{ui}}$ can have orthonormal basis matrices $X,\, Y \in \mathbb{C}^{n\times k}$ that are far apart in the sense that $\|X - Y\|_{\mathrm{ui}} \gg \|\sin\Theta(\mathcal{X}, \mathcal{Y})\|_{\mathrm{ui}}$. The next lemma whose proof can be found in [27, Lemma 4.1] says that one can always choose the basis matrices that differ from each other by $O(\|\sin\Theta(\mathcal{X}, \mathcal{Y})\|_{\mathrm{ui}})$.

Lemma 2 ([27, Lemma 4.1])**.** *Suppose $X,\, Y \in \mathbb{U}^{n\times k}$. Then there exists a unitary matrix $Q \in \mathbb{R}^{k\times k}$ such that*

$$\|\sin\Theta(X, Y)\|_{\mathrm{ui}} \le \|XQ - Y\|_{\mathrm{ui}} \le \sqrt{2}\,\|\sin\Theta(X, Y)\|_{\mathrm{ui}}, \tag{2.6}$$

for any unitarily invariant norm $\|\cdot\|_{\mathrm{ui}}$.

Each $\mathcal{X} \in \mathscr{G}_k(\mathbb{C}^n)$ can be represented uniquely by the orthogonal projector $P_{\mathcal{X}}$ onto the subspace $\mathcal{X}$. Given $X^{\mathrm{H}}X = I_k$ such that $\mathcal{X} = \mathcal{R}(X)$, we have

$$P_{\mathcal{X}} = P_X := XX^{\mathrm{H}}.$$

Note that even though P_X is explicitly defined by X, it is independent of the choice of X so long as $\mathcal{R}(X) = \mathcal{X}$. Therefore, any norm on the differences between the orthogonal projectors induces a metric on $\mathscr{G}_k(\mathbb{C}^n)$. Naturally, we would ask if there is any relation between $\|\sin\Theta(X, Y)\|_{\mathrm{ui}}$ and $\|P_X - P_Y\|_{\mathrm{ui}}$. Indeed, for any $X,\, Y \in \mathbb{U}^{n\times k}$, we have

$$\|\sin\Theta(X, Y)\|_2 = \|P_X - P_Y\|_2, \quad \|\sin\Theta(X, Y)\|_{\mathrm{F}} = \frac{1}{\sqrt{2}}\|P_X - P_Y\|_{\mathrm{F}}. \tag{2.7}$$

Both equalities in (2.7) are the simple consequences of the fact that the singular values of $P_X - P_Y$ consists of each $\sin\theta_i(X, Y)$ repeated twice and $n - 2k$ zeros [20, p.43]. In addition, we have

$$\frac{1}{2}\|P_X - P_Y\|_{\mathrm{ui}} \le \|\sin\Theta(X, Y)\|_{\mathrm{ui}} \le \|P_X - P_Y\|_{\mathrm{ui}} \tag{2.8}$$

for any unitarily invariant norm $\|\cdot\|_{\mathrm{ui}}$. Closely related, the singular values of $P_X(I - P_Y)$ consists of all $\sin\theta_i(X, Y)$ and $n - k$ zeros [20, p.43], and therefore

$$\|\sin\Theta(X, Y)\|_{\mathrm{ui}} = \|P_X(I - P_Y)\|_{\mathrm{ui}} = \|P_Y(I - P_X)\|_{\mathrm{ui}}. \tag{2.9}$$

In the rest of this paper we will treat $\mathcal{X}$, $P_{\mathcal{X}}$, and P_X indistinguishably whenever convenient.

Perturbation of Hermitian matrices. A well-known theorem of Weyl is the following.

Lemma 3 ([20, p.203])**.** *For two Hermitian matrices A, $\widetilde{A} \in \mathbb{C}^{n\times n}$, we have*

$$|\lambda_j(A) - \lambda_j(\widetilde{A})| \le \|A - \widetilde{A}\|_2 \quad \text{for } 1 \le j \le n.$$

The next lemma is essentially [6, Theorem 5.1].

Lemma 4 ([6])**.** *Let H and M be two Hermitian matrices, and let S be a matrix of a compatible size as determined by the Sylvester equation*

$$HY - YM = S.$$

If either all eigenvalues of H are contained in a closed interval that contains no eigenvalue of M or vice versa, then the Sylvester equation has a unique solution Y, and moreover

$$\|Y\|_{\mathrm{ui}} \le \frac{1}{\delta}\|S\|_{\mathrm{ui}},$$

where $\delta = \min|\mu - \omega|$ over all eigenvalues μ of M and all eigenvalues ω of H.

The Davis-Kahan $\sin\Theta$ theorem [6] (see also [12, 20]) will play a central role in our later analysis. However, we will not explicitly state the theorem here for two reasons. The first reason is that it does not take up much of page space to infer the Davis-Kahan $\sin\Theta$ theorem from Lemma 4, and the second one is that we can derive a better locally convergent rate of the SCF iteration by going through the actual proof of the theorem.

3 Existence and Uniqueness

Recall (1.2) about the dependency of $H(V)$ on V, which makes $H(\cdot)$ a Hermitian matrix-valued function on the complex Grassmann manifold $\mathscr{G}_k(\mathbb{C}^n)$. For convenience, we will treat $H(V)$ and $H(\mathcal{V})$ indistinguishably, where $\mathcal{V} = \mathcal{R}(V)$. As a convention, $\lambda_j(H(V))$ for $1 \le j \le n$ are the eigenvalues of $H(V)$, arranged in nondecreasing order. The following theorem gives sufficient conditions for the existence and uniqueness of the solution of NEPv (1.1).

Theorem 1. *Assume that for given unitarily invariant norm $\|\cdot\|_{\mathrm{ui}}$ and the spectral norm $\|\cdot\|_2$, there exist positive constants ξ_{ui} and ξ_2 such that for any V, $\widetilde{V} \in \mathbb{U}^{n\times k}$,*

$$\|H(V) - H(\widetilde{V})\|_{\mathrm{ui}} \le \xi_{\mathrm{ui}}\|\sin\Theta(V,\widetilde{V})\|_{\mathrm{ui}}, \tag{3.1a}$$

$$\|H(V) - H(\widetilde{V})\|_2 \le \xi_2\|\sin\Theta(V,\widetilde{V})\|_2, \tag{3.1b}$$

and also assume that there exists a positive constant δ such that for any $V \in \mathbb{U}^{n\times k}$,

$$\lambda_{k+1}(H(V)) - \lambda_k(H(V)) \ge \delta. \tag{3.2}$$

If $\delta > \xi_{\mathrm{ui}} + \xi_2$, then NEPv (1.1) has a unique solution.

Remark 1. Before we provide a proof, three comments are in order. (a) The conditions in (3.1) are Lipschitz-like conditions. The proof below needs (3.1b) but at the place where (3.1a) is used, it can be simply replaced by using (3.1b) instead. Thus it may seem that the theorem is made unnecessarily more complicated by including both conditions in (3.1) than just (3.1b) alone. But we argue that there are situations where the theorem as is is stronger, namely, if and when $\xi_{\mathrm{ui}} < \xi_2$ for some $\|\cdot\|_{\mathrm{ui}}$

other than the spectral norm and hence the condition on $\delta > \xi_{\rm ui} + \xi_2$ is weaker than $\delta > 2\xi_2$. By the same logic, if $\xi_{\rm ui} \geq \xi_2$, then we should just use the version of this theorem with $\|\cdot\|_{\rm ui}$ being also the spectral norm.

(b) Any one of the conditions in (3.1) yields one for the other by using (2.2). For example, (3.1a) leads to (3.1b) with $\xi_2 = k\xi_{\rm ui}$, and likewise (3.1b) leads to (3.1a) with $\xi_{\rm ui} = n\xi_2$. Conceivably, the resulting ξ-constant is likely worse than obtained through a direct estimation.

(c) The assumption (3.2) requires a uniform gap between the kth and $(k+1)$st eigenvalues of every $H(V)$ for $V \in \mathbb{U}^{n\times k}$. This is known as the "uniform well-posedness" property for using the SCF iteration to solve the Hartree-Fock equation in electronic structure calculations [4]. It is undoubtedly strong and may be hard to verify.

Proof of Theorem 1. We prove the theorem by constructing a mapping on $\mathscr{G}_k(\mathbb{C}^n)$ whose fixed-point is a solution to NEPv (1.1) and vice versa. For any $\mathcal{V} \in \mathscr{G}_k(\mathbb{C}^n)$, let $V \in \mathbb{U}^{n\times k}$ such that $\mathcal{R}(V) = \mathcal{V}$. Because of (3.2), $H(V)$ has a unique invariant subspace associated with its k smallest eigenvalues. We define $\phi(\mathcal{V})$ to be that subspace. Any solution V to NEPv (1.1) satisfies $\mathcal{R}(V) = \phi(\mathcal{R}(V))$, i.e., $\mathcal{R}(V)$ is a fixed-point of ϕ, and vice versa.

In what follows, we will prove ϕ is strictly contractive on $\mathscr{G}_k(\mathbb{C}^n)$ endowed with the distance metric

$$\operatorname{dist}(\mathcal{V}, \widetilde{\mathcal{V}}) = \|\sin\Theta(\mathcal{V}, \widetilde{\mathcal{V}})\|_{\rm ui}.$$

To this end, we consider $\mathcal{V}$, $\widetilde{\mathcal{V}} \in \mathscr{G}_k(\mathbb{C}^n)$, and let V, $\widetilde{V} \in \mathbb{U}^{n\times k}$ such that $\mathcal{R}(V) = \mathcal{V}$ and $\mathcal{R}(\widetilde{V}) = \widetilde{\mathcal{V}}$, respectively. Write the eigen-decompositions of $H(V)$ and $H(\widetilde{V})$ as

$$H(V) = [U, U_{\rm c}]\operatorname{Diag}(\Lambda, \Lambda_{\rm c})[U, U_{\rm c}]^{\rm H} \quad \text{and} \quad H(\widetilde{V}) = [\widetilde{U}, \widetilde{U}_{\rm c}]\operatorname{Diag}(\widetilde{\Lambda}, \widetilde{\Lambda}_{\rm c})[\widetilde{U}, \widetilde{U}_{\rm c}]^{\rm H},$$

where $[U, U_{\rm c}]$, $[\widetilde{U}, \widetilde{U}_{\rm c}] \in \mathbb{C}^{n\times n}$ are unitary, U, $\widetilde{U} \in \mathbb{U}^{n\times k}$, and

$$\begin{aligned}
\Lambda &= \operatorname{Diag}(\lambda_1(H(V)), \ldots, \lambda_k(H(V)), \quad \Lambda_{\rm c} = \operatorname{Diag}(\lambda_{k+1}(H(V)), \ldots, \lambda_n(H(V)),\\
\widetilde{\Lambda} &= \operatorname{Diag}(\lambda_1(H(\widetilde{V})), \ldots, \lambda_k(H(\widetilde{V})), \quad \Lambda_{\rm c} = \operatorname{Diag}(\lambda_{k+1}(H(\widetilde{V})), \ldots, \lambda_n(H(\widetilde{V})).
\end{aligned}$$

By Lemma 3 and the Lipschitz-like condition (3.1), we have

$$|\lambda_j(H(V)) - \lambda_j(H(\widetilde{V}))| \leq \|H(V) - H(\widetilde{V})\|_2 \leq \xi_2\|\sin\Theta(V, \widetilde{V})\|_2 \quad \text{for } 1 \leq j \leq n.$$

It follows from (3.2) and $\delta > \xi_{\rm ui} + \xi_2$ that

$$\begin{aligned}
\lambda_{k+1}(H(V)) - \lambda_k(H(\widetilde{V})) &= \lambda_{k+1}(H(V)) - \lambda_k(H(V)) + \lambda_k(H(V)) - \lambda_k(H(\widetilde{V}))\\
&\geq \delta - \xi_2\|\sin\Theta(V, \widetilde{V})\|_2\\
&\geq \delta - \xi_2 > 0,
\end{aligned} \tag{3.3}$$

since $\|\sin\Theta(V, \widetilde{V})\|_2 \leq 1$ always. Now define $R = H(V)\widetilde{U} - \widetilde{U}\widetilde{\Lambda}$. We have

$$U_{\rm c}^{\rm H} R = \Lambda_{\rm c} U_{\rm c}^{\rm H}\widetilde{U} - U_{\rm c}^{\rm H}\widetilde{U}\widetilde{\Lambda}.$$

On the other hand, it can be seen that $R = [H(V) - H(\widetilde{V})]\widetilde{U}$. Therefore

$$\Lambda_{\rm c} U_{\rm c}^{\rm H}\widetilde{U} - U_{\rm c}^{\rm H}\widetilde{U}\widetilde{\Lambda} = U_{\rm c}^{\rm H}[H(V) - H(\widetilde{V})]\widetilde{U}.$$

Next we apply Lemmas 1 and 4 to get

$$\begin{aligned}
\operatorname{dist}(\phi(\mathcal{V}), \phi(\widetilde{\mathcal{V}})) &= \|\sin\Theta(U, \widetilde{U})\|_{\mathrm{ui}} = \|U_{\mathrm{c}}^{\mathrm{H}}\widetilde{U}\|_{\mathrm{ui}} && (3.4\mathrm{a})\\
&\le \frac{1}{\lambda_{k+1}(H(V)) - \lambda_k(H(\widetilde{V}))}\|U_{\mathrm{c}}^{\mathrm{H}}[H(V) - H(\widetilde{V})]\widetilde{U}\|_{\mathrm{ui}} && (3.4\mathrm{b})\\
&\le \frac{\xi_{\mathrm{ui}}}{\delta - \xi_2}\|\sin\Theta(V, \widetilde{V})\|_{\mathrm{ui}} && (3.4\mathrm{c})\\
&= \frac{\xi_{\mathrm{ui}}}{\delta - \xi_2}\operatorname{dist}(\mathcal{V}, \widetilde{\mathcal{V}}),
\end{aligned}$$

where we have used Lemma 1 for (3.4a), Lemma 4 for (3.4b), and (3.3) for (3.4c). This completes the proof that the mapping ϕ is strictly contractive on $\mathscr{G}_k(\mathbb{C}^n)$ since the factor $\xi_{\mathrm{ui}}/(\delta - \xi_2) < 1$. By Banach fixed-point theorem [10], ϕ has a unique fixed-point in $\mathscr{G}_k(\mathbb{C}^n)$, or equivalently, NEPv (1.1) has a unique solution. □

In section 5, we will verify the satisfiability of the Lipschitz-like conditions (3.1) for two NEPvs arising in electronic structure calculations and linear discriminant analysis.

4 SCF iteration and convergence analysis

4.1 SCF iteration

A natural and most widely used method to solve NEPv (1.1) is the so-called *self-consistent field (SCF) iteration* shown in Algorithm 1, see [14, 19] and references therein for its usage and variants in electronic structure calculations.

Algorithm 1 SCF iteration for solving NEPv (1.1)

Input: $V_0 \in \mathbb{C}^{n\times k}$ with orthonormal columns, i.e., $V_0^{\mathrm{H}}V_0 = I_k$;
Output: a solution to NEPv (1.1).

1: **for** $i = 1, 2, \ldots$ until convergence **do**
2: construct $H_i = H(V_{i-1})$
3: compute the partial eigenvalue decomposition $H_iV_i = V_i\Lambda_i$, where $V_i \in \mathbb{U}^{n\times k}$ and $\Lambda_i = \operatorname{Diag}(\lambda_1(H_i), \ldots, \lambda_k(H_i))$.
4: **end for**
5: **return** the last V_i as a solution to NEPv (1.1).

At each iterative step of SCF, a linear eigenvalue problem for $H_i = H(V_{i-1})$ is partially solved. It is hoped that eventually $\mathcal{R}(V_i)$ converges to some subspace $\mathcal{V}_* \in \mathscr{G}_k(\mathbb{C}^n)$. When it does, the orthonormal basis matrix V_* of $\mathcal{V}_*$ will satisfy NEPv (1.1), provided $H(V)$ is continuous at V_*. Note that at line 2 of Algorithm 1, we use the word "construct" to mean that sometimes H_i may not be explicitly computed but rather exists in some form in such a way that matrix-vector products by H_i can be efficiently performed.

To monitor the progress of convergence, we can compute the normalized residual

$$\mathrm{NRes}_i = \frac{\|H_{i+1}V_i - V_i(V_i^{\mathrm{H}}H_{i+1}V_i)\|}{\|H_{i+1}\| + \|\Lambda_i\|}, \tag{4.1}$$

where $\|\cdot\|$ is some matrix norm that is easy to compute, e.g., the Frobenius norm. But some of the quantities in defining NRes_i may not be necessarily needed in the SCF iteration, e.g., $H_{i+1}V_i$ and $\|H_{i+1}\|$, and they may not be cheap to compute. Therefore we should not compute NRes_i too often, especially for the first many iterations of SCF when convergence hasn't happened yet. Also only a very rough estimate of $\|H_{i+1}\|$ is enough.

There are metrics other than (4.1) that have been used to monitor the convergence of SCF iteration when it comes to a particular application, e.g., the use of $\rho(V_i) = \mathrm{Diag}(V_iV_i^{\mathrm{H}})$ that corresponds to the charge density of electrons in electronic structure calculations (see [3, 4] and references therein). The idea is to examine the difference of $\|\rho(V_i)-\rho(V_{i-1})\|$. When it falls below a prescribed tolerance, convergence is claimed and not yet if otherwise. This criteria often works well, but there is a potential pitfall of premature terminations during a stretch of iterations when $\rho(V_i)$ moves extremely slowly. So it must be used with remedies. One of the remedies is as follows: whenever $\|\rho(V_i)-\rho(V_{i-1})\|$ falls below a prescribed tolerance, check if NRes_i defined by (4.1) is sufficiently tiny. If it is, terminate the SCF iteration; otherwise continue the iteration.

We note that in some applications, such as the one to be discussed in section 5.2, the solutions V of interest to (1.1) are those such that the eigenvalues of $\Lambda = V^{\mathrm{H}}H(V)V$ correspond to the k largest eigenvalues of $H(V)$. We mentioned before but we emphasize here again that we will state all our results explicitly for the case of the smallest eigenvalues as we did and do in the previous and next sections. However, they are equally valid if the word "smallest" is simply replaced by "largest".

4.2 Local convergence of SCF

Let V_* be a solution to NEPv (1.1). The three assumptions we will make are as follows.

(A1) The eigenvalue gap

$$\delta_* = \lambda_{k+1}(H(V_*)) - \lambda_k(H(V_*)) > 0; \tag{4.2}$$

(A2) The matrix-valued function $H(V)$ is continuous at $V = V_*$;

(A3) There exists a nonnegative constant $\chi < \infty$ such that for some $q \geq 1$,

$$\limsup_{\|\sin\Theta(V,V_*)\|_{\mathrm{ui}}\to 0} \frac{\|(I-P_*)[H(V)-H(V_*)]P_*\|_{\mathrm{ui}}}{\|\sin\Theta(V,V_*)\|_{\mathrm{ui}}^q} \leq \chi, \tag{4.3}$$

where $P_* = V_*V_*^{\mathrm{H}}$ is the orthogonal projector onto $\mathcal{R}(V_*)$.

Theorem 2 is our main result on the local convergence of the SCF iteration.

Theorem 2. *Assume* (A1), (A2), (A3), *and* $\chi < \delta_*$ *if* $q = 1$ *(not necessary to assume* $\chi < \delta_*$ *if* $q > 1$*), and let* $\{V_i\}_i$ *be the sequence generated by the SCF iteration (*Algorithm 1*) with initial guess* V_0*. If* V_0 *is sufficiently close to* V_* *in the sense that* $\|\sin\Theta(V_0,V_*)\|_{\mathrm{ui}}$ *is sufficiently small, then there exists a sequence* $\{\tau_i\}_i$ *such that*

$$\|\sin\Theta(V_i,V_*)\|_{\mathrm{ui}} \leq \tau_{i-1}\|\sin\Theta(V_{i-1},V_*)\|_{\mathrm{ui}}^q, \tag{4.4}$$

and

$$\lim_{i\to\infty}\tau_i = \frac{\chi}{\delta_*}, \tag{4.5}$$

and

1. *for $q = 1$, all $\tau_i < 1$ and thus the SCF iteration is locally linearly convergent to $\mathcal{R}(V_*)$ with the linear convergence rate no larger than χ/δ_*;*

2. *for $q > 1$, the SCF iteration is locally convergent to $\mathcal{R}(V_*)$ of order q.*

Proof. For $q = 1$, as $\chi < \delta_*$, we can pick two positive constants $\epsilon_1 < \delta_*/3$ and ϵ_2 such that

$$\tau := \frac{\chi + \epsilon_2}{\delta_* - 3\epsilon_1} < 1; \tag{4.6}$$

otherwise, for $q > 1$, any two positive constants $\epsilon_1 < \delta_*/3$ and ϵ_2 will do. By (A2) and (A3), there exists a positive number Δ with

$$\Delta \le 1 \text{ for } q = 1, \quad \text{or } \Delta < \min\left\{1, \left(\frac{\delta_* - 3\epsilon_1}{\chi + \epsilon_2}\right)^{1/(q-1)}\right\} \text{ for } q > 1 \tag{4.7}$$

such that, whenever $\|\sin\Theta(V, V_*)\|_{\mathrm{ui}} \le \Delta$,

$$\|H(V) - H(V_*)\|_2 \le \epsilon_1, \tag{4.8a}$$

$$\frac{\|(I - P_*)[H(V) - H(V_*)]P_*\|_{\mathrm{ui}}}{\|\sin\Theta(V, V_*)\|_{\mathrm{ui}}^q} \le \chi + \epsilon_2. \tag{4.8b}$$

Now suppose that $\|\sin\Theta(V_0, V_*)\|_{\mathrm{ui}} \le \Delta$, and set

$$\varepsilon_{i1} = \|H(V_i) - H(V_*)\|_2, \tag{4.9a}$$

$$\varepsilon_{i2} = \max\left\{\frac{\|(I - P_*)[H(V_i) - H(V_*)]P_*\|_{\mathrm{ui}}}{\|\sin\Theta(V_i, V_*)\|_{\mathrm{ui}}^q} - \chi, 0\right\}, \tag{4.9b}$$

$$\tau_i = \frac{\chi + \varepsilon_{i2}}{\delta_* - 3\varepsilon_{i1}}. \tag{4.9c}$$

Then $\varepsilon_{01} \le \epsilon_1 < \delta_*/3$, $\varepsilon_{02} \le \epsilon_2$ and hence

$$\tau_0 \Delta^{q-1} \le \tau \Delta^{q-1} < 1. \tag{4.10}$$

To see this, for $q = 1$, $\tau_0 \le \tau < 1$ by (4.6). If $q > 1$, (4.10) is a consequence of (4.7).

In what follows, we will prove that $\|\sin\Theta(V_i, V_*)\|_{\mathrm{ui}} \le \Delta$ and (4.4) hold for all $i \ge 1$. As a consequence, the inequalities in (4.8) hold for $V = V_i$, and $\varepsilon_{i1} \le \epsilon_1 < \delta_*/3$, $\varepsilon_{i2} \le \epsilon_2$ and hence $\tau_i \Delta^{q-1} \le \tau \Delta^{q-1} < 1$. These, in particular, imply that $\|\sin\Theta(V_i, V_*)\|_{\mathrm{ui}} \to 0$ because

$$\|\sin\Theta(V_i, V_*)\|_{\mathrm{ui}} \le \tau \Delta^{q-1} \|\sin\Theta(V_{i-1}, V_*)\|_{\mathrm{ui}},$$

and the limiting equality in (4.4) holds.

Due to similarity, it suffices to show $\|\sin\Theta(V_1, V_*)\|_{\mathrm{ui}} \le \Delta$ and (4.4) hold for $i = 1$. Let the eigen-decompositions of $H(V_0)$ and $H(V_*)$ be

$$H(V_0) = [V_1, V_{1\,\mathrm{c}}]\,\mathrm{Diag}(\Lambda_1, \Lambda_{1\,\mathrm{c}})[V_1, V_{1\,\mathrm{c}}]^{\mathrm{H}},$$
$$H(V_*) = [V_*, V_{*\,\mathrm{c}}]\,\mathrm{Diag}(\Lambda_*, \Lambda_{*\,\mathrm{c}})[V_*, V_{*\,\mathrm{c}}]^{\mathrm{H}},$$

where $[V_1, V_{1\,\mathrm{c}}]$, $[V_*, V_{*\,\mathrm{c}}] \in \mathbb{C}^{n\times n}$ are unitary, V_1, $V_* \in \mathbb{U}^{n\times k}$, and

$$\begin{aligned}
\Lambda_1 &= \mathrm{Diag}(\lambda_1(H(V_0)), \ldots, \lambda_k(H(V_0)), \quad \Lambda_{1\,\mathrm{c}} = \mathrm{Diag}(\lambda_{k+1}(H(V_0)), \ldots, \lambda_n(H(V_0)),\\
\Lambda_* &= \mathrm{Diag}(\lambda_1(H(V_*)), \ldots, \lambda_k(H(V_*)), \quad \Lambda_{*\,\mathrm{c}} = \mathrm{Diag}(\lambda_{k+1}(H(V_*)), \ldots, \lambda_n(H(V_*)).
\end{aligned}$$

By Lemma 3, we know that

$$|\lambda_j(H(V_0)) - \lambda_j(H(V_*))| \le \|H(V_0) - H(V_*)\|_2 = \varepsilon_{01}$$

for $1 \le j \le n$. It follows that

$$\begin{aligned}
\lambda_{k+1}(H(V_0)) - \lambda_k(H(V_*)) &= \lambda_{k+1}(H(V_0)) - \lambda_k(H(V_0)) + \lambda_k(H(V_0)) - \lambda_k(H(V_*))\\
&\ge \delta_* - \varepsilon_{01} > \frac{2\delta_*}{3} > 0. \qquad (4.11)
\end{aligned}$$

Now define $R_1 = H(V_0)V_* - V_*\Lambda_*$. We have

$$(V_{1\,\mathrm{c}})^{\mathrm{H}} R_1 = \Lambda_{1\,\mathrm{c}}(V_{1\,\mathrm{c}})^{\mathrm{H}} V_* - (V_{1\,\mathrm{c}})^{\mathrm{H}} V_* \Lambda_*. \qquad (4.12)$$

On the other hand, it can be seen that $R_1 = [H(V_0) - H(V_*)]V_*$. Therefore

$$\Lambda_{1\,\mathrm{c}}(V_{1\,\mathrm{c}})^{\mathrm{H}} V_* - (V_{1\,\mathrm{c}})^{\mathrm{H}} V_* \Lambda_* = (V_{1\,\mathrm{c}})^{\mathrm{H}}[H(V_0) - H(V_*)]V_*.$$

Next we apply Lemmas 1 and 4, (2.1), and (2.8) to get

$$\begin{aligned}
\|\sin\Theta(V_1, V_*)\|_{\mathrm{ui}} &= \|(V_{1\,\mathrm{c}})^{\mathrm{H}} V_*\|_{\mathrm{ui}}\\
&\le \frac{1}{\lambda_{k+1}(H(V_0)) - \lambda_k(H(V_*))}\|(V_{1\,\mathrm{c}})^{\mathrm{H}}[H(V_0) - H(V_*)]V_*\|_{\mathrm{ui}}\\
&= \frac{1}{\lambda_{k+1}(H(V_0)) - \lambda_k(H(V_*))}\|(I - P_1)[H(V_0) - H(V_*)]P_*\|_{\mathrm{ui}}\\
&\le \frac{1}{\delta_* - \varepsilon_{01}}\|(I - P_1)[H(V_0) - H(V_*)]P_*\|_{\mathrm{ui}}\\
&\le \frac{1}{\delta_* - \varepsilon_{01}}\big(\|(P_1 - P_*)[H(V_0) - H(V_*)]P_*\|_{\mathrm{ui}}\\
&\quad + \|(I - P_*)[H(V_0) - H(V_*)]P_*\|_{\mathrm{ui}}\big)\\
&\le \frac{1}{\delta_* - \varepsilon_{01}}\big(2\|\sin\Theta(V_1, V_*)\|_{\mathrm{ui}}\varepsilon_{01} + (\chi + \varepsilon_{02})\|\sin\Theta(V_0, V_*)\|_{\mathrm{ui}}^q\big).
\end{aligned}$$

Solving the above inequality for $\|\sin\Theta(V_1, V_*)\|_{\mathrm{ui}}$, we obtain

$$\|\sin\Theta(V_1, V_*)\|_{\mathrm{ui}} \le \tau_0\|\sin\Theta(V_0, V_*)\|_{\mathrm{ui}}^q \le \tau\Delta^{q-1}\|\sin\Theta(V_0, V_*)\|_{\mathrm{ui}}, \qquad (4.13)$$

where we have used $\|\sin\Theta(V_0, V_*)\|_{\mathrm{ui}} \le \Delta$ for the second inequality. The first inequality in (4.13) is (4.4) for $i = 1$ and the second inequality there implies $\|\sin\Theta(V_1, V_*)\|_{\mathrm{ui}} \le \Delta \le 1$ because $\tau\Delta^{q-1} < 1$. □

Remark 2. (a) The assumption (A1) is similar to the "uniform well-posedness" assumption (3.2) in Theorem 1 for the eigenvalue gap of $H(V)$.

(b) Let $[V_*,\, V_{*\,\mathrm{c}}] \in \mathbb{C}^{n\times n}$ be unitary. Notice that

$$\|(I-P_*)[H(V)-H(V_*)]P_*\|_{\mathrm{ui}} = \|V_{*\,\mathrm{c}}^{\mathrm{H}}[H(V)-H(V_*)]V_*\|_{\mathrm{ui}}.$$

The assumption (A3) is on the closeness of the $(2,1)$-block of $[V_*, V_{*\,\mathrm{c}}]^{\mathrm{H}} H(V)[V_*, V_{*\,\mathrm{c}}]$ in limit to the $(2,1)$-block, which is 0, of $[V_*,\, V_{*\,\mathrm{c}}]^{\mathrm{H}} H(V_*)[V_*,\, V_{*\,\mathrm{c}}]$.

(c) The assumption (A3) with $q=1$ is weaker than the Lipschitz-like condition (3.1a). This is because (3.1a) implies

$$\begin{aligned}\limsup_{\|\sin\Theta(V,V_*)\|_{\mathrm{ui}}\to 0} & \frac{\|(I-P_*)[H(V)-H(V_*)]P_*\|_{\mathrm{ui}}}{\|\sin\Theta(V,V_*)\|_{\mathrm{ui}}} \\ & \le \limsup_{\|\sin\Theta(V,V_*)\|_{\mathrm{ui}}\to 0} \frac{\|H(V)-H(V_*)\|_{\mathrm{ui}}}{\|\sin\Theta(V,V_*)\|_{\mathrm{ui}}} \le \xi_{\mathrm{ui}}.\end{aligned}$$

In other words, the Lipschitz-like condition (3.1a) implies the assumption (A3) with $q=1$ and $\chi=\xi_{\mathrm{ui}}$.

(d) From the proof of Theorem 2, we can see that the assumption (A3) can be relaxed to

$$\limsup_{i\to\infty} \frac{\|(I-P_*)[H(V_i)-H(V_*)]P_*\|_{\mathrm{ui}}}{\|\sin\Theta(V_i,V_*)\|_{\mathrm{ui}}} \le \chi, \tag{4.14}$$

instead for all $V\in\mathbb{U}^{n\times k}$ that go to V_*.

Example 1. We give an example to show the local convergence rate revealed in Theorem 2 is nearly achievable, which implies its near-optimality. Consider the following single-particle Hamiltonian in electronic structure calculations studied in [13, 25, 30]:

$$H(V) = L + \alpha\cdot \mathrm{Diag}(L^{-1}\rho(V)), \tag{4.15}$$

where $L=\mathrm{tridiag}(-1,2,-1)$ is a discrete 1-D Laplacian, α is some constant, $\rho(V)=\mathrm{diag}(VV^{\mathrm{T}})$. Here all numbers are real, and $V^{\mathrm{T}}V=I_k$. This is a good place for us to point out again that all our developments in this paper are valid for real NEPv (1.1), i.e., $H(V)$ is an $n\times n$ real symmetric matrix-valued function of real V with $V^{\mathrm{T}}V=I_k$, and at the same time Q in (1.2) is restricted to any orthogonal matrix.

To numerically demonstrate the local convergence rate, we use the following approach to compute an estimated theoretical convergence rate and the corresponding observed convergence rate for solving NEPv (1.1) with $H(V)$ here by the SCF iteration (Algorithm 1). For a given α, we compute an "exact" solution $\widehat{V}_*$ by setting a small tolerance of 10^{-14} in SCF. At convergence, the eigenvalue gap δ_* of the assumption (A1) is estimated by

$$\widehat{\delta}_* = \lambda_{k+1}(H(\widehat{V}_*)) - \lambda_k(H(\widehat{V}_*)).$$

We approximate the quantity χ in the assumption (A3) by using the quantity

$$\frac{\|(I-\widehat{P}_*)[H(V_i)-H(\widehat{V}_*)]\widehat{P}_*\|_2}{\|\sin\Theta(V_i,\widehat{V}_*)\|_2}$$

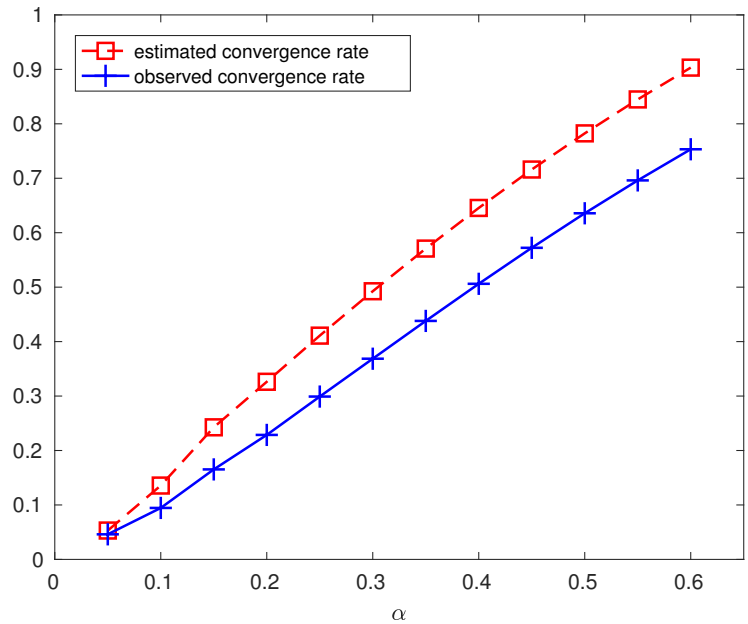

Figure 4.1: Convergence rate: estimated $\widehat{\chi}/\widehat{\delta}_*$ and observed $\widehat{\tau}$ for different values of α for the single-particle Hamiltonian $H(V)$ defined in (4.15) with $n = 10$ and $k = 2$.

near the end of the SCF iteration when it stays almost unchanged at a constant $\widehat{\chi}$, where $\widehat{P}_* = \widehat{V}_*\widehat{V}_*^{\mathrm{T}}$. Consequently, an estimated theoretical convergence rate in Theorem 2 is the quantity $\widehat{\chi}/\widehat{\delta}_*$. The corresponding observed convergence rate $\widehat{\tau}$ is the *numerical* limit of the sequence

$$\widehat{\tau}_i = \frac{\|\sin\Theta(V_i, \widehat{V}_*)\|_2}{\|\sin\Theta(V_{i-1}, \widehat{V}_*)\|_2}.$$

Figure 4.1 shows the estimated convergence rates $\widehat{\chi}/\widehat{\delta}_*$ and observed convergence rates $\widehat{\tau}$ for different values of α. We can see that the estimated convergence rates are *tight* upper bounds for the observed convergence rates for all tested values of α. In particular, for $\alpha = 0.05$, the bound is essentially reached.

4.3 Global convergence of SCF

Our main results for the global convergence of the SCF iteration are based on the following two inequalities to be established:

$$\|\sin\Theta(V_i, V_{i+1})\|_2 \le \tau_2 \|\sin\Theta(V_{i-1}, V_i)\|_2, \tag{4.16}$$

$$\|\sin\Theta(V_i, V_{i+1})\|_{\mathrm{ui}} \le \tau_{\mathrm{ui}} \|\sin\Theta(V_{i-1}, V_i)\|_{\mathrm{ui}}, \tag{4.17}$$

for some constants τ_2, $\tau_{\mathrm{ui}} < 1$ to be specified in the theorem below.

Theorem 3. *Assume the Lipschitz-like condition* (3.1) *hold and* ξ_{ui} *and* ξ_2 *are the corresponding positive constants, and let* $\{V_i\}_i$ *be generated by the SCF iteration (*Algorithm 1*). Suppose that there*

exists a positive constant δ such that

$$\lambda_{k+1}(H(V_i)) - \lambda_k(H(V_i)) \geq \delta \quad \textit{for all } i = 1, 2, \ldots. \tag{4.18}$$

(a) If $\delta > \xi_{\mathrm{ui}} + \xi_2$, then the inequality (4.17) *holds for all i with*

$$\tau_{\mathrm{ui}} = \frac{\xi_{\mathrm{ui}}}{\delta - \xi_2} < 1. \tag{4.19}$$

(b) If $\delta > \xi_2 + \|\sin\Theta(V_0, V_1)\|_2 \xi_2$, then the inequality (4.16) *holds for all i with*

$$\tau_2 = \frac{\xi_2}{\delta - \xi_2 \|\sin\Theta(V_0, V_1)\|_2} < 1. \tag{4.20}$$

(c) If $\delta > \max\{\xi_{\mathrm{ui}}, \xi_2\} + \|\sin\Theta(V_0, V_1)\|_2 \xi_2$, then both inequalities (4.16) *and* (4.17) *hold for all i with*

$$\tau_2 = \frac{\xi_2}{\delta - \xi_2 \|\sin\Theta(V_0, V_1)\|_2} < 1 \quad \textit{and} \quad \tau_{\mathrm{ui}} = \frac{\xi_{\mathrm{ui}}}{\delta - \xi_2 \|\sin\Theta(V_0, V_1)\|_2} < 1. \tag{4.21}$$

These inequalities imply, in their respective cases, $\mathcal{R}(V_i)$ converges and the limit, denoted by $\mathcal{V}_* \in \mathscr{G}_k(\mathbb{C}^n)$, is the solution of NEPv (1.1)*. In other words, $\sin\Theta(V_{i-1}, V_i) \to 0$ as $i \to \infty$ and the SCF iteration is globally linearly convergent.*

Proof. Let the eigen-decomposition of $H(V_i)$ be

$$H(V_i) = [V_{i+1},\, V_{i+1,\mathrm{c}}]\, \mathrm{Diag}(\Lambda_{i+1}, \Lambda_{i+1,\mathrm{c}})[V_{i+1},\, V_{i+1,\mathrm{c}}]^{\mathrm{H}},$$

where $[V_{i+1},\, V_{i+1,\mathrm{c}}] \in \mathbb{C}^{n\times n}$ is unitary, $V_{i+1} \in \mathbb{U}^{n\times k}$,

$$\begin{aligned} \Lambda_{i+1} &= \mathrm{Diag}(\lambda_1(H(V_i)), \ldots, \lambda_k(H(V_i)), \\ \Lambda_{i+1,\mathrm{c}} &= \mathrm{Diag}(\lambda_{k+1}(H(V_i)), \ldots, \lambda_n(H(V_i)). \end{aligned}$$

For convenience, introduce $\eta_i = \|\sin\Theta(V_{i-1}, V_i)\|_2$. By Lemma 3 and (3.1), it holds that

$$|\lambda_j(H(V_{i-1})) - \lambda_j(H(V_i))| \leq \|H(V_{i-1}) - H(V_i)\|_2 \leq \xi_2 \eta_i \tag{4.22}$$

for all j. Combine (4.18) and (4.22) to get

$$\begin{aligned} \lambda_{k+1}(H(V_{i-1})) - \lambda_k(H(V_i)) &= \lambda_{k+1}(H(V_{i-1})) - \lambda_k(H(V_{i-1})) \\ &\quad + \lambda_k(H(V_{i-1})) - \lambda_k(H(V_i)) \\ &\geq \delta - \xi_2 \eta_i. \end{aligned} \tag{4.23}$$

Define $R_i = H(V_{i-1})V_{i+1} - V_{i+1}\Lambda_{i+1}$. We have

$$V_{i\,\mathrm{c}}^{\mathrm{H}} R_i = \Lambda_{i\,\mathrm{c}} V_{i\,\mathrm{c}}^{\mathrm{H}} V_{i+1} - V_{i\,\mathrm{c}}^{\mathrm{H}} V_{i+1} \Lambda_{i+1}.$$

On the other hand, it can be verified that $R_i = [H(V_{i-1}) - H(V_i)]V_{i+1}$. Therefore

$$\Lambda_{i\,\mathrm{c}} V_{i\,\mathrm{c}}^{\mathrm{H}} V_{i+1} - V_{i\,\mathrm{c}}^{\mathrm{H}} V_{i+1} \Lambda_{i+1} = V_{i\,\mathrm{c}}^{\mathrm{H}} [H(V_{i-1}) - H(V_i)] V_{i+1}. \tag{4.24}$$

Apply Lemmas 1 and 4 to get, provided that we can prove $\delta - \xi_2\eta_i > 0$,

$$\begin{aligned}
\|\sin\Theta(V_i, V_{i+1})\|_{\mathrm{ui}} &= \|V_{i\,\mathrm{c}}^{\mathrm{H}} V_{i+1}\|_{\mathrm{ui}} && (4.25\mathrm{a})\\
&\le \frac{1}{\lambda_{k+1}(H(V_{i-1})) - \lambda_k(H(V_i))}\|V_{i\,\mathrm{c}}^{\mathrm{H}}[H(V_{i-1}) - H(V_i)]V_{i+1}\|_{\mathrm{ui}} && (4.25\mathrm{b})\\
&\le \frac{1}{\delta - \xi_2\eta_i}\|H(V_{i-1}) - H(V_i)\|_{\mathrm{ui}}\\
&\le \frac{\xi_{\mathrm{ui}}}{\delta - \xi_2\eta_i}\|\sin\Theta(V_{i-1}, V_i)\|_{\mathrm{ui}}, && (4.25\mathrm{c})
\end{aligned}$$

where we have used Lemma 1 for (4.25a), and Lemma 4 for (4.25b).

Item (a) is an immediate consequence of (4.25c) because all

$$\eta_i = \|\sin\Theta(V_{i-1}, V_i)\|_2 \le 1$$

and thus $\delta - \xi_2\eta_i \ge \delta - \xi_2 > 0$ by assumption.

For item (b), specialize (4.25c) to the case $\|\cdot\|_{\mathrm{ui}} = \|\cdot\|_2$ to get

$$\eta_{i+1} = \|\sin\Theta(V_i, V_{i+1})\|_2 \le \frac{\xi_2}{\delta - \xi_2\eta_i}\|\sin\Theta(V_{i-1}, V_i)\|_2 = \frac{\xi_2}{\delta - \xi_2\eta_i}\eta_i. \tag{4.26}$$

By assumption $\delta > \xi_2 + \xi_2\eta_1$, $\delta - \xi_2\eta_1 > \xi_2 > 0$ and thus $\eta_2 \le \xi_2\eta_1/(\delta - \xi_2\eta_1) < \eta_1$. Now assume $\eta_{i+1} < \eta_i$ for all $i \le \ell - 1$ ($\ell \ge 2$). Using (4.26), we get

$$\eta_{\ell+1} \le \frac{\xi_2\eta_\ell}{\delta - \xi_2\eta_\ell} < \frac{\xi_2\eta_\ell}{\delta - \xi_2\eta_1} < \eta_\ell.$$

Thus, by mathematical induction, we conclude that $\eta_{i+1} < \eta_i$ for all $i \ge 1$. Finally, by (4.26), we obtain $\delta - \xi_2\eta_i \ge \delta - \xi_2\eta_1 > 0$ and

$$\eta_{i+1} \le \frac{\xi_2\eta_i}{\delta - \xi_2\eta_i} < \frac{\xi_2}{\delta - \xi_2\eta_1}\eta_i = \tau_2\eta_i,$$

where τ_2 is given by (4.20).

Finally for item (c), the assumption on δ is stronger than the one in item (b). Hence (4.16) holds for all i with τ_2 given by (4.20) which is the same as the one in (4.21). Now we still have (4.25c) and, by what we have shown for item (b), $\delta - \xi_2\eta_i \ge \delta - \xi_2 > 0$. The inequality (4.17) with τ_{ui} given in (4.21) is implied by (4.25c). □

Theorem 3 looks similar to Theorem 1 on the existence and uniqueness of the solution of NEPv (1.1). Both use the Lipschitz-like conditions in (3.1), but differ on gap assumptions between the kth and $(k+1)$st eigenvalues. Theorem 3 only requires the uniform gap assumption (4.18) with three different assumptions on the size of the gap δ on $H(V_i)$ for all V_i generated by the SCF iteration, whereas the gap assumption in Theorem 1 is for all $V \in \mathbb{U}^{n\times k}$. This seems weaker, but it is not clear that if (4.18) is any easier to use than (3.2). Depending on how large $\|\sin\Theta(V_0, V_1)\|_2$ is, $\delta > \xi_{\mathrm{ui}} + \xi_2$ needed for item (a) can be a significantly stronger assumption than the ones for items (b) and (c). Another difference is that under the conditions of Theorem 1, NEPv (1.1) has a unique solution, whereas for Theorem 3, it only guarantees that NEPv (1.1) has a solution which is the limit of $\mathcal{R}(V_i)$.

5 Applications

In this section, we apply our previous convergence analysis to the discretized Kohn-Sham NEPv in electronic structure calculations and the NEPv arising from linear discriminant analysis for dimension reduction. When applicable, we compare with the existing results. We note that both examples take the form (1.1) but for real numbers, i.e., $H(V) \in \mathbb{R}^{n\times n}$ are symmetric and $V \in \mathbb{R}^{n\times k}$ has orthonormal columns. As we commented before, our general theory so far remains valid after replacing all $\mathbb{C}$ by $\mathbb{R}$ and $\mathbb{U}^{n\times k}$ by $\mathbb{O}^{n\times k} := \{V \mid V \in \mathbb{R}^{n\times k}, V^{\mathrm{T}}V = I_k\}$.

5.1 The discretized Kohn-Sham NEPv

Consider the following discretized Kohn-Sham NEPv studied in [16, 26, 13] and references therein:

$$H(V)V = V\Lambda, \tag{5.1a}$$

where the matrix-valued function

$$H(V) = \frac{1}{2}L + V_{\mathrm{ion}} + \sum_{\ell} w_\ell w_\ell^{\mathrm{T}} + \mathrm{Diag}(L^\dagger \rho) + \mathrm{Diag}(\mu_{\mathrm{xc}}^{\mathrm{T}}(\rho)\mathbf{1}) \tag{5.1b}$$

is the plane-wave discretized Hamiltonian of the total energy functional. The first term corresponds to the kinetic energy and L is a finite dimensional representation of the Laplacian operator. The second term V_{ion} is for the ionic pseudopotential sampled on the suitably chosen Cartesian grid in the local ionic potential energy. The third term represents a discretized pseudopotential reference projection function in the nonlocal ionic potential energy. The fourth term is for the Hartree potential energy, where $\rho \equiv \rho(V) := \mathrm{Diag}(VV^{\mathrm{T}}) \in \mathbb{R}^n$ and $L^\dagger$ is the pseudoinverse of L. The last term is for the exchange correlation energy, where $\mu_{\mathrm{xc}}(\rho) = \frac{\partial \epsilon_{\mathrm{xc}}(\rho)}{\partial \rho} \in \mathbb{R}^{n\times n}$, $\epsilon_{\mathrm{xc}}(\rho)$ is an exchange correlation functional and $\mathbf{1}$ is a vector of all ones.

The discretized Kohn-Sham NEPv (5.1) is of the NEPv form (1.1). To apply the results in the previous sections, we first have to estimate how $H(V)$ changes with respect to V. For this purpose, it suffices to know how $\mu_{\mathrm{xc}}(\rho)$ changes with respect to $\rho \equiv \rho(V)$ since the first three terms in $H(V)$ are independent of V. We assume that there exist positive constants σ_2 and σ_{F} such that

$$\|\,\mathrm{Diag}(\mu_{\mathrm{xc}}(\rho)^{\mathrm{T}}\mathbf{1}) - \mathrm{Diag}(\mu_{\mathrm{xc}}(\tilde\rho)^{\mathrm{T}}\mathbf{1})\|_2 = \|[\mu_{\mathrm{xc}}(\rho) - \mu_{\mathrm{xc}}(\tilde\rho)]^{\mathrm{T}}\mathbf{1}\|_\infty \le \sigma_2\|\rho - \tilde\rho\|_\infty \tag{5.2a}$$

$$\|\,\mathrm{Diag}(\mu_{\mathrm{xc}}(\rho)^{\mathrm{T}}\mathbf{1}) - \mathrm{Diag}(\mu_{\mathrm{xc}}(\tilde\rho)^{\mathrm{T}}\mathbf{1})\|_{\mathrm{F}} = \|[\mu_{\mathrm{xc}}(\rho) - \mu_{\mathrm{xc}}(\tilde\rho)]^{\mathrm{T}}\mathbf{1}\|_2 \le \sigma_{\mathrm{F}}\|\rho - \tilde\rho\|_2 \tag{5.2b}$$

for all $\rho \equiv \rho(V)$ and $\tilde\rho \equiv \rho(\widetilde{V})$, where $\|\cdot\|_\infty$ is either the ℓ_∞ of a vector or the ℓ_∞-operator norm of a matrix. With these assumptions (5.2), we can verify that $H(V)$ satisfy the Lipschitz-like conditions (3.1):

$$\begin{aligned}
\|H(V) - H(\widetilde{V})\|_2 &\le \|\,\mathrm{Diag}(L^\dagger(\rho - \tilde\rho))\|_2 + \|\,\mathrm{Diag}(\mu_{\mathrm{xc}}^{\mathrm{T}}(\rho)\mathbf{1} - \mu_{\mathrm{xc}}^{\mathrm{T}}(\tilde\rho)\mathbf{1})\|_2 \\
&= \|L^\dagger(\rho - \tilde\rho)\|_\infty + \|[\mu_{\mathrm{xc}}(\rho) - \mu_{\mathrm{xc}}(\tilde\rho)]^{\mathrm{T}}\mathbf{1})\|_\infty \\
&\le \|L^\dagger\|_\infty\|\rho - \tilde\rho\|_\infty + \sigma_2\|\rho - \tilde\rho\|_\infty \\
&= (\|L^\dagger\|_\infty + \sigma_2)\max_i |e_i^{\mathrm{T}}(VV^{\mathrm{T}} - \widetilde{V}\widetilde{V}^{\mathrm{T}})e_i| \\
&\le (\|L^\dagger\|_\infty + \sigma_2)\|\sin\Theta(V, \widetilde{V})\|_2 \\
&\equiv \xi_2^{\mathrm{ks}}\|\sin\Theta(V, \widetilde{V})\|_2
\end{aligned} \tag{5.3a}$$

and

$$\begin{aligned}\|H(V)-H(\widetilde{V})\|_{\mathrm{F}} &\le \|\operatorname{Diag}(L^{\dagger}(\rho-\tilde{\rho}))\|_{\mathrm{F}} + \|\operatorname{Diag}(\mu_{\mathrm{xc}}^{\mathrm{T}}(\rho)\mathbf{1}-\mu_{\mathrm{xc}}^{\mathrm{T}}(\tilde{\rho})\mathbf{1})\|_{\mathrm{F}}\\ &= \|L^{\dagger}(\rho-\tilde{\rho})\|_2 + \|[\mu_{\mathrm{xc}}(\rho)-\mu_{\mathrm{xc}}(\tilde{\rho})]^{\mathrm{T}}\mathbf{1})\|_2\\ &\le \|L^{\dagger}\|_2\|\rho-\tilde{\rho}\|_2 + \sigma_{\mathrm{F}}\|\rho-\tilde{\rho}\|_2\\ &\le (\|L^{\dagger}\|_2+\sigma_{\mathrm{F}})\|VV^{\mathrm{T}}-\widetilde{V}\widetilde{V}^{\mathrm{T}}\|_{\mathrm{F}}\\ &\equiv \xi_{\mathrm{F}}^{\mathrm{ks}}\|\sin\Theta(V,\widetilde{V})\|_{\mathrm{F}}, \end{aligned} \tag{5.3b}$$

where $\xi_2^{\mathrm{ks}} = \|L^{\dagger}\|_{\infty}+\sigma_2$ and $\xi_{\mathrm{F}}^{\mathrm{ks}} = \sqrt{2}(\|L^{\dagger}\|_2+\sigma_{\mathrm{F}})$. By Theorem 1, we have the following theorem on the existence and uniqueness of (5.1).

Theorem 4. *Under the assumption* (5.2)*, if for any* $V\in\mathbb{O}^{n\times k}$

$$\lambda_{k+1}(H(V))-\lambda_k(H(V)) > \min\{\xi_2^{\mathrm{ks}}+\xi_{\mathrm{F}}^{\mathrm{ks}}, 2\xi_2^{\mathrm{ks}}\},$$

then the discretized Kohn-Sham NEPv (5.1) *has a unique solution.*

Next we consider the convergence of the SCF iteration for solving the NEPv (5.1). For applying the local and global convergence results of Theorems 2 and 3, we note the assumption (A3) in (4.3) becomes

$$\limsup_{\|\sin\Theta(V,V_*)\|_{\mathrm{ui}}\to 0} \frac{\|(I-P_*)[\operatorname{Diag}(L^{\dagger}[\rho-\rho_*]) + \operatorname{Diag}([\mu_{\mathrm{xc}}(\rho)-\mu_{\mathrm{xc}}(\rho_*)]^{\mathrm{T}}\mathbf{1})]P_*\|_{\mathrm{ui}}}{\|\sin\Theta(V,V_*)\|_{\mathrm{ui}}^{q}} \le \chi, \tag{5.4}$$

where $\rho_* := \rho(V_*)$. Evidently, by Lipschitz-like conditions (5.3), we can take $\chi = \xi_{\mathrm{ui}}^{\mathrm{ks}}$ for $\mathrm{ui}\in\{2,\mathrm{F}\}$. The following theorem summarizes the local and global convergence of the SCF iteration.

Theorem 5. *Let* V_* *be a solution of the discretized Kohn-Sham NEPv* (5.1)*, and let* $\{V_i\}_i$ *be the sequence generated by the SCF iteration (*Algorithm 1*). Then under the assumption* (5.2)*,*

(i) If for $\mathrm{ui}\in\{2,\mathrm{F}\}$ *and* $\delta_* := \lambda_{k+1}(H(V_*))-\lambda_k(H(V_*)) > \xi_{\mathrm{ui}}^{\mathrm{ks}}$ *and* $\mathcal{R}(V_0)$ *is sufficiently close* $\mathcal{R}(V_*)$*, then* $\mathcal{R}(V_i)$ *linearly converges to* $\mathcal{R}(V_*)$*. Moreover,*

$$\|\sin\Theta(V_{i+1},V_*)\|_{\mathrm{ui}} \le \tau_i^{\mathrm{ks}}\|\sin\Theta(V_{i+1},V_*)\|_{\mathrm{ui}}, \tag{5.5}$$

where $\tau_i^{\mathrm{ks}} < 1$ *and* $\lim_{i\to\infty}\tau_i^{\mathrm{ks}} = \frac{\xi_{\mathrm{ui}}^{\mathrm{ks}}}{\delta_*}$.

(ii) Theorem 3 *is valid for* $\mathrm{ui}=\mathrm{F}$ *and with* ξ_2 *and* ξ_{ui} *replaced by* ξ_2^{ks} *and* $\xi_{\mathrm{ui}}^{\mathrm{ks}}$*, respectively.*

Proof. Result (i) follows from Theorem 2 since the subspace approximation condition (4.3) holds for the constants $\chi = \xi_{\mathrm{ui}}^{\mathrm{ks}}$. Result (ii) immediately follows from Theorem 3. □

Let us compare Theorem 5 with the previous convergence results of the SCF iteration on NEPv (5.1) obtained by Liu, Wang, Wen and Yuan [13]. We first restate the following main results of [13].

Theorem 6 ([13, Theorem 4.2])**.** *For NEPv* (5.1)*, suppose that there exists a constant* σ *such that for all* $\rho,\tilde{\rho}\in\mathbb{R}^n$*,*

$$\|\operatorname{Diag}(\mu_{\mathrm{xc}}(\rho)^{\mathrm{T}}\mathbf{1}) - \operatorname{Diag}(\mu_{\mathrm{xc}}(\tilde{\rho})^{\mathrm{T}}\mathbf{1})\|_{\mathrm{F}} \le \sigma\|\rho-\tilde{\rho}\|_2, \tag{5.6a}$$

$$\left\|\frac{\partial^2\epsilon_{\mathrm{xc}}}{\partial\rho^2}\mathbf{1}\right\|_2 \le \sigma. \tag{5.6b}$$

*Let $\{V_i\}_i$ be the sequence generated by the SCF iteration (*Algorithm 1*), V_* be a solution of NEPv* (5.1)*, and $\delta_* = \lambda_{k+1}(H(V_*)) - \lambda_k(H(V_*)) > 0$. If V_i is sufficiently close to V_*, i.e., $\|\sin\Theta(V_i, V_*)\|_2$ is sufficiently small, then*

$$\|\sin\Theta(V_{i+1}, V_*)\|_2 \le \frac{2\sqrt{n}(\|L^\dagger\|_2 + \sigma)}{\delta_*}\|\sin\Theta(V_i, V_*)\|_2 + \mathcal{O}(\|\sin\Theta(V_i, V_*)\|_2^2). \tag{5.7}$$

Theorem 7 ([13, Theorem 3.3])**.** *Assume* (5.6) *and suppose there exists a constant $\delta > 0$ such that $\lambda_{k+1}(H(V_i)) - \lambda_k(H(V_i)) \ge \delta > 0$ for all i, where $\{V_i\}_i$ is the sequence generated by the SCF iteration (*Algorithm 1*). If $\delta > 12k\sqrt{n}(\|L^\dagger\|_2 + \sigma)$, $\mathcal{R}(V_i)$ converges to a solution $\mathcal{R}(V_*)$ of NEPv* (5.1)*.*

First of all, we note that the assumption (5.6b) on the twice differentiability of the exchange correlation functional $\epsilon_{\rm xc}$ is not necessary for the new Theorem 5. On the local convergence, Theorem 6 requires the eigenvalue gap $\delta_* > 2\sqrt{n}(\|L^\dagger\|_2 + \sigma)$. In contrast, for $\mathrm{ui} = 2$, Theorem 5(i) only requires $\delta_* > \|L^\dagger\|_\infty + \sigma_2$, which is a much weaker condition. This can be verified as follows. By the assumption (5.2) of Theorem 5(i), let

$$\widehat{\sigma}_2 = \sup_{\rho\neq\widetilde{\rho}} \frac{\|[\mu_{\rm xc}(\rho) - \mu_{\rm xc}(\tilde{\rho})]^{\rm T}\mathbf{1}\|_\infty}{\|\rho - \tilde{\rho}\|_\infty}, \quad \text{and} \quad \widehat{\sigma}_{\rm F} = \sup_{\rho\neq\widetilde{\rho}} \frac{\|[\mu_{\rm xc}(\rho) - \mu_{\rm xc}(\tilde{\rho})]^{\rm T}\mathbf{1}\|_2}{\|\rho - \tilde{\rho}\|_2}.$$

Then $\frac{1}{\sqrt{n}}\widehat{\sigma}_{\rm F} \le \widehat{\sigma}_2 \le \sqrt{n}\widehat{\sigma}_{\rm F}$ and $\widehat{\sigma}_{\rm F} \le \sigma$. Since Theorem 5 also holds for $\sigma_2 = \widehat{\sigma}_2$ and $\sigma_{\rm F} = \widehat{\sigma}_{\rm F}$, we have

$$\|L^\dagger\|_\infty + \widehat{\sigma}_2 \le \sqrt{n}(\|L^\dagger\|_2 + \widehat{\sigma}_{\rm F}) < 2\sqrt{n}(\|L^\dagger\|_2 + \sigma). \tag{5.8}$$

Therefore, Theorem 5(i) has a weaker condition on the eigenvalue gap δ_* than the one required by Theorem 6. In fact, since the first inequality in (5.8) is overestimated, Theorem 5(i) has significantly weaker condition on the eigenvalue gap by removing the factor $\sqrt{n}$. By the inequalities (5.8), we can also see that the new bound (5.5) on $\|\sin\Theta(V_{i+1}, V_*)\|_2$ is much sharper than the first-order bound (5.7) of Theorem 6. In addition, we note that for $\mathrm{ui} = \mathrm{F}$, Theorem 5(i) provides the convergence rate $\xi_{\rm F}^{\rm ks}/\delta_*$ of the SCF iteration, which is absent in [13].

On the global convergence, the condition $\delta > 12k\sqrt{n}(\|L^\dagger\|_2 + \sigma)$ in Theorem 7 is a much more stringent condition than the one required by our Theorem 5(ii). This is due to the fact that

$$\xi_{\rm ui}^{\rm ks} + \xi_2^{\rm ks} \le (\sqrt{n} + 1)(\|L^\dagger\|_2 + \widehat{\sigma}_{\rm F}) < 12k\sqrt{n}(\|L^\dagger\|_2 + \sigma). \tag{5.9}$$

Now let us examine the implications of these results for the simple single-particle Hamiltonian (4.15) with the nonlinearity controlled by the parameter α. To ensure the local convergence of the SCF iteration, the sufficient condition from the analysis in [13] is that the parameter α must satisfy

$$\alpha < \alpha_L := \frac{\delta_*}{2\sqrt{n}\|L^\dagger\|_2}. \tag{5.10}$$

In constrast, by new Theorem 5(ii), the upper bound is

$$\alpha < \widetilde{\alpha}_L := \max\left\{\frac{\delta_*}{\|L^\dagger\|_1}, \frac{\delta_*}{\sqrt{2}\|L^\dagger\|_2}\right\}. \tag{5.11}$$

Since[1] $\|L^\dagger\|_1 \le \sqrt{n}\|L^\dagger\|_2$, $\widetilde{\alpha}_L$ is always larger than α_L does not explicitly depend on n, it implies that our sufficient condition (5.11) is less stringent than (5.10). This is also confirmed by numerical results seen Example 1 and [25, Table 1].

[1]Numerical observation suggests that $\|L^\dagger\|_1/\|L^\dagger\|_2 \le 1.7072$. We do not have a rigourous proof.

For the global convergence, an earlier one of Yang, Gao and Meza [25] requires

$$\alpha < \alpha_{\mathrm{F}} := \frac{\delta}{\ln \frac{1-\gamma}{\gamma} \cdot n^4 \|L^\dagger\|_1}, \tag{5.12}$$

where γ is a constant and $\gamma \ll 1$, and δ is the one as in Theorems 3 and 7. Liu, Wang, Wen and Yuan [13] improved the upper bound (5.12) to

$$\alpha < \alpha_G := \frac{\delta}{12k\sqrt{n}\|L^\dagger\|_2}. \tag{5.13}$$

In contrast, the result in Theorem 5 (ii) requires

$$\alpha < \widetilde{\alpha}_G := \max\left\{ \frac{\delta}{(1 + \|\sin\Theta(V_0, V_1)\|_2)\|L^\dagger\|_1}, \frac{\delta}{\|L^\dagger\|_1 + \sqrt{2}\|L^\dagger\|_2} \right\}. \tag{5.14}$$

As we can see, unlike the previous bounds α_{F} and α_G, $\widetilde{\alpha}_G$ does not explicitly depend on n. Furthermore, $\widetilde{\alpha}_G$ is always larger than α_G, which in turn is larger than α_{F} for

$$\gamma < \left[1 + \exp\left(\frac{12k}{n^{7/2}} \cdot \frac{\|L^\dagger\|_2}{\|L^\dagger\|_1}\right)\right]^{-1},$$

i.e, $\widetilde{\alpha}_G > \alpha_G > \alpha_{\mathrm{F}}$. This means our result (5.14) predicts a much larger range of α than (5.12) and (5.13) could, within which the SCF iteration converges. This is again confirmed by numerical experiments reported in Example 1 and [25, Table 1].

5.2 The trace ratio problem

In this section, we discuss an application to a trace ratio maximization problem (TRP) arising from the linear discriminant analysis for dimension reduction [15, 28, 29]. Given symmetric matrices $A, B \in \mathbb{R}^{n\times n}$ and $B \succ 0$ (positive definite), TRP solves the following optimization problem

$$\max_{V \in \mathbb{O}^{n\times k}} \frac{\mathrm{tr}(V^{\mathrm{T}}AV)}{\mathrm{tr}(V^{\mathrm{T}}BV)}, \tag{5.15}$$

where $\mathrm{tr}(\,\cdot\,)$ denotes the trace of a square matrix. Employing the first-order optimality condition (i.e., the KKT condition) yields that any critical point $V \in \mathbb{O}^{n\times k}$ of (5.15) is a solution of the following NEPv:

$$H(V)V = V\Lambda, \tag{5.16a}$$

where

$$H(V) = A - \psi(V)B \in \mathbb{R}^{n\times n} \quad \text{and} \quad \psi(V) = \frac{\mathrm{tr}(V^{\mathrm{T}}AV)}{\mathrm{tr}(V^{\mathrm{T}}BV)}. \tag{5.16b}$$

Necessarily, $\Lambda = V^{\mathrm{T}}H(V)V \in \mathbb{R}^{k\times k}$ is symmetric. Evidently, $H(VQ) \equiv H(V)$ for any orthogonal $Q \in \mathbb{R}^{k\times k}$. NEPv (5.16) takes the form of NEPv (1.1). We have the following theorem that characterizes the relation between any global solution V_* of (5.15) and solutions of NEPv (5.16).

Theorem 8 ([29, Theorem 2.1])**.** *$V \in \mathbb{O}^{n\times k}$ is a global maximizer of* (5.15) *if and only if it solves NEPv* (5.16) *and the eigenvalues of $\Lambda \equiv V^{\mathrm{T}}H(V)V$ correspond to the k largest eigenvalues of $H(V)$.*

Theorem 8 transforms TRP (5.15) into NEPv (5.16), and naturally it leads to an SCF iteration for finding the desired solution. The SCF iteration is the same as Algorithm 1, except a simple modification of Λ_i at line 3 to:

$$\Lambda_i = \mathrm{Diag}(\lambda_{n-k+1}(H_i), \ldots, \lambda_n(H_i)),$$

namely, consisting of the k largest eigenvalues of H_i.

In [28, Theorem 5.1], it is shown that SCF iteration is globally convergent to a global maximizer V_* of (5.15) for any given initial guess V_0. In what follows, we will apply the convergence results in section 4 to estimate the local convergence rate of the SCF iteration. To that end, we need to establish the assumption (A3) at the beginning of section 4.2. The next lemma is similar to [28, Theorem 5.2] but with tighter constants.

Lemma 5. *Let $V_* \in \mathbb{O}^{n\times k}$ solve NEPv* (5.16)*. For any $V \in \mathbb{O}^{n\times k}$, we have*

$$|\psi(V_*) - \psi(V)| \le \kappa_{\mathrm{F}} \|\sin\Theta(V, V_*)\|_{\mathrm{F}}^2, \tag{5.17a}$$

$$|\psi(V_*) - \psi(V)| \le \kappa_2 \|\sin\Theta(V, V_*)\|_2^2, \tag{5.17b}$$

where

$$\kappa_{\mathrm{F}} = \frac{\lambda_n(H(V_*)) - \lambda_1(H(V_*))}{\sum_{i=1}^k \lambda_i(B)} \quad \text{and} \quad \kappa_2 = \frac{\sum_{i=1}^k [\lambda_{n-i+1}(H(V_*)) - \lambda_i(H(V_*))]}{\sum_{i=1}^k \lambda_i(B)}. \tag{5.18}$$

Proof. We note that $\mathcal{R}(V_*)$ is the invariant subspace of $H(V_*)$ and

$$\mathrm{tr}(V_*^{\mathrm{T}} H(V_*) V_*) = 0.$$

Viewing $\mathcal{R}(V)$ as an approximate invariant subspace of $H(V_*)$, we have by [11, Theorem 2.2] (see also [22, item 2 of Theorem 3.1])

$$\begin{aligned} \left|\mathrm{tr}(V^{\mathrm{T}} H(V_*) V)\right| &= \left|\mathrm{tr}(V^{\mathrm{T}} H(V_*) V) - \mathrm{tr}(V_*^{\mathrm{T}} H(V_*) V_*)\right| \\ &\le [\lambda_n(H(V_*)) - \lambda_1(H(V_*))] \, \|\sin\Theta(V, V_*)\|_{\mathrm{F}}^2, \end{aligned} \tag{5.19a}$$

$$\left|\mathrm{tr}(V^{\mathrm{T}} H(V_*) V)\right| \le \Big(\sum_{i=1}^k [\lambda_{n-i+1}(H(V_*)) - \lambda_i(H(V_*))]\Big) \, \|\sin\Theta(V, V_*)\|_2^2. \tag{5.19b}$$

On the other hand, since $\left|\mathrm{tr}(V^{\mathrm{T}} H(V_*) V)\right| = \left|\mathrm{tr}(V^{\mathrm{T}} A V) - \psi(V_*)\,\mathrm{tr}(V^{\mathrm{T}} B V)\right|$, we get

$$|\psi(V_*) - \psi(V)| = \frac{\left|\mathrm{tr}(V^{\mathrm{T}} H(V_*) V)\right|}{\mathrm{tr}(V^{\mathrm{T}} B V)} \le \frac{\left|\mathrm{tr}(V^{\mathrm{T}} H(V_*) V)\right|}{\sum_{i=1}^k \lambda_i(B)}$$

which, combined with (5.19), yield (5.17). □

We now are able to establish the quadratic convergence of the SCF iteration for NEPv (5.16).

Theorem 9. *Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric with $B \succ 0$, and V_* be any global solution to TRP* (5.15)*. If $\delta_* := \lambda_{n-k+1}(H(V_*)) - \lambda_{n-k}(H(V_*)) > 0$, then for any given $V_0 \in \mathbb{O}^{n\times k}$, the SCF iteration (*Algorithm 1 *with a modification $\Lambda_i = \mathrm{Diag}(\lambda_{n-k+1}(H_i), \ldots, \lambda_n(H_i))$ at line 3) converges quadratically to $\mathcal{R}(V_*)$. Moreover,*

$$\limsup_{i\to\infty} \frac{\|\sin\Theta(V_i, V_*)\|_{\mathrm{ui}}}{\|\sin\Theta(V_{i-1}, V_*)\|_{\mathrm{ui}}^2} \le \frac{\chi_{\mathrm{ui}}}{\delta_*}, \tag{5.20}$$

where $\mathrm{ui} \in \{2, \mathrm{F}\}$*, $\chi_{\mathrm{ui}} = \kappa_{\mathrm{ui}} \|B\|_{\mathrm{ui}}$ with κ_{ui} given by* (5.18)*.*

Proof. Note by (5.16b) that $\|H(V) - H(V_*)\|_{\mathrm{ui}} = |\psi(V) - \psi(V_*)| \cdot \|B\|_{\mathrm{ui}}$. Thus for $\mathrm{ui} \in \{2, \mathrm{F}\}$, we can use Lemma 5 to obtain

$$\|H(V) - H(V_*)\|_{\mathrm{ui}} = |\psi(V) - \psi(V_*)| \cdot \|B\|_{\mathrm{ui}} \le \kappa_{\mathrm{ui}} \|B\|_{\mathrm{ui}} \cdot \|\sin\Theta(V, V_*)\|_{\mathrm{ui}}^2.$$

Consequently, the assumption (A3) of Theorem 2 for local convergence is satisfied for $q = 2$ due to the fact that

$$\begin{aligned} \limsup_{\|\sin\Theta(V,V_*)\|_{\mathrm{ui}} \to 0} & \frac{\|(I - P_*)[H(V) - H(V_*)]P_*\|_{\mathrm{ui}}}{\|\sin\Theta(V, V_*)\|_{\mathrm{ui}}^2} \\ & \le \limsup_{\|\sin\Theta(V,V_*)\|_{\mathrm{ui}} \to 0} \frac{\|H(V) - H(V_*)\|_{\mathrm{ui}}}{\|\sin\Theta(V, V_*)\|_{\mathrm{ui}}^2} \\ & \le \kappa_{\mathrm{ui}} \|B\|_{\mathrm{ui}} = \chi_{\mathrm{ui}}. \end{aligned} \tag{5.21}$$

Thus, by Theorem 2, the locally quadratic convergence of the SCF iteration immediately follows under the assumption of the eigenvalue gap $\delta_* = \lambda_{n-k+1}(H(V_*)) - \lambda_{n-k}(H(V_*)) > 0$. □

Example 2. We present an example to demonstrate the local quadratic convergence revealed in Theorem 9. Let $A = Z\,\mathrm{Diag}(1, 2, \ldots, n)Z$ and $B = L_m \otimes I_m + I_m \otimes L_m + \alpha I_n$, where $Z = I_n - 2\mathbf{1}\mathbf{1}^{\mathrm{T}}/n$ is a Householder matrix, $\mathbf{1}$ is a vector of all ones, $B \in \mathbb{R}^{n\times n}$ is a regularized standard 2-D discrete Laplacian on the unit square based upon a 5-point stencil with equally-spaced mesh points and $L_m = \mathrm{tridiag}(-1, 2, -1)$. $\alpha > 0$ is a reguarlization parameter, usually determined by the cross-validation technique over a prescribed set [28].

To numerically demonstrate the quadratic convergence rate, we take $n = 400$ (i.e., $m = 20$), $k = 10$ and $\alpha_i = 2i$ for $i = 0, 1, \ldots, 10$. For each α_i, the SCF iteration starts with $V_0 = [e_1, \ldots, e_k]$ and terminates and returns $\widehat{V}_*$ whenever NRes_i in (4.1) is no larger than 10^{-14}. $\widehat{V}_*$ is then treated as an "exact" solution. The "observed"' quadratic rate is taken to be

$$\widehat{\tau} \approx \frac{\|\sin\Theta(V_i, \widehat{V}_*)\|_2}{\|\sin\Theta(V_{i-1}, \widehat{V}_*)\|_2^2},$$

for i near the end of the SCF iteration. Correspondingly, the "estimated" theoretical quadratic rate is taken to be $\widehat{\chi}/\widehat{\delta}_*$, where

$$\widehat{\chi} \approx \frac{\|(I - \widehat{P}_*)[H(V_i) - H(\widehat{V}_*)]\widehat{P}_*\|_2}{\|\sin\Theta(V_i, \widehat{V}_*)\|_2^2} \quad \text{and} \quad \widehat{\delta}_* = \lambda_{n-k+1}(H(\widehat{V}_*)) - \lambda_{n-k}(H(\widehat{V}_*))$$

for i near the end of the SCF iteration. Figure 5.1 shows both $\widehat{\tau}$ and $\widehat{\chi}/\widehat{\delta}_*$ for different values of α. As we can see that the "estimated" quadratic convergence rates are generally tight as upper bounds for the "observed" ones.

6 Concluding remarks

In this paper, we identified two sufficient conditions for the existence and uniqueness of the NEPv (1.1), namely Lipschitz-like conditions (3.1) and a uniform eigenvalue gap condition (3.2). The latter one is undoubtedly strong and may be hard to verify in general unless the coefficient matrix $H(V)$ is very special such as the one on the Hartree-Fock differential equations by Cancès and Le Bris [4].

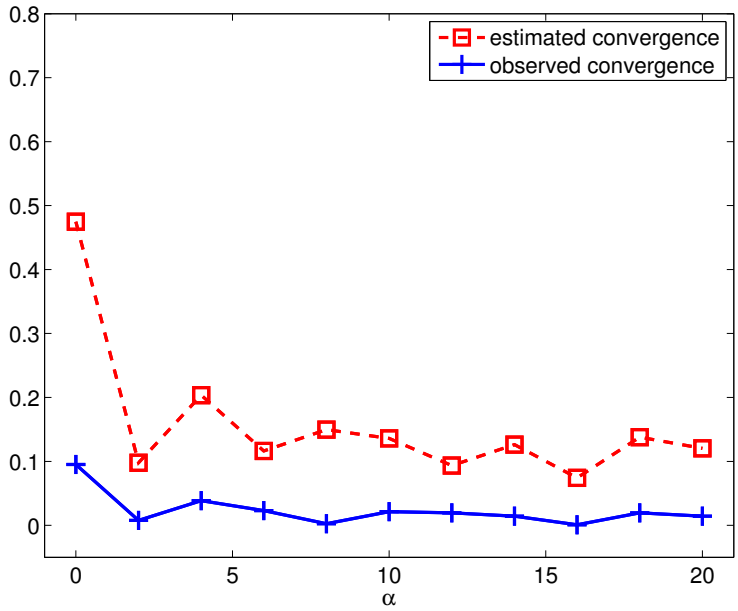

Figure 5.1: Estimated quadratic convergence rate $\widehat{\chi}/\widehat{\delta}_*$ and observed rate $\widehat{\tau}$ of the SCF iteration for solving NEPv (5.16) as α varies .

Throughout the paper, we have assumed (1.2), i.e., $H(V) \equiv H(VQ)$ for any unitary $Q \in \mathbb{C}^{k\times k}$ which makes $H(V)$ a matrix-valued function on the Grassmann manifold of k-dimensional subspaces. As a result, Lipschitz-like conditions and the convergence results of the SCF iteration are stated in terms of the sine of the canonical angles between the subspaces. Looking beyond (1.2), we point out that most of our developments can still be adapted to the situations where (1.2) is no longer true. Possible modifications include, in general, replacing all $\|\sin\Theta(V,\widetilde{V})\|$ by $\|V-\widetilde{V}\|$.

We presented local and global convergence analysis for the plain SCF iteration (Algorithm 1.1) for solving NEPv (1.1), and showed their applications to discrete Kohn-Sham NEPv (5.1) and the trace ratio problem (5.15). For these applications, we are able to demonstrate the near-optimality of the convergence rates revealed in this paper. Furthermore, for the the instance of the Kohn-Sham problem (5.1), we have significantly improved the previous results in [25, 13]. Our analysis so far has been on the plain SCF iteration, i.e., without incorporating any accelerated schemes such as the ones in [17, 23]. It would be an interesting topic to examine whether our analysis can be carried over to those accelerated SCF iterations.

References

[1] W. Bao and Q. Du, *Computing the ground state solution of Bose–Einstein condensates by a normalized gradient flow*, SIAM J. Sci. Comput., 25 (2006), pp. 1674–1697.

[2] R. Bhatia, *Matrix Analysis*, Springer, New York, 1996.

[3] Y. Cai, Z. Bai, J. E. Pask, and N. Sukumar, *Hybrid preconditioning for iterative diagonalization of ill-conditioned generalized eigenvalue problems in electronic structure calculations*, J. Comput. Phys., 255 (2013), pp. 16 –30.

[4] E. Cancès and C. L. Bris, *On the convergence of SCF algorithms for the Hatree–Fock equations*, Math. Model. Numer. Anal., 34 (2000), pp. 749–774.

[5] H. Chen, X. Dai, X. Gong, L. He, and A. Zhou, *Adaptive finite element approximations for Kohn-Sham models*, Multiscale Model. Simul., 12 (2014), pp. 1828–1869.

[6] C. Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation. III*, SIAM J. Numer. Anal., 7 (1970), pp. 1–46.

[7] E. Jarlebring, S. Kvaal, and W. Mechiels, *An inverse iteration method for eigenvalue problems with eigenvectors nonlinearities*, SIAM J. Sci. Comput., 36 (2014), pp. A1978–A2001.

[8] S.-H. Jia, H.-H. Xie, M.-T. Xie, and F. Xu, *A full multigrid method for nonlinear eigenvalue problems*, SCIENCE CHINA Math., 59 (2016), pp. 2037–2048.

[9] L. Jost, S. Setzer, and M. Hein, *Nonlinear eigenproblems in data analysis: balanced graph cuts and the RatioDCA-Prox*, in Extraction of Quantifiable Information from Complex Systems, S. Dahlke, W. Dahmen, M. Griebel, W. Hackbusch, R. Schneider, C. Schwab, and H. Yserentant, eds., vol. 102 of Lecture Notes in Computational Science and Engineering, Springer, Berlin, 2014, pp. 263–279.

[10] M. A. Khamsi, *Introduction to metric fixed point theory*, in Topics in Fixed Point Theory, S. Almezel, Q. H. Ansari, and M. A. Khamsi, eds., Springer International Publishing, Cham, 2014, pp. 1–32.

[11] A. V. Knyazev and M. E. Argentati, *Rayleigh-Ritz majorization error bounds with applications to FEM*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 1521–1537.

[12] R.-C. Li, *Matrix perturbation theory*, in Handbook of Linear Algebra, L. Hogben, ed., CRC Press, Boca Raton, FL, 2nd ed., 2014, p. Chapter 21.

[13] X. Liu, X. Wang, Z. Wen, and Y. Yuan, *On the convergence of the self-consistent field iteration in Kohn–Sham density functional theory*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 546–558.

[14] R. M. Martin, *Electronic structure: basic theory and practical methods*, Cambridge University Press, Cambridge, UK, 2004.

[15] T. Ngo, M. Bellalij, and Y. Saad, *The trace ratio optimization problem for dimensionality reduction*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2950–2971.

[16] M. C. Payne, M. P. Teter, D. C. Allen, T. A. Arias, and J. D. Joannopoulos, *Iterative minimization techniques for ab initio total energy calculation: Molecular dynamics and conjugate gradients*, Rev. Modern Phys., 64 (1992), pp. 1045–1097.

[17] P. Pulay, *Improved SCF convergence*, J. Comput. Chem., 3 (1982), pp. 556–560.

[18] L. Qiu, Y.-X. Zhang, and C.-K. Li, *Unitarily invariant metrics on the Grassmann space*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 507–531.

[19] Y. Saad, J. R. Chelikowsky, and S. M. Shontz, *Numerical methods for electronic structure calculations of materials*, SIAM Rev., 52 (2010), pp. 3–54.

[20] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.

[21] J.-G. Sun, *Matrix Perturbation Analysis*, Academic Press, Beijing, 1987. In Chinese.

[22] Z.-M. Teng, L.-Z. Lu, and R.-C. Li, *Cluster-robust accuracy bounds for Ritz subspaces*, Linear Algebra Appl., 480 (2015), pp. 11–26.

[23] H. Walker and P. Ni, *Anderson acceleration for fixed-point iterations*, SIAM J. Numer. Anal., 49 (2011), pp. 1715–1735.

[24] P.-Å. Wedin, *On angles between subspaces*, in Matrix Pencils, B. Kågström and A. Ruhe, eds., New York, 1983, Springer, pp. 263–285.

[25] C. Yang, W. Gao, and J. C. Meza, *On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 30 (2009), pp. 1773–1788.

[26] C. Yang, J. C. Meza, and L.-W. Wang, *A trust region direct constrained minimization algorithm for the Kohn-Sham equation*, SIAM J. Sci. Comput., 29 (2007), pp. 1854–1875.

[27] L.-H. Zhang and R.-C. Li, *Maximization of the sum of the trace ratio on the Stiefel manifold, II: Computation*, SCIENCE CHINA Math., 58 (2015), pp. 1549–1566.

[28] L.-H. Zhang, L.-Z. Liao, and M. K. Ng, *Fast algorithms for the generalized Foley-Sammon discriminant analysis*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 1584–1605.

[29] L.-H. Zhang, W. Yang, and L.-Z. Liao, *A note on the trace quotient problem*, Opt. Lett., 8 (2014), pp. 1637–1645.

[30] Z. Zhao, Z.-J. Bai, and X.-Q. Jin, *A Riemannian Newton algorithm for nonlinear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 752–774.