SIAM J. MATRIX ANAL. APPL.
Vol. 43, No. 1, pp. 301–327

SHARP ESTIMATION OF CONVERGENCE RATE FOR SELF-CONSISTENT FIELD ITERATION TO SOLVE EIGENVECTOR-DEPENDENT NONLINEAR EIGENVALUE PROBLEMS*

ZHAOJUN BAI†, REN-CANG LI‡, AND DING LU§

Abstract. We present a comprehensive convergence analysis for the self-consistent field (SCF) iteration to solve a class of nonlinear eigenvalue problems with eigenvector dependency (NEPvs). Using the tangent-angle matrix as an intermediate measure for approximation error, we establish new formulas for two fundamental quantities that characterize the local convergence behavior of the plain SCF: the local contraction factor and the local asymptotic average contraction factor. In comparison with previously established results, new convergence rate estimates provide much sharper bounds on the convergence speed. As an application, we extend the convergence analysis to a popular SCF variant—the level-shifted SCF. The effectiveness of the convergence rate estimates is demonstrated numerically for NEPvs arising from solving the Kohn–Sham equation in electronic structure calculation and the Gross–Pitaevskii equation for modeling of the Bose–Einstein condensation.

Key words. nonlinear eigenvalue problem, self-consistent field iteration, convergence factor, level-shifted SCF

AMS subject classifications. 65F15, 65H17

DOI. 10.1137/20M136606X

1. Introduction.

We consider the following nonlinear eigenvalue problem with eigenvector dependency (NEPv): find an orthonormal matrix $V \in \mathbb{C}^{n\times k}$, i.e., $V^{\mathrm{H}}V = I_k$, and a square matrix $\Lambda \in \mathbb{C}^{k\times k}$ satisfying

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

where $H\colon \mathbb{C}^{n\times k} \to \mathbb{C}^{n\times n}$ is a continuous Hermitian matrix-valued function of V. Necessarily, $\Lambda = V^{\mathrm{H}}H(V)V$ and the eigenvalues of Λ are k of the eigenvalues of $H(V)$, often either the k smallest or largest ones. Our later analysis will focus on Λ associated with the k smallest eigenvalues of $H(V)$, but it works equally well for the case when Λ is associated with the k largest ones. We assume throughout this paper that $H(V)$ is right-unitarily invariant in V, i.e.,

$$H(VQ) = H(V) \quad \text{for any unitary } Q \in \mathbb{U}^{k\times k}, \tag{1.2}$$

where $\mathbb{U}^{k\times k}$ is the set of all $k \times k$ unitary matrices. This property (1.2) essentially says that NEPv (1.1) is eigenspace-dependent, to be more precise. However, we will adopt the notion of the *nonlinear eigenvalue problem with eigenvector dependency*, as commonly used in literature. Furthermore, the assumption (1.2) implies that if (V, Λ)

*Received by the editors September 10, 2020; accepted for publication (in revised form) by K. Meerbergen October 25, 2021; published electronically February 28, 2022.
https://doi.org/10.1137/20M136606X
Funding: The first author was supported by NSF grant DMS-1913364. The second author was supported in part by NSF grants DMS-1719620 and DMS-2009689. The third author was supported by NSF grant DMS-2110731.

†Department of Computer Science, University of California, Davis, Davis, CA 95616 USA (bai@cs.ucdavis.edu).
‡Department of Mathematics, University of Texas at Arlington, Arlington, TX 76019 USA (rcli@uta.edu).
§Department of Mathematics, University of Kentucky, Lexington, KY 40506 USA (ding.lu@uky.edu).

is a solution of NEPv (1.1), then so is $(VQ, Q^{\mathrm{H}}\Lambda Q)$ for any unitary Q. We therefore view V and $\widetilde{V}$ as an identical solution if the two share a common range $\mathcal{R}(V) = \mathcal{R}(\widetilde{V})$.

NEPvs in the form of (1.1) arise in a number of areas of computational science and engineering. They are the discrete representations of the Kohn–Sham equation of the density functional theory in electronic structure calculations [19, 35] and the Gross–Pitaevskii equation in modeling the ground state wave function in a Bose–Einstein condensate [4, 11]. In particular, $H(V) = \phi(P)$, where ϕ is a Hermitian matrix-valued function of $P = VV^{\mathrm{H}}$, known as the density matrix in the density functional theory [19, 35]. NEPvs have also long played an important role in the classical methods for data analysis, such as multidimensional scaling [21]. It has become increasingly popular recently in the fields of machine learning and network science, such as the trace ratio maximizations for dimensional reduction [22, 42], balanced graph cut [13], robust Rayleigh quotient maximization for handling data uncertainty [2], core-periphery detection in networks [37], and orthogonal canonical correlation analysis [43]. The unitary invariance (1.2) holds in all those practical NEPvs except a few.

The self-consistent field (SCF) iteration is the most general and widely used method to solve NEPv (1.1). SCF, first introduced in molecular quantum mechanics in the 1950s [29], serves as an entrance to all other approaches. Starting with an orthonormal matrix $V_0 \in \mathbb{U}^{n\times k}$, SCF computes iteratively V_{i+1} and Λ_{i+1} satisfying

$$H(V_i)V_{i+1} = V_{i+1}\Lambda_{i+1} \quad \text{for} \quad i = 0, 1, 2, \ldots, \tag{1.3}$$

where $V_{i+1} \in \mathbb{C}^{n\times k}$ is orthonormal and Λ_{i+1} is a diagonal matrix consisting of the k smallest eigenvalues of $H(V_i)$. Since unit eigenvectors associated with simple eigenvalues can differ by scalar factors of unimodular complex numbers and those associated with multiple eigenvalues have even more freedom, the iteration matrix V_{i+1} cannot be uniquely defined. But thanks to the property (1.2), the computed subspaces $\mathcal{R}(V_1), \mathcal{R}(V_2), \ldots$ are always the same, provided the kth and $(k+1)$st eigenvalues of $H(V_i)$ are distinct at the ith iteration. Because of this, SCF can be interpreted as an iteration of subspaces of dimension k, i.e., elements in the Grassmann manifold $\mathbf{Gr}(k, \mathbb{C}^n)$ of all k-dimensional subspaces of $\mathbb{C}^n$.

The procedure in (1.3) is an SCF in its simplest form, also known as the plain SCF iteration. In practice, such a procedure is prone to slow convergence and sometimes may not converge [14]. Therefore, it has been a fundamental problem of intensive research for decades to understand when and how the plain SCF converges so as to develop remedies to stabilize and accelerate the SCF iteration.

For the applications of solving the Kohn–Sham equation in physics and quantum chemistry, the solution of the associated NEPv corresponds to the minimizer of an energy function. In such context, optimization techniques can be employed to establish convergence results of SCF. A number of convergence conditions have been investigated [7, 17, 18, 40]. For solving general NEPvs, one may view the plain SCF (1.3) as a simple fixed-point iteration. Sufficient conditions for the fixed-point map being a contraction have been studied in [5], where the authors revealed a convergence rate of SCF based on the Davis–Kahan SinΘ theorem [8]. Another approach for the fixed-point analysis is to examine the spectral radius of the Jacobian supermatrix of the fixed-point map. When $H(V)$ is a smooth function in the density matrix $P = VV^{\mathrm{H}}$, a closed-form expression of the Jacobian has been obtained in a recent work [38]. Similar analysis appeared in an earlier work [32] on the Hartree–Fock equation.

What is often different among the existing convergence analyses is the way of measuring the approximation error. Since SCF is a subspace iteration, how to assess

the distance between two subspaces $\mathcal{R}(V)$ and $\mathcal{R}(V_*)$ is the key to the convergence analysis. Various distance measures have been applied in the literature, leading to different approaches of analysis and different types of convergence results. In particular, the difference in density matrices in 2-norm is used as a measure of distance in [40]. A chordal 2-norm is used in [17]. More recent work [5] turned to the sines of the canonical angles between subspaces. The work [7] (as well as [38] though not explicitly specified) used the difference of density matrices in the Frobenius norm. We believe that those distance measures may not necessarily be the best to capture the local convergence rate of the SCF iteration.

The results presented in this paper are a refinement and extension of the previous ones in [5, 17, 38, 40]. We aim to provide a comprehensive and unified local convergence analysis of SCF. Rather than resorting to a specific distance measure, our development is based on the tangent-angle matrix, associated with the tangents of canonical angles of two subspaces. Such matrices can precisely capture the error recurrence of SCF when close to convergence, and they can act as intermediate measurements by which various distance measures can be evaluated as needed. Although they are less popular than sines, the tangents of canonical angles have been used to assess the distance between subspaces and can lead to tighter bounds when applicable; see [8, 45] and references therein.

The use of the tangent-angle matrix allows us to take a closer examination of the local error recursion of SCF, leading to the following contributions presented in this paper:

(a) A precise characterization for the local contraction factor of SCF for both continuous and differentiable $H(V)$. This improves over the previous work [5, 17, 40], where only upper bounds of such a quantity were obtained.
(b) A closed-form formula for the local asymptotic average contraction factor of SCF in terms of the spectral radius of an underlying linear operator when $H(V)$ is differentiable. The formula is sharp for providing a sufficient and almost necessary local convergence condition of SCF. It extends the previous work in [32, 38] to general $H(V)$ functions and has a compact expression that is convenient to work with in both theory and computation.
(c) A new justification for a commonly used level-shifting scheme for the stabilization and acceleration of SCF [7]. A closed-form lower bound on the shifting parameter to guarantee local convergence is obtained.

The rest of the paper is organized as follows. Section 2 presents some preliminaries to set up basic definitions and assumptions. Section 3 introduces the tangent-angle matrix and establishes the recurrence relation of such matrices in consecutive SCF iterations. Section 4 is devoted to the local convergence theory of the plain SCF iteration. Section 5 deals with the level-shifted SCF and its convergence. Numerical illustrations are in section 6, followed by conclusions in section 7.

We follow the notation convention in matrix analysis: $\mathbb{R}^{m\times n}$ and $\mathbb{C}^{m\times n}$ are the sets of $m\times n$ real and complex matrices, respectively, and $\mathbb{R}^n = \mathbb{R}^{n\times 1}$ and $\mathbb{C}^n = \mathbb{C}^{n\times 1}$. $\mathbb{U}^{m\times n} \subset \mathbb{C}^{m\times n}$ denotes the set of $m \times n$ complex orthonormal matrices. A^{T} and A^{H} are the transpose and conjugate transpose of a matrix or a vector A, respectively, and $\overline{A}$ takes entrywise conjugate. $H_1 \succeq H_2$ means that H_1 and H_2 are Hermitian matrices, and $H_1 - H_2$ is positive semidefinite. For a matrix $H \in \mathbb{C}^{n\times n}$ known to have real eigenvalues only, $\lambda_i(H)$ is the ith eigenvalue of H in the ascending order, i.e., $\lambda_1(H) \leqslant \lambda_2(H) \leqslant \cdots \leqslant \lambda_n(H)$, and $\lambda_{\min}(H) = \lambda_1(H)$ and $\lambda_{\max}(H) = \lambda_n(H)$. $\mathrm{Diag}(x)$ is a diagonal matrix made of the vector x, and $\mathrm{diag}(X)$ is a vector consisting of the diagonal elements of a matrix X; $\mathcal{R}(X)$ is the range of X; $\sigma(X)$ is the collection

of all singular values of X. $\Re(\cdot)$ and $\Im(\cdot)$ extract the real and imaginary parts of a complex number, and, when applied to a matrix/vector, they are understood in the elementwise sense. Standard big-O and little-o notations in mathematical analysis are used: for functions $f(x), g(x) \to 0$ as $x \to 0$, write $f(x) = \mathcal{O}(g(x))$ if $|f(x)| \leqslant c|g(x)|$ for some constant c as $x \to 0$, and write $f(x) = \mathbf{o}(g(x))$ if $|f(x)|/|g(x)| \to 0$ as $x \to 0$. Other notations will be explained at their first appearance.

2. Preliminaries. Throughout this paper, we denote by $V_* \in \mathbb{U}^{n\times k}$ a solution of NEPv (1.1). The eigendecomposition of $H(V_*)$ is given by

$$H(V_*)\,[V_*, V_{*\perp}] = [V_*, V_{*\perp}] \begin{bmatrix} \Lambda_* & \\ & \Lambda_{*\perp} \end{bmatrix}, \tag{2.1}$$

where $[V_*, V_{*\perp}] \in \mathbb{U}^{n\times n}$ is unitary, and

$$\Lambda_* = \mathrm{diag}(\lambda_1, \ldots, \lambda_k) \text{ and } \Lambda_{*\perp} = \mathrm{diag}(\lambda_{k+1}, \ldots, \lambda_n)$$

are diagonal matrices containing the eigenvalues of $H(V_*)$ in the ascending order, i.e., $\lambda_i = \lambda_i(H(V_*))$. We make the following assumption for the solution V_* of NEPv (1.1) under consideration.

Assumption 1. There is a positive eigenvalue gap:

$$\delta_* := \lambda_{k+1}(H(V_*)) - \lambda_k(H(V_*)) > 0. \tag{2.2}$$

Such an assumption, which is commonly required in the convergence analysis of SCF, guarantees the uniqueness of the eigenspace corresponding to the k smallest eigenvalues of $H(V_*)$ [5, 7, 17, 38, 40].

Sylvester equation. The following Sylvester equation in $X \in \mathbb{C}^{n\times k}$ will be needed in our analysis:

$$\Lambda_{*\perp} X - X\Lambda_* = V_{*\perp}^{\mathrm{H}}[H(V_*) - H(V)]V_*. \tag{2.3}$$

Under Assumption 1, this equation has a unique solution $X \equiv S(V)$ for each $V \in \mathbb{U}^{n\times k}$, given by

$$S(V) = D(V_*) \odot \left(V_{*\perp}^{\mathrm{H}}[H(V_*) - H(V)]V_*\right), \tag{2.4}$$

where

$$D(V_*) \in \mathbb{R}^{(n-k)\times k} \quad \text{with } D(V_*)_{ij} = (\lambda_{k+i}(H(V_*)) - \lambda_j(H(V_*)))^{-1}, \tag{2.5}$$

and $\odot$ denotes the Hadamard product, i.e., elementwise multiplication.

Unitarily invariant norm. We denote by $\|\cdot\|_{\mathrm{ui}}$ a unitarily invariant norm, which, besides being a matrix norm, also satisfies the following two additional conditions:

(1) $\|XAY\|_{\mathrm{ui}} = \|A\|_{\mathrm{ui}}$ for any unitary matrices X and Y;

(2) $\|A\|_{\mathrm{ui}} = \|A\|_2$ whenever A is rank-1, where $\|\cdot\|_2$ is the spectral norm.

It is well known that $\|A\|_{\mathrm{ui}}$ is dependent only on the singular values of A. In this paper, we assume any $\|\cdot\|_{\mathrm{ui}}$ we use is applicable to matrices of all sizes in a compatible way, i.e., $\|A\|_{\mathrm{ui}} = \|B\|_{\mathrm{ui}}$ for A, B sharing the same set of nonzero singular values (see, e.g., [34, Theorem 3.6, page 78]). The spectral norm $\|\cdot\|_2$ and Frobenius norm $\|\cdot\|_{\mathrm{F}}$ are two examples of such unitarily invariant norms. Unitary invariant norms satisfy

$$\|ABC\|_{\mathrm{ui}} \leqslant \|A\|_2 \cdot \|B\|_{\mathrm{ui}} \cdot \|C\|_2 \tag{2.6}$$

for any matrices A, B, and C of compatible sizes (see, e.g., [34, Theorem 3.9, page 80]).

Canonical angles between subspaces. Let $X, Y \in \mathbb{U}^{n\times k}$. The k canonical angles between subspaces $\mathcal{X} = \mathcal{R}(X)$ and $\mathcal{Y} = \mathcal{R}(Y)$ are defined as

$$0 \leqslant \theta_j(\mathcal{X}, \mathcal{Y}) := \arccos \sigma_j \leqslant \frac{\pi}{2} \quad \text{for } 1 \leqslant j \leqslant k, \tag{2.7}$$

where $\sigma_1 \geqslant \cdots \geqslant \sigma_k$ are singular values of the matrix $Y^{\mathrm{H}}X$ (see, e.g., [34, section 4.2.1]). Put k canonical angles all together to define

$$\Theta(\mathcal{X}, \mathcal{Y}) = \operatorname{diag}(\theta_1(\mathcal{X}, \mathcal{Y}), \ldots, \theta_k(\mathcal{X}, \mathcal{Y})). \tag{2.8}$$

Since the canonical angles defined above are independent of the basis matrices X and Y, for convenience, we use the notation $\Theta(X, Y)$ interchangeably with $\Theta(\mathcal{X}, \mathcal{Y})$.

Canonical angles provide a natural distance measure for subspaces. For any unitarily invariant norm $\|\cdot\|_{\mathrm{ui}}$, it holds that both $\|\Theta(X, Y)\|_{\mathrm{ui}}$ and $\|\sin\Theta(X, Y)\|_{\mathrm{ui}}$ are unitarily invariant metrics on the Grassmann manifold $\mathbf{Gr}(k, \mathbb{C}^n)$ (see, e.g., [34, Theorem 4.10, page 93] and [27]). In our analysis, the tangents of canonical angles will play an important role. By trigonometric function properties, tangents provide good approximation to the canonical angles as $\Theta(X, Y) \to 0$:

$$\tan\Theta(X, Y) = \Theta(X, Y) + \mathcal{O}(\|\Theta(X, Y)\|_{\mathrm{ui}}^3). \tag{2.9}$$

$\mathbb{R}$-linear mapping. A mapping $\mathscr{L}\colon \mathbb{C}^{n\times k} \to \mathbb{C}^{p\times q}$ is called $\mathbb{R}$-linear if it satisfies

$$\mathscr{L}(X + Y) = \mathscr{L}(X) + \mathscr{L}(Y) \quad \text{and} \quad \mathscr{L}(\alpha\, X) = \alpha\, \mathscr{L}(X) \tag{2.10}$$

for all $X, Y \in \mathbb{C}^{n\times k}$ and $\alpha \in \mathbb{R}$. When we talk about an $\mathbb{R}$-linear mapping, the complex matrix space $\mathbb{C}^{m\times n}$ is viewed as a vector space over the field $\mathbb{R}$ of real numbers, denoted by $\mathbb{C}^{m\times n}(\mathbb{R})$. By elementary linear algebra, $\mathbb{C}^{m\times n}(\mathbb{R})$ is a $(2mn)$-dimensional inner product space, equipped with the inner product $\langle X, Y\rangle := \Re\operatorname{tr}(X^{\mathrm{H}}Y)$ and the induced norm $\|X\|_{\mathrm{F}} = (\Re\operatorname{tr}(X^{\mathrm{H}}X))^{1/2}$. We can see that $\mathscr{L}\colon \mathbb{C}^{n\times k}(\mathbb{R}) \to \mathbb{C}^{p\times q}(\mathbb{R})$ is a linear mapping (over $\mathbb{R}$). For convenience, we use $\mathbb{C}^{n\times k}$ and $\mathbb{C}^{n\times k}(\mathbb{R})$ interchangeably in future discussions when referring to an $\mathbb{R}$-linear mapping.

The *spectral radius* of an $\mathbb{R}$-linear operator $\mathscr{L} : \mathbb{C}^{n\times k} \to \mathbb{C}^{n\times k}$ is defined as the largest eigenvalue in magnitude of a matrix representation $\mathbf{L} \in \mathbb{R}^{(2nk)\times(2nk)}$ of $\mathscr{L}$:

$$\rho(\mathscr{L}) := \max\left\{\ |\lambda|\colon \mathbf{L}\,\mathbf{x} = \lambda\,\mathbf{x},\ \mathbf{x} \in \mathbb{C}^{2nk}\ \right\}. \tag{2.11}$$

Notice that $\mathbf{x}$ is allowed to be a complex vector because a real matrix can have complex eigenvalues. Here we do not make any assumption on the basis used to obtain $\mathbf{L}$; the choice of the basis does not affect the spectrum of $\mathbf{L}$ and therefore $\rho(\mathscr{L})$.

Derivative operator. Let $V = V_{\mathbf{r}} + \imath V_{\mathbf{i}} \in \mathbb{C}^{n\times k}$ with $V_{\mathbf{r}}, V_{\mathbf{i}} \in \mathbb{R}^{n\times k}$ being the real and imaginary parts of V, respectively. A Hermitian matrix-valued function $H(V)$ is called differentiable if each element $h_{ij}(V)$ is a smooth function in the real and imaginary parts $(V_{\mathbf{r}}, V_{\mathbf{i}})$ of V. Such differentiability is also known as real differentiability in the literature (see, e.g., [28, 10]), and it is different from the one in the holomorphic sense, which generally cannot hold for $H(V)$ with real diagonal elements.

For $H(V)$ differentiable at V_*, we can define a derivative operator

$$\mathbf{D}H(V_*)[\cdot]\colon \mathbb{C}^{n\times k} \to \mathbb{C}^{n\times n} \quad \text{with} \quad \mathbf{D}H(V_*)[X] = \left[\frac{d}{dt}H(V_* + tX)\right]_{t=0}, \tag{2.12}$$

where $t \in \mathbb{R}$. The $\mathbf{D}H(V_*)[X]$ represents the directional derivative of $H(V)$ at V_* in the direction of $X = X_{\mathbf{r}} + \imath X_{\mathbf{i}} \in \mathbb{C}^{n\times k}$. By definition, the (i,j) entry of $\mathbf{D}H(V_*)[X]$ is given by

(2.13)
$$\mathbf{D}h_{ij}(V_*)[X] = \left[\frac{d}{dt}h_{ij}(V_* + tX)\right]_{t=0} = \left\langle \frac{\partial h_{ij}}{\partial V_{\mathbf{r}}}(V_*)\,,\, X_{\mathbf{r}}\right\rangle + \left\langle \frac{\partial h_{ij}}{\partial V_{\mathbf{i}}}(V_*)\,,\, X_{\mathbf{i}}\right\rangle,$$

where $\langle X, Y\rangle := \mathrm{tr}(X^{\mathrm{T}}Y)$. It follows that the elements $\mathbf{D}h_{ij}(V_*)[\cdot]$, and hence the full $\mathbf{D}H(V_*)[\cdot]$, are $\mathbb{R}$-linear mappings satisfying (2.10).

By Taylor's expansion, as V comes close to V_* (in the Euclidean sense), it holds that

$$\begin{aligned} h_{ij}(V) &= h_{ij}(V_*) + \left\langle \frac{\partial h_{ij}}{\partial V_{\mathbf{r}}}(V_*)\,,\, V_{\mathbf{r}} - V_{*,\mathbf{r}}\right\rangle + \left\langle \frac{\partial h_{ij}}{\partial V_{\mathbf{i}}}(V_*)\,,\, V_{\mathbf{i}} - V_{*,\mathbf{i}}\right\rangle + \mathbf{o}(\|V - V_*\|) \\ &= h_{ij}(V_*) + \mathbf{D}h_{ij}(V_*)[V - V_*] + \mathbf{o}(\|V - V_*\|), \end{aligned}$$

where $\|\cdot\|$ is any matrix norm, and the last equation is due to (2.13). It then follows that

$$H(V) = H(V_*) + \mathbf{D}H(V_*)[V - V_*] + \mathbf{o}(\|V - V_*\|). \tag{2.14}$$

Namely, $\mathbf{D}H(V_*)[\,\cdot\,]$ is the Fréchet derivative of $H : \mathbb{C}^{n\times k}(\mathbb{R}) \to \mathbb{C}^{n\times n}(\mathbb{R})$. Note that the expansion (2.14) does not take into account the unitary invariance (1.2) of $H(V)$, and that is why the remainder is in terms of the Euclidean difference $V - V_*$.

The Fréchet derivatives for matrix-valued functions and the $\mathbb{R}$-linear operators are essential tools for numerical analysis over the Grassmann manifold; see, e.g., [1] and references therein.

3. Tangent-angle matrix. Let $V \in \mathbb{U}^{n\times k}$ be an approximation to the solution V_* of NEPv (1.1). Each V represents an orthonormal basis matrix of a subspace. As far as a solution of NEPv (1.1) is concerned, it is the subspace that matters. To assess the distance of V to the solution V_* in terms of the subspaces their columns span, we define the *tangent-angle matrix* from V to V_* as

$$T(V) := (V_{*\perp}^{\mathrm{H}}V)(V_*^{\mathrm{H}}V)^{-1} \in \mathbb{C}^{(n-k)\times k}, \tag{3.1}$$

provided $V_*^{\mathrm{H}}V$ is invertible, and we recall (2.1) for $V_{*\perp}$. By definition, $T(V)$ can be viewed as a function of $\mathbb{U}^{n\times k} \to \mathbb{C}^{(n-k)\times k}$. The name of "tangent-angle matrix" comes from the fact that

$$\|\tan\Theta(V, V_*)\|_{\mathrm{ui}} = \|(V_{*\perp}^{\mathrm{H}}V)(V_*^{\mathrm{H}}V)^{-1}\|_{\mathrm{ui}} = \|T(V)\|_{\mathrm{ui}} \tag{3.2}$$

for all unitarily invariant norms. Recall that the unitarily invariant norm $\|A\|_{\mathrm{ui}}$ is defined by the singular values of A; (3.2) is a direct consequence of the identity of singular values $\sigma(\tan\Theta(V, V_*)) = \sigma\left((V_\perp^{\mathrm{H}}V_*)(V^{\mathrm{H}}V_*)^{-1}\right)$, which follows from the definition of canonical angles in (2.8) (see, e.g., [33, Theorems 2.2, 2.4, Chapter 4] and [45]). The tangents of canonical angles have long been used in numerical matrix analysis, and we refer to [45] and references therein.

By definition (2.7), the singular values of $V_*^{\mathrm{H}}V$ consist of those of the matrix $\cos\Theta(V, V_*) = I + \mathcal{O}(\|\Theta(V, V_*)\|_{\mathrm{ui}}^2)$. Therefore, it can be seen from (3.2) that $T(V)$ is well defined for sufficiently small canonical angles $\Theta(V, V_*)$. Meanwhile, $\Theta(V, V_*) \to 0$ if and only if $T(V) \to 0$. By the unitary invariance (1.2) and the continuity of $H(V)$, we have $H(V) \to H(V_*)$ as the tangent-angle matrix $T(V) \to 0$. This is more precisely described in the following lemma.

LEMMA 3.1. *Let $V \in \mathbb{U}^{n\times k}$. Then as $T(V) \to 0$, for a solution V_* of NEPv* (1.1)*, it holds that*

$$H(V) = H\big(V_* + V_{*\perp}T(V) + \mathcal{O}(\|T(V)\|_{\mathrm{ui}}^2)\big). \tag{3.3}$$

If $H(V)$ is also differentiable, then

$$H(V) = H(V_*) + \mathbf{D}H(V_*)[V_{*\perp}T(V)] + \mathbf{o}(\|T(V)\|_{\mathrm{ui}}). \tag{3.4}$$

Proof. The singular values of $V_*^{\mathrm{H}}V$ consist of $\cos\Theta(V, V_*) = I + \mathcal{O}(\|\Theta(V, V_*)\|_{\mathrm{ui}}^2)$. So we have $V_*^{\mathrm{H}}V = W + \mathcal{O}(\|\Theta(V, V_*)\|_{\mathrm{ui}}^2)$ for some unitary $W \in \mathbb{U}^{k\times k}$. It follows that

$$VW^{-1} = V(V_*^{\mathrm{H}}V)^{-1} + \mathcal{O}(\|\Theta(V, V_*)\|_{\mathrm{ui}}^2) = V_* + V_{*\perp}\,T(V) + \mathcal{O}(\|T(V)\|_{\mathrm{ui}}^2), \tag{3.5}$$

where we used $V = V_*(V_*^{\mathrm{H}}V) + V_{*\perp}(V_{*\perp}^{\mathrm{H}}V)$ and $T(V) = \mathcal{O}(\|\Theta(V, V_*)\|_{\mathrm{ui}})$ in the last equation. The unitary invariance property $H(V) = H(VW^{-1})$ leads to (3.3). Combining (3.5) with (2.14), we obtain (3.4). ∎

The following lemma, which is the key to establish our local convergence results, describes the relation between the tangent-angle matrices of two consecutive SCF iterations.

LEMMA 3.2. *Suppose Assumption* 1 *holds. Let $\widetilde{V}$ be an orthonormal basis matrix associated with the k smallest eigenvalues of $H(V)$, and let $S(V)$ be the unique solution of the Sylvester equation defined in* (2.4)*. Then*

(a) *$S(V) \to 0$ as $T(V) \to 0$;*

(b) *the tangent-angle matrix $T(\widetilde{V})$ of $\widetilde{V}$ satisfies*

$$T(\widetilde{V}) = S(V) + \mathbf{o}(\|S(V)\|_{\mathrm{ui}}); \tag{3.6}$$

(c) *if $H(V)$ is differentiable at V_*, then*

$$T(\widetilde{V}) = \mathscr{L}\,(T(V)) + \mathbf{o}(\|T(V)\|_{\mathrm{ui}}), \tag{3.7}$$

where $\mathscr{L} : \mathbb{C}^{(n-k)\times k} \to \mathbb{C}^{(n-k)\times k}$ defined by

$$\mathscr{L}(Z) = D(V_*) \odot (V_{*\perp}^{\mathrm{H}}\,\mathbf{D}H(V_*)[V_{*\perp}Z]\,V_*) \tag{3.8}$$

is an $\mathbb{R}$-linear operator, called the local $\mathbb{R}$-linear operator *of the plain SCF.*

Proof. For item (a), by (3.3) and the continuity of H, it holds that $H(V) \to H(V_*)$ as $T(V) \to 0$. Hence, $S(V) \to 0$ by the definition of $S(V)$.

For item (b), we begin with the eigendecomposition of $H(V)$:

$$H(V)\begin{bmatrix}\widetilde{V} & \widetilde{V}_\perp\end{bmatrix} = \begin{bmatrix}\widetilde{V} & \widetilde{V}_\perp\end{bmatrix}\begin{bmatrix}\widetilde{\Lambda} & \\ & \widetilde{\Lambda}_\perp\end{bmatrix},$$

where $[\widetilde{V}\ \widetilde{V}_\perp] \in \mathbb{U}^{n\times n}$ is unitary, $\widetilde{\Lambda} = \mathrm{diag}(\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_k)$, and $\widetilde{\Lambda}_\perp = \mathrm{diag}(\widetilde{\lambda}_{k+1}, \ldots, \widetilde{\lambda}_n)$ with $\widetilde{\lambda}_i = \lambda_i(H(V))$. Due to Assumption 1, as $H(V) \to H(V_*)$, we can apply the standard perturbation analysis of eigenspaces [34, section V.2] to obtain

$$\begin{bmatrix}\widetilde{V} & \widetilde{V}_\perp\end{bmatrix} = \begin{bmatrix}V_* & V_{*\perp}\end{bmatrix}\begin{bmatrix}I_k & -Z^{\mathrm{H}} \\ Z & I_{n-k}\end{bmatrix}\begin{bmatrix}(I_k + Z^{\mathrm{H}}Z)^{-1/2} & \\ & (I_{n-k} + ZZ^{\mathrm{H}})^{-1/2}\end{bmatrix}\begin{bmatrix}Q & \\ & P\end{bmatrix}, \tag{3.9}$$

where $Z \in \mathbb{R}^{(n-k)\times k}$, $Q \in \mathbb{U}^{k\times k}$, and $P \in \mathbb{U}^{(n-k)\times(n-k)}$ are parameter matrices and

$$Z \to 0 \quad \text{as} \quad H(V) \to H(V_*). \tag{3.10}$$

The parameterization from (3.9) can be equivalently put as

$$\begin{aligned}\widetilde{V} &= (V_* + V_{*\perp} Z)\,(I_k + Z^{\mathrm{H}} Z)^{-1/2} Q,\\ \widetilde{V}_\perp &= (-V_* Z^{\mathrm{H}} + V_{*\perp})\,(I_{n-k} + Z Z^{\mathrm{H}})^{-1/2} P.\end{aligned}$$

By the first equation, Z is identical to the tangent-angle matrix from $\widetilde{V}$ to V_*:

$$T(\widetilde{V}) = (V_{*\perp}^{\mathrm{H}} \widetilde{V})(V_*^{\mathrm{H}} \widetilde{V})^{-1} = Z, \tag{3.11}$$

where we have used $V_*^{\mathrm{H}} \widetilde{V} = (I_k + Z^{\mathrm{H}} Z)^{-1/2} Q$ and $V_{*\perp}^{\mathrm{H}} \widetilde{V} = Z(I_k + Z^{\mathrm{H}} Z)^{-1/2} Q$.

Next, we establish an equation to characterize Z. From $\widetilde{V}_\perp^{\mathrm{H}} H(V) \widetilde{V} = \widetilde{V}_\perp^{\mathrm{H}} \widetilde{V} \widetilde{\Lambda} = 0$, we get

$$\begin{aligned}0 &= \begin{bmatrix} -Z & I_{n-k}\end{bmatrix} [V_*, V_{*\perp}]^{\mathrm{H}} H(V) [V_*, V_{*\perp}] \begin{bmatrix} I_k \\ Z \end{bmatrix}\\ &= \begin{bmatrix} -Z & I_{n-k}\end{bmatrix} [V_*, V_{*\perp}]^{\mathrm{H}} [H(V_*) + (H(V) - H(V_*))] [V_*, V_{*\perp}] \begin{bmatrix} I_k \\ Z \end{bmatrix}\\ &= \Lambda_{*\perp} Z - Z \Lambda_* + (-Z V_*^{\mathrm{H}} + V_{*\perp}^{\mathrm{H}})[H(V) - H(V_*)](V_* + V_{*\perp} Z).\end{aligned}$$

Therefore, Z satisfies the Sylvester equation (view the right-hand side as fixed)

$$\Lambda_{*\perp} Z - Z \Lambda_* = (Z V_*^{\mathrm{H}} - V_{*\perp}^{\mathrm{H}})[H(V) - H(V_*)](V_* + V_{*\perp} Z).$$

By Assumption 1, we can solve the Sylvester equation to obtain

$$Z = S(V) + \Phi(Z), \tag{3.12}$$

where

$$\Phi(Z) = D(V_*) \odot \big(Z V_*^{\mathrm{H}} [H(V) - H(V_*)](V_* + V_{*\perp} Z) - V_{*\perp}^{\mathrm{H}} [H(V) - H(V_*)] V_{*\perp} Z \big)$$

and $D(V_*)$ is defined as in (2.5). A quick calculation shows that

$$\|\Phi(Z)\|_{\mathrm{F}} \leqslant \delta_*^{-1} \|H(V) - H(V_*)\|_{\mathrm{F}} (2\|Z\|_2 + \|Z\|_2^2) = \mathbf{o}(\|Z\|_{\mathrm{ui}}), \tag{3.13}$$

where the last equation is due to $H(V) \to H(V_*)$ and $Z \to 0$, as $T(V) \to 0$, and the equivalency of matrix norms. Recall $T(\widetilde{V}) = Z$. Equations (3.12) and (3.13) lead directly to (3.6).

For item (c), we derive from the definition of $S(V)$ and the expansion (3.4) that

$$S(V) = D(V_*) \odot \big(V_{*\perp}^{\mathrm{H}}\, \mathbf{D} H(V_*)[V_{*\perp} T(V)]\, V_* \big) + \mathbf{o}(\|T(V)\|_{\mathrm{ui}}).$$

Plugging it into (3.6), and exploiting $\|\mathscr{L}(T(V))\|_{\mathrm{ui}} = \mathcal{O}(\|T(V)\|_{\mathrm{ui}})$ since $\mathscr{L}$ is an $\mathbb{R}$-linear operator of finite dimension (which is bounded), we complete the proof. ∎

We mention that the tangent-angle matrix in the form of (3.1) appeared in the so-called McWeeny transformation [20, 31, 32] in the density matrix theory for electronic structure calculations, where the matrix was treated as an independent parameter that is not connected with canonical angles of subspaces. This lack of geometric interpretation makes it difficult to produce a comprehensive convergence analysis as developed in the following sections and extend to the treatment of a continuous $H(V)$.

4. Convergence analysis. Because of the invariance property (1.2), the plain SCF iteration (1.3) should be inherently understood as a subspace iterative scheme, and the convergence of the basis matrices $\{V_i\}_{i=0}^{\infty}$ to a solution V_* should be measured by a metric on the Grassmann manifold $\mathbf{Gr}(k, \mathbb{C}^n)$. Let $d(\cdot, \cdot)$ be a metric on $\mathbf{Gr}(k, \mathbb{C}^n)$. Without causing any ambiguity, in what follows we will not distinguish an element $\mathcal{R}(V) \in \mathbf{Gr}(k, \mathbb{C}^n)$ from its representation $V \in \mathbb{U}^{n\times k}$. We adopt the following notions for the local convergence analysis:

(i) SCF (1.3) is called *locally (R-linearly) convergent* to V_* if there exist $c < 1$, $\varepsilon > 0$, and $m_0 > 0$ such that, for any initial V_0 with $d(V_0, V_*) \leqslant \varepsilon$ and integer $i > m_0$, it holds that $d(V_i, V_*) \leqslant \gamma \cdot c^i \cdot d(V_0, V_*)$, where γ is a constant independent of i and V_0;

(ii) SCF (1.3) is called *locally divergent* from V_* if there exists $c > 1$ such that for any $\varepsilon > 0$ and arbitrarily large $m_0 > 0$, there exist an initial V_0 with $d(V_0, V_*) \leqslant \varepsilon$ and integer $i > m_0$ satisfying $d(V_i, V_*) \geqslant c^i \cdot d(V_0, V_*)$.

The local (R-linear) convergence implies $d(V_i, V_*) \to 0$ as $i \to \infty$ for any V_0 that is sufficiently close to V_* in the metric, i.e., $d(V_0, V_*)$ is sufficiently small. Moreover, the asymptotic rate of convergence is at least linear with a convergence factor c.

4.1. Contraction factors. There are two fundamental quantities that provide convergence measures of SCF on $\mathbf{Gr}(k, \mathbb{C}^n)$: the *local contraction factor* and the *local asymptotic average contraction factor*. The former, which is a quantity to assess local convergence, accounts for the worst case error reduction of SCF per iterative step. The latter captures the asymptotic average convergence rate of SCF and provides a sufficient and almost necessary condition for the local convergence.

Since SCF is a fixed-point iteration on the Grassmann manifold $\mathbf{Gr}(k, \mathbb{C}^n)$, the *local contraction factor of SCF* is defined as

$$\eta_{\sup} := \limsup_{\substack{V_0 \in \mathbb{U}^{n\times k} \\ d(V_0, V_*) \to 0}} \frac{d(V_1, V_*)}{d(V_0, V_*)}. \tag{4.1}$$

Such a constant can be viewed as the (best) local Lipschitz constant for the fixed-point mapping of SCF. We observe that the condition $\eta_{\sup} < 1$, which implies SCF is locally error reductive, is sufficient for local convergence. In the convergent case, it follows from the definition (4.1) that

$$\limsup_{i\to\infty} \frac{d(V_{i+1}, V_*)}{d(V_i, V_*)} \leqslant \eta_{\sup};$$

namely, the *(asymptotic) convergence rate* of SCF is bounded by $\eta_{\sup}$.

To take into account oscillation and to obtain tighter convergence bounds, the one-step contraction factor (4.1) can be generalized to multiple iterative steps. Let m be a given positive integer, and define

$$\eta_{\sup,m} := \limsup_{\substack{V_0 \in \mathbb{U}^{n\times k} \\ d(V_0, V_*) \to 0}} \left(\frac{d(V_m, V_*)}{d(V_0, V_*)} \right)^{1/m}. \tag{4.2}$$

Then $\eta_{\sup,m}$ is a (geometric) average contraction factor of m consecutive iterative steps of SCF (1.3). The limit of the average contraction factor as $m \to \infty$,

$$\eta_{\sup,\infty} := \limsup_{m\to\infty} \eta_{\sup,m}, \tag{4.3}$$

defines a *local asymptotic average contraction factor* of SCF. We mention that the quantities $\eta_{\sup}$, $\eta_{\sup,m}$, and $\eta_{\sup,\infty}$ from above are always well defined under Assumption 1, which implies $\mathcal{R}(V_1), \mathcal{R}(V_m) \in \mathbf{Gr}(k, \mathbb{C}^n)$, for any fixed m, are unique for sufficient small $d(V_0, V_*)$. We should also caution the reader that all those quantities depend on the metric $d(\cdot,\cdot)$ and that the dependency is suppressed for notational clarity.

By definition, the number $\eta_{\sup,\infty}$ measures the average convergence rate of SCF. The average convergence rate is a conventional tool to study matrix iterative methods [39] and typically leads to tight convergence rates in practice.[1] It follows from item (b) of Lemma 4.1 below that $\eta_{\sup,\infty}$ is precisely the local convergence factor of SCF. The properties of contraction factors as shown in Lemma 4.1 have long been known for linear iterative methods; see, e.g., [39, section 3.2]. It is not surprising they also hold for contraction factors defined for a general iterative scheme. Lemma 4.1 is tailored for SCF, and, for the sake of completeness, we present a proof here. It has two assumptions: Assumption 1, which is needed for the SCF iteration to proceed in a neighborhood of V_* to ensure the well-definedness of the contraction factors, and $\eta_{\sup} < \infty$, which will be proved in Theorem 4.2 later under the condition that $H(V)$ is Lipschitz continuous at V_*.

LEMMA 4.1. *Suppose Assumption* 1 *and* $\eta_{\sup} < \infty$.

(a) *It holds that for any* $m > 1$

$$\eta_{\sup,\infty} \leqslant \eta_{\sup,m} \leqslant \eta_{\sup}. \tag{4.4}$$

(b) *If* $\eta_{\sup,\infty} < 1$, *then SCF is locally convergent to* V_*, *with its asymptotic average convergence rate bounded by* $\eta_{\sup,\infty}$. *If* $\eta_{\sup,\infty} > 1$, *then SCF is locally divergent from* V_*.

Proof. For item (a), first from definition (4.1) and $\eta_{\sup} < \infty$, we conclude that $d(V_p, V_*) \to 0$ for $p = 0, 1, \ldots, m-1$ as $d(V_0, V_*) \to 0$. Therefore,

$$\begin{aligned}\limsup_{\substack{V_0 \in \mathbb{U}^{n\times k} \\ d(V_0,V_*)\to 0}} \left(\frac{d(V_m,V_*)}{d(V_0,V_*)}\right)^{1/m} &= \limsup_{\substack{V_0 \in \mathbb{U}^{n\times k} \\ d(V_0,V_*)\to 0}} \left(\prod_{p=0}^{m-1} \frac{d(V_{p+1},V_*)}{d(V_p,V_*)}\right)^{1/m} \\ &\leqslant \left(\prod_{p=0}^{m-1} \limsup_{\substack{V_p \in \mathbb{U}^{n\times k} \\ d(V_p,V_*)\to 0}} \frac{d(V_{p+1},V_*)}{d(V_p,V_*)}\right)^{1/m},\end{aligned}$$

and $\eta_{\sup,m} \leqslant \eta_{\sup}$ follows.

Now fix m. Any integer $m' > m$ can be expressed as $m' = sm + p$ for some $s \geqslant 0$ and $0 \leqslant p \leqslant m-1$. Using the same arguments as from above, and noticing that

$$\begin{aligned}\left(\frac{d(V_{m'},V_*)}{d(V_0,V_*)}\right)^{1/m'} &= \left(\frac{d(V_{m'},V_*)}{d(V_p,V_*)}\frac{d(V_p,V_*)}{d(V_0,V_*)}\right)^{1/m'} \\ &= \left(\prod_{\ell=0}^{s-1} \frac{d(V_{m(\ell+1)+p},V_*)}{d(V_{m\ell+p},V_*)} \cdot \frac{d(V_p,V_*)}{d(V_0,V_*)}\right)^{1/m'},\end{aligned}$$

we obtain by taking lim sup that

$$\eta_{\sup,m'} \leqslant (\eta_{\sup,m})^{sm/m'} \cdot (\eta_{\sup,p})^{p/m'}.$$

[1] The average convergence rate in [39] is defined with an extra logarithm, i.e., $-\ln(\eta_{\sup,m})$.

Assume for the moment $\eta_{\sup,m} \neq 0$. Letting $m' \to \infty$ and noticing that $\eta_{\sup,p} \leqslant \eta_{\sup}$ is bounded, we get $\eta_{\sup,\infty} = \limsup_{m'\to\infty} \eta_{\sup,m'} \leqslant \eta_{\sup,m}$. If, however, $\eta_{\sup,m} = 0$, then $\eta_{\sup,m'} = 0$ for any $m' \geqslant m$, and so $\eta_{\sup,\infty} = 0 \leqslant \eta_{\sup,m}$.

For item (b), consider first $\eta_{\sup,\infty} < 1$. Pick a constant c such that $\eta_{\sup,\infty} < c < 1$. Because of how $\eta_{\sup,m}$ is defined in (4.3), we see that $[d(V_m, V_*)/d(V_0, V_*)]^{1/m} \leqslant c$ for m sufficiently large and for all V_0 sufficiently close to V_* in the metric $d(\cdot,\cdot)$. More precisely, there exist $m_0 > 0$ and $\varepsilon_1 > 0$ such that

$$d(V_{m_0}, V_*) \leqslant c^{m_0}\, d(V_0, V_*) \quad \text{for all } V_0 \text{ with } d(V_0, V_*) < \varepsilon_1.$$

Recall $c < 1$, so we have $d(V_{m_0}, V_*) < \varepsilon_1$. By induction, it holds for all $s = 0, 1, 2, \ldots$ and all V_0 with $d(V_0, V_*) < \varepsilon_1$ that

$$d(V_{sm_0}, V_*) \leqslant c^{m_0}\, d(V_{(s-1)m_0}, V_*) \leqslant \cdots \leqslant c^{sm_0}\, d(V_0, V_*) < \varepsilon_1. \tag{4.5}$$

Recall that $\eta_{\sup} < \infty$, and pick a finite constant $c_2 > \max\{1, \eta_{\sup}\} \geqslant 1$. By (4.1), there exists $\varepsilon_2 \in (0, \varepsilon_1)$ such that

$$d(V_1, V_*) \leqslant c_2\, d(V_0, V_*) \quad \text{for all } V_0 \text{ with } d(V_0, V_*) < \varepsilon_2. \tag{4.6}$$

Let $\varepsilon_3 = c_2^{-(m_0-1)} \times \varepsilon_2 < \varepsilon_2 < \varepsilon_1$. For any V_0 with $d(V_0, V_*) < \varepsilon_3$, we have by (4.6)

$$d(V_1, V_*) \leqslant c_2\, d(V_0, V_*) < c_2\varepsilon_3 \leqslant \varepsilon_2, \tag{4.7a}$$

$$d(V_2, V_*) \leqslant c_2 d(V_1, V_*) \leqslant c_2^2 d(V_0, V_*) < c_2^2\varepsilon_3 \leqslant \varepsilon_2, \tag{4.7b}$$

$$\vdots$$

$$d(V_{m_0-1}, V_*) \leqslant c_2^{m_0-1} d(V_0, V_*) < c_2^{m_0-1}\varepsilon_3 \leqslant \varepsilon_2. \tag{4.7c}$$

Whereas for any $m \geqslant m_0$, we can write $m = sm_0 + p$ for some $0 \leqslant p \leqslant m_0 - 1$. We have by (4.5) and (4.7) that for any V_0 with $d(V_0, V_*) < \varepsilon_3$,

$$d(V_m, V_*) \leqslant c^{sm_0} \cdot d(V_p, V_*) = c^m \cdot \frac{d(V_p, V_*)}{c^p} \leqslant c^m \cdot \Big(\frac{c_2}{c}\Big)^p d(V_0, V_*).$$

Observe that $c < 1$ and $[c_2/c]^p \leqslant [c_2/c]^{m_0-1}$ is bounded by a constant independent of m and V_0. So SCF is locally (R-linearly) convergent.

On the other hand, if $\eta_{\sup,\infty} > 1$, then there exist $c > 1$ and a subsequence $\{m_i\}_{i=0}^{\infty}$ of positive integers such that $\eta_{\sup,m_i} \geqslant c$ as $i \to \infty$. Let $\delta > 0$ be a constant satisfying $c - \delta > 1$. It follows from the definition of $\eta_{\sup,m}$ that for all $\varepsilon > 0$ there exists V_0, with $d(V_0, V_*) \leqslant \varepsilon$, such that $d(V_{m_i}, V_*)/d(V_0, V_*) \geqslant (c - \delta)^{m_i}$, which is arbitrarily large as $m_i \to \infty$. Hence the iteration is locally divergent. ∎

4.2. Characterization of contraction factors. The definitions of $\eta_{\sup}$ in (4.1) and $\eta_{\sup,\infty}$ in (4.3) are generic. A meaningful characterization of $\eta_{\sup}$ and $\eta_{\sup,\infty}$ will have to involve the specific choice of the metric $d(\cdot,\cdot)$ and the detail of $H(V)$. Theorem 4.2 below contains the main contributions of this paper. It reveals a direct characterization of $\eta_{\sup}$ by $H(V)$ for a class of metrics, as compared to the previous works on the upper bounds of $\eta_{\sup}$ [5, 17, 40]. Furthermore, for differentiable $H(V)$, it provides closed-form expressions for $\eta_{\sup}$ and the true convergence factor $\eta_{\sup,\infty}$.

THEOREM 4.2. *Suppose Assumption* 1*, and let* $d(\cdot,\cdot) := \|\Theta(\cdot,\cdot)\|_{\mathrm{ui}}$.

(a) *If $H(V)$ is Lipschitz continuous at V_*, then*

$$\eta_{\sup} = \limsup_{\substack{V\in\mathbb{U}^{n\times k}\\ \|\tan\Theta(V,V_*)\|_{\rm ui}\to 0}} \frac{\|S(V)\|_{\rm ui}}{\|\tan\Theta(V,V_*)\|_{\rm ui}} \quad < \quad \infty, \tag{4.8}$$

where $S(V)$ is the unique solution of the Sylvester equation defined in (2.4).

(b) *If $H(V)$ is differentiable at V_*, then*

$$\eta_{\sup} = |||\mathscr{L}|||_{\rm ui} \geqslant \eta_{\sup,\infty} = \rho(\mathscr{L}), \tag{4.9}$$

where $\mathscr{L}$ is the local $\mathbb{R}$-linear operator of the plain SCF defined in (3.8) *and $|||\mathscr{L}|||_{\rm ui}$ is the operator norm of $\mathscr{L}$ induced by the unitarily invariant norm $\|\cdot\|_{\rm ui}$, i.e., $|||\mathscr{L}|||_{\rm ui} := \sup_{Z\neq 0}\frac{\|\mathscr{L}(Z)\|_{\rm ui}}{\|Z\|_{\rm ui}}$. Consequently, the plain SCF* (1.3) *is locally convergent to V_* with its asymptotic average convergence rate bounded by $\rho(\mathscr{L})$ if $\rho(\mathscr{L})<1$ and locally divergent at V_* if $\rho(\mathscr{L})>1$.*

Proof. For item (a), by definition (4.1) with $d(\cdot,\cdot) := \|\Theta(\cdot,\cdot)\|_{\rm ui}$, we obtain

$$\eta_{\sup} = \limsup_{\substack{V_0\in\mathbb{U}^{n\times k}\\ \|\Theta(V_0,V_*)\|_{\rm ui}\to 0}} \frac{\|\Theta(V_1,V_*)\|_{\rm ui}}{\|\Theta(V_0,V_*)\|_{\rm ui}} = \limsup_{\substack{V_0\in\mathbb{U}^{n\times k}\\ \|\tan\Theta(V_0,V_*)\|_{\rm ui}\to 0}} \frac{\|\tan\Theta(V_1,V_*)\|_{\rm ui}}{\|\tan\Theta(V_0,V_*)\|_{\rm ui}}, \tag{4.10}$$

where the second equality is a consequence of (2.9), together with $\Theta(V_1,V_*)\to 0$ as $\Theta(V_0,V_*)\to 0$ due to items (a) and (b) of Lemma 3.2. Then, a direct application of (3.6) leads to (4.8).

For the boundedness of $\eta_{\sup}<\infty$, we first observe that

$$\begin{aligned}
\|\Lambda_{*\perp}X - X\Lambda_*\|_{\rm ui} &= \|(\Lambda_{*\perp}-\lambda_1 I)X - X(\Lambda_*-\lambda_1 I)\|_{\rm ui}\\
&\geqslant \|(\Lambda_{*\perp}-\lambda_1 I)X\|_{\rm ui} - \|X(\Lambda_*-\lambda_1 I)\|_{\rm ui}\\
&\geqslant \frac{\|X\|_{\rm ui}}{\|(\Lambda_{*\perp}-\lambda_1 I)^{-1}\|_2} - \|(\Lambda_*-\lambda_1 I)\|_2\|X\|_{\rm ui}\\
&= \delta_*\|X\|_{\rm ui},
\end{aligned}$$

where the second inequality is due to the 2-norm consistency with $\|\cdot\|_{\rm ui}$ in (2.6) and the last one is by the definition of δ_* in (2.2). So by taking norms on the Sylvester equation (2.3) and noticing that $X\equiv S(V)$ and $\delta_*>0$ by Assumption 1, we have

$$\|S(V)\|_{\rm ui} \leqslant \delta_*^{-1}\|V_{*\perp}^{\rm H}[H(V_*)-H(V)]V_*\|_{\rm ui}. \tag{4.11}$$

On the other hand, it follows from the Lipschitz continuity of $H(V)$ and (3.3) that

$$\|H(V)-H(V_*)\|_{\rm ui} \leqslant \alpha\,\left(\|\tan\Theta(V,V_*)\|_{\rm ui} + \mathcal{O}(\|\tan\Theta(V,V_*)\|_{\rm ui}^2)\right)$$

for some constant $\alpha<\infty$. Combining this with (4.11) and (4.8), we conclude $\eta_{\sup}<\infty$.

For item (b), the inequality in (4.9) has already been established in (4.4), and the formula of $\eta_{\sup}$ follows directly from (4.8) and the expansion (3.7). It remains to find the expression for $\eta_{\sup,\infty}$.

For notation simplicity, we denote by $T_m := T(V_m) = (V_{*\perp}^{\rm H}V_m)(V_*^{\rm H}V_m)^{-1}$ for $m=0,1,\ldots$. It follows from Lemma 3.2 that

$$T_m = \mathscr{L}^m(T_0) + \mathbf{o}(c_m\|T_0\|_{\rm ui}),$$

where $\mathscr{L}^m = \mathscr{L} \circ \cdots \circ \mathscr{L}$ represents the composition of the linear operator $\mathscr{L}$ for m times and c_m is a constant independent of T_0. Hence for any given m

$$\eta_{\sup,m} = \limsup_{\|\Theta(V_0,V_*)\|_{\mathrm{ui}}\to 0} \left(\frac{\|\Theta(V_m,V_*)\|_{\mathrm{ui}}}{\|\Theta(V_0,V_*)\|_{\mathrm{ui}}}\right)^{1/m} = \limsup_{\|T_0\|_{\mathrm{ui}}\to 0} \left(\frac{\|T_m\|_{\mathrm{ui}}}{\|T_0\|_{\mathrm{ui}}}\right)^{1/m}$$
$$= \limsup_{T_0\to 0} \left(\frac{\|\mathscr{L}^m(T_0)\|_{\mathrm{ui}}}{\|T_0\|_{\mathrm{ui}}}\right)^{1/m},$$

where the second equation is due to (2.9), together with the continuity $T_m \to 0$ as $T_0 \to 0$, implied by (3.6). Since $\mathscr{L}$ is a finite-dimensional linear operator, we have that $\eta_{\sup,m} = (|||\mathscr{L}^m|||_{\mathrm{ui}})^{1/m}$. The expression for $\eta_{\sup,\infty}$ in (4.9) is a consequence of Gelfand's formula, which says $\lim_{m\to\infty} |||\mathscr{L}^m|||^{1/m} = \rho(\mathscr{L})$ for any operator norm $|||\cdot|||$ in a finite-dimensional vector space (see, e.g., [15, Theorem 17.4]). ∎

In recent years, a series of works has been published to improve the upper bounds of the local contraction factor $\eta_{\sup}$, e.g., [5, 17, 40]. Those bounds were typically established for particular choices of the metric $d(\cdot,\cdot)$ between subspaces and for a class of $H(V)$. Let us revisit particularly the following estimation of the convergence factor of the plain SCF iteration presented recently in [5]:

$$\eta_{\mathrm{czbl}} := \limsup_{\substack{V\in\mathbb{U}^{n\times k}\\ \|\sin\Theta(V,V_*)\|_{\mathrm{ui}}\to 0}} \frac{\delta_*^{-1}\|V_{*\perp}^{\mathrm{H}}[H(V_*)-H(V)]V_*\|_{\mathrm{ui}}}{\|\sin\Theta(V,V_*)\|_{\mathrm{ui}}}. \tag{4.12}$$

We use η_{czbl} as a baseline for comparison since it improves the previous estimates of the convergence factors presented in [17, 40].

For a differentiable $H(V)$ with the expansion (3.4), η_{czbl} can be expressed as

$$\eta_{\mathrm{czbl}} = \delta_*^{-1}\cdot |||\mathscr{L}_{\mathrm{czbl}}|||_{\mathrm{ui}}, \tag{4.13}$$

where $\mathscr{L}_{\mathrm{czbl}} : \mathbb{C}^{(n-k)\times k} \to \mathbb{C}^{(n-k)\times k}$ is an $\mathbb{R}$-linear operator:

$$\mathscr{L}_{\mathrm{czbl}}(Z) = V_{*\perp}^{\mathrm{H}}\,\mathbf{D}H(V_*)[V_{*\perp}Z]\,V_*. \tag{4.14}$$

The convergence factor η_{czbl} in (4.12) significantly improves several previously established results in [17, 40]. However, it follows from the characterization of $\eta_{\sup}$ in (4.8) and the bound of $S(V)$ in (4.11) that

$$\eta_{\sup} \leqslant \eta_{\mathrm{czbl}}. \tag{4.15}$$

Therefore, the quantity η_{czbl} is an upper bound of $\eta_{\sup}$ and can substantially overestimate the convergence rate of SCF in practice; see numerical examples in section 6.

We have already seen from Lemma 4.1 that $\eta_{\sup,\infty}$ is the true convergence factor for SCF and $\eta_{\sup,\infty} \leqslant \eta_{\sup}$. To see why there may be a difference between $\eta_{\sup,\infty}$ and $\eta_{\sup}$ in (4.9), we take a look at $\eta_{\sup}$ in the commonly used Frobenius norm:

$$|||\mathscr{L}|||_{\mathrm{F}} := \sup_{Z\neq 0}\frac{\|\mathscr{L}(Z)\|_{\mathrm{F}}}{\|Z\|_{\mathrm{F}}} = \sup_{Z\neq 0}\frac{\langle\mathscr{L}(Z),\mathscr{L}(Z)\rangle^{1/2}}{\langle Z,Z\rangle^{1/2}}$$
$$= \sup_{Z\neq 0}\frac{\langle Z,\mathscr{L}^*\circ\mathscr{L}(Z)\rangle^{1/2}}{\langle Z,Z\rangle^{1/2}} = \left(\lambda_{\max}(\mathscr{L}^*\circ\mathscr{L})\right)^{1/2}, \tag{4.16}$$

where $\langle X,Y\rangle = \Re(\mathrm{tr}(X^{\mathrm{H}}Y))$ denotes the inner product on $\mathbb{C}^{(n-k)\times k}$ (ℝ) and $\mathscr{L}^*$ is the adjoint of $\mathscr{L}$. It follows from (4.9) and (4.16) that for the Frobenius norm

$$\eta_{\sup} = |\lambda_{\max}(\mathscr{L}^*\circ\mathscr{L})|^{1/2} \geqslant |\rho(\mathscr{L})| = \eta_{\sup,\infty}. \tag{4.17}$$

By the standard matrix analysis, the equality in (4.17) holds if $\mathscr{L}$ is a *normal* linear operator on $\mathbb{C}^N(\mathbb{R})$, and the difference between the two numbers can be arbitrarily large when $\mathscr{L}$ is far from normal. For practical NEPvs, such as the ones in section 6, we have observed that $\mathscr{L}$ is usually a slightly nonnormal operator, causing a small difference between the two contraction factors.

Finally, we comment on another recent work [38] on the local convergence analysis of SCF using the spectral radius. In [38], SCF is viewed as a fixed-point iteration $P_{m+1} = \psi(P_m)$ in the density matrix $P_m = V_m V_m^{\mathrm{H}} \in \mathbb{C}^{n\times n}$, rather than in V_m directly. The authors showed that the fixed-point mapping $\psi(P)$ has a closed-form Jacobian supermatrix J, assuming $H(V)$ is a linear function in $P = VV^{\mathrm{H}}$. So the spectral radius of J also provides a convergence criterion. Since P has $p = (n+1)n/2$ free variables, the corresponding supermatrix J is of size p-by-p. This is in contrast to the $\mathbb{R}$-linear operator $\mathscr{L}$ (3.8) in tangent-angle matrices, which is only of size q-by-q with $q = 2(n-k)k = \mathcal{O}(p^{1/2})$ for modest k. In addition to the reduced size, the use of a linear operator, rather than a supermatrix, allows for more convenient computation of the spectral radius in practice; see subsection 6.1. Furthermore, $\mathscr{L}$ is also easier to work with theoretically and numerically, thanks to its simplicity in formulation and more explicit dependencies on key variables, such as derivatives and eigenvalue gaps. In the next section, we will show how to apply the spectral radius $\rho(\mathscr{L})$ to analyze the so-called *level-shifting scheme* for stabilizing and accelerating the plain SCF iteration.

5. Level-shifted SCF iteration. In the previous section, we have discussed that if the spectral radius $\rho(\mathscr{L}) > 1$ (or more generally $\eta_{\sup,\infty} > 1$ in the case when $H(V)$ is just continuous), then the plain SCF (1.3) is locally divergent at V_*. Even if $\rho(\mathscr{L}) < 1$, the process is prone to slow convergence or oscillation before reaching local convergence. To address those issues, the plain SCF may be applied in practice with some stabilizing schemes to help with convergence. Among the most popular choices is the level-shifting strategy initially developed in computation chemistry [30, 36, 41]. In this section, we discuss why such a scheme works through the lens of spectral radius when $H(V)$ is differentiable.

5.1. Level-shifting. The level-shifting scheme modifies the plain SCF (1.3) with a parameter σ as follows:

$$[H(V_i) - \sigma\, V_i V_i^{\mathrm{H}}]V_{i+1} = V_{i+1}\Lambda_{i+1} \quad \text{for} \quad i = 0, 1, 2, \ldots, \tag{5.1}$$

where V_{i+1} is an orthonormal basis matrix of the invariant subspace associated with the k smallest eigenvalues of the matrix $H(V_i) - \sigma\, V_i V_i^{\mathrm{H}}$. It can be viewed simply as the plain SCF (1.3) applied to the level-shifted NEPv:

$$H_\sigma(V)V = V\Lambda \quad \text{with} \quad H_\sigma(V) := H(V) - \sigma VV^{\mathrm{H}}. \tag{5.2}$$

Note that $H_\sigma(V)$ is again unitarily invariant as in (1.2). The level-shifting transformation does not alter the solutions of the original NEPv (1.1) but shifts related eigenvalues of $H(V)$ by σ:

$$H(V)V = V\Lambda \qquad \Longleftrightarrow \qquad H_\sigma(V)V = V(\Lambda - \sigma I_k).$$

Hence, if (V_*, Λ_*) is a solution of the original NEPv (1.1), then $(V_*, \Lambda_* - \sigma I_k)$ will solve the level-shifted NEPv (5.2). In the following discussion, we assume the parameter σ is a constant for convenience. In practice, it can change iteration-by-iteration.

One direct consequence of the level-shifting transformation is that it enlarges the eigenvalue gap at the solution V_*. By the eigendecomposition (2.1), we obtain

$$H_\sigma(V_*)\,[V_*, V_{*\perp}] = [V_*, V_{*\perp}] \begin{bmatrix} \Lambda_* - \sigma I_k & \\ & \Lambda_{*\perp} \end{bmatrix}. \tag{5.3}$$

Recall that $\Lambda_* = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ and $\Lambda_{*\perp} = \mathrm{diag}(\lambda_{k+1}, \ldots, \lambda_n)$ consist of the ordered eigenvalues of $H(V_*)$ as in (2.1). Therefore, the gap between the kth and $(k+1)$st eigenvalue of $H_\sigma(V_*)$ becomes

$$\delta_{\sigma *} := \lambda_{k+1} - (\lambda_k - \sigma) = \delta_* + \sigma, \tag{5.4}$$

where δ_* denotes the eigenvalue gap (2.2) of the original NEPv (1.1) at V_*. So the level-shifted NEPv (5.2) always has a larger eigenvalue gap $\delta_{\sigma *}$ if $\sigma > 0$.

For the standard Hermitian eigenvalue problem, it is well known that the larger the eigenvalue gap between the desired eigenvalues and the others, the easier and more robust it will become to compute the desired eigenvalues and the associated eigenspace [8, 24, 34]. Therefore, it is desirable to have a large eigenvalue gap $\delta_{\sigma *}$ for the sequence of matrix eigenvalue problems in the SCF iteration (5.1), but on the other hand if the shift σ is too large, it will negatively affect the local convergence rate of SCF as numerical evidences suggest. Presently, there are heuristic schemes to choose the level-shift parameter σ in practice; see, e.g., [41]. However, those heuristics cannot explain how the convergence behavior of SCF (5.1) is affected by the level-shifting parameter σ.

We mention that the conventional restriction of $\sigma > 0$ for the level-shifting parameter [30, 36, 41] is not necessary. From the eigendecomposition (5.3) we see that the eigenvector matrix V_* always corresponds to the k smallest eigenvalues of $H_\sigma(V_*)$ so long as $\sigma \in (-\delta_*, +\infty)$.

5.2. Local convergence of level-shifted SCF. In what follows, we investigate the local convergence behavior of the level-shifting scheme by examining the spectral radius $\rho(\mathscr{L}_\sigma)$ for the local $\mathbb{R}$-linear operator $\mathscr{L}_\sigma$ of the level-shifted SCF (5.1). We will focus on a class of NEPv where certain conditions on the derivatives of $H(V)$ are satisfied. Those conditions hold for NEPvs arising in optimization problems with orthogonality constraints, as is usually the case for most practical NEPvs.

5.2.1. NEPvs from optimization with orthogonality constraints. Let us review a class of NEPvs arising from the following optimization problems with orthogonality constraints

$$\min_{V \in \mathbb{C}^{n\times k}} E(V) \quad \text{s.t.} \quad V^{\mathrm{H}} V = I_k, \tag{5.5}$$

where E is some energy function satisfying $\nabla E(V) = H(V)V$ (see, e.g., [3, 41, 42]). We will make no assumption on the specific form of $E(\cdot)$ to be used. For the constrained optimization problem (5.5), the associated Lagrangian function is given by

$$L(V) := E(V) + \frac{1}{2}\,\mathrm{tr}\left(\Lambda^{\mathrm{H}}(V^{\mathrm{H}}V - I_k)\right),$$

where $\Lambda = \Lambda^{\mathrm{H}}$ is the k-by-k matrix of Lagrange multipliers. We have suppressed L's dependency on Λ for notation simplicity. The first-order optimization condition $\nabla_V L(V) = H(V)V - V\Lambda = 0$ leads immediately to NEPv (1.1).

Because the target solution V_* of interest is also a minimizer of (5.5), it needs to satisfy certain second-order conditions as well. Assuming $E(V)$ is also second-order differentiable, by straightforward derivation, the Hessian operator of $L(V)$ is given by

$$\nabla_V^2 L(V_*)[X] = H(V_*)X + (\mathbf{D}H(V_*)[X])\, V_* - X\Lambda_*,$$

where X denotes the direction for the evaluation and $\mathbf{D}H(V_*)[\cdot]$ denotes the directional derivative of H as defined in (2.12). Then by the standard second-order optimization condition [23], this operator needs to be at least positive semidefinite when restricted to $X = V_{*\perp}Z$ for all $Z \in \mathbb{C}^{(n-k)\times k}$; namely, within the tangent space of the feasible set $V^{\mathrm{H}}V = I_k$ at V_*, the operator

$$V_{*\perp}^{\mathrm{H}}\left(\nabla_V^2 L(V_*)[V_{*\perp}Z]\right) = V_{*\perp}^{\mathrm{H}}\,\mathbf{D}H(V_*)[V_{*\perp}Z]\, V_* + \Lambda_{*\perp}Z - Z\Lambda_*$$

is self-adjoint and at least positive semidefinite.

In general, NEPv (1.1) may or may not be associated with an optimization problem like (5.5). The discussion above nonetheless can still motivate us to introduce the following $\mathbb{R}$-linear operator $\mathscr{Q}\colon \mathbb{C}^{(n-k)\times k} \to \mathbb{C}^{(n-k)\times k}$:

$$\mathscr{Q}(Z) := V_{*\perp}^{\mathrm{H}}\,\mathbf{D}H(V_*)[V_{*\perp}Z]\, V_* + \Lambda_{*\perp}Z - Z\Lambda_*. \tag{5.6}$$

It is well defined without the need of an associated optimization problem (5.5) so long as $H(V)$ is differentiable at V_* with respect to the real and imaginary parts of V. We call $\mathscr{Q}$ a *restricted derivative operator* of NEPv (1.1). In this more general situation, there is no implied self-adjointness as a result of being the Hessian of $L(V)$, however, not to mention that it is positive definite. For that reason, we need to make the following assumption.

Assumption 2. The linear operator $\mathscr{Q}$ is self-adjoint and positive definite with respect to the standard inner product on $\mathbb{C}^{(n-k)\times k}$, i.e.,

$$\Re(\mathrm{tr}(Z^{\mathrm{H}}\mathscr{Q}(Z))) = \Re(\mathrm{tr}([\mathscr{Q}(Z)]^{\mathrm{H}}Z)) \quad \text{and} \quad \Re(\mathrm{tr}(Z^{\mathrm{H}}\mathscr{Q}(Z))) > 0 \text{ for all } Z \neq 0.$$

5.2.2. Spectral radius of level-shifted local $\mathbb{R}$-linear operator. We can immediately draw from Lemma 3.2 and Theorem 4.2 a conclusion that the local convergence behavior of the level-shifted SCF (5.1) is characterized by the local $\mathbb{R}$-linear operator corresponding to the level-shifted NEPv (5.2). To show the dependency on σ, we note that the local $\mathbb{R}$-linear operator associated with $H_\sigma(\,\cdot\,)$, as defined in Lemma 3.2, is

$$\mathscr{L}_\sigma(Z) = D_\sigma(V_*) \odot (V_{*\perp}^{\mathrm{H}}\,\mathbf{D}H_\sigma(V_*)[V_{*\perp}Z]\, V_*), \tag{5.7}$$

where $D_\sigma(V_*) \in \mathbb{R}^{(n-k)\times k}$ has elements

$$D_\sigma(V_*)_{(i,j)} = (\lambda_{k+i}(H(V_*)) - \lambda_j(H(V_*)) + \sigma)^{-1}.$$

A representation of $\mathscr{L}_\sigma$ in terms of restricted derivative operator $\mathscr{Q}$ and a bound of the spectral radius of $\mathscr{L}_\sigma$ is given in the following theorem.

THEOREM 5.1. *Suppose Assumptions* 1, *and* 2 *and that* $\sigma \in (-\delta_*, +\infty)$. *The local* $\mathbb{R}$*-linear operator* $\mathscr{L}_\sigma(\cdot)$ *of the level-shifted SCF* (5.1) *for the level-shifted NEPv* (5.2) *is given by*

$$\mathscr{L}_\sigma(\cdot) = D_\sigma(V_*) \odot \mathscr{Q}(\cdot) - I_{\mathrm{id}}, \tag{5.8}$$

where $\mathscr{Q}$ is the restricted derivative operator defined in (5.6) *and* I_{id} *denotes the identity operator on the vector space* $\mathbb{C}^{(n-k)\times k}(\mathbb{R})$. *Moreover, the spectral radius of* $\mathscr{L}_\sigma$ *is bounded:*

$$\rho(\mathscr{L}_\sigma) \leqslant \max\left\{\left|\frac{\mu_{\max}}{\sigma+\delta_*}-1\right|, \left|\frac{\mu_{\min}}{\sigma+s_*}-1\right|\right\}, \tag{5.9}$$

where $\mu_{\max} \geqslant \mu_{\min} > 0$ *denote the largest and smallest eigenvalues of the* $\mathbb{R}$*-linear operator* $\mathscr{Q}$ *and* δ_* *and* s_* *are the spectral gap and span, respectively, i.e.,*

$$\delta_* = \lambda_{k+1}(H(V_*)) - \lambda_k(H(V_*)) \quad \textit{and} \quad s_* = \lambda_n(H(V_*)) - \lambda_1(H(V_*)).$$

In particular, $\rho(\mathscr{L}_\sigma) < 1$ *if*

$$\sigma > \frac{\mu_{\max}}{2} - \delta_*. \tag{5.10}$$

Proof. By the definition of $H_\sigma(V)$ in (5.2) and the derivative operator (2.12), it holds that

$$\mathbf{D}H_\sigma(V_*)[X] = \mathbf{D}H(V_*)[X] - \sigma\mathbf{D}(V_*V_*^{\mathrm{H}})[X] = \mathbf{D}H(V_*)[X] - \sigma(V_*X^{\mathrm{H}} + XV_*^{\mathrm{H}}).$$

Hence

$$\begin{aligned} V_{*\perp}^{\mathrm{H}}\,\mathbf{D}H_\sigma(V_*)[V_{*\perp}Z]\,V_* &= V_{*\perp}^{\mathrm{H}}\,\mathbf{D}H(V_*)[V_{*\perp}Z]\,V_* - \sigma\, Z \\ &= \mathscr{Q}(Z) + Z(\Lambda_* - \sigma I_k) - \Lambda_{*\perp}Z = \mathscr{Q}(Z) - Z \oslash D_\sigma(V_*), \end{aligned} \tag{5.11}$$

where the second equation is by (5.6) and $\oslash$ denotes the elementwise division. Plug (5.11) into (5.7) to obtain

$$\mathscr{L}_\sigma(Z) = D_\sigma(V_*) \odot [\mathscr{Q}(Z) - Z \oslash D_\sigma(V_*)] = D_\sigma(V_*) \odot \mathscr{Q}(Z) - Z.$$

This proves (5.8).

The vector space $\mathbb{C}^{(n-k)\times k}(\mathbb{R})$ has a natural basis $\mathcal{B} := \{E_{ij},\, \imath E_{ij} : i = 1, \ldots, n-k,\ j = 1, \ldots, k\}$, where the entries of $E_{ij} \in \mathbb{R}^{(n-k)\times k}$ are all zeros but 1 is its (i,j)th entry. Let $\mathcal{L}_\sigma, \mathcal{D}_\sigma, \mathcal{Q} \in \mathbb{R}^{2N\times 2N}$ be the matrix representations of the operators $\mathscr{L}_\sigma(\cdot)$, $D_\sigma(V_*)\odot(\cdot)$, and $\mathscr{Q}(\cdot)$ with respect to the basis $\mathcal{B}$, respectively, where $N = (n-k)\times k$. It follows from (5.8) that

$$\mathcal{L}_\sigma = \mathcal{D}_\sigma\mathcal{Q} - I_{2N}.$$

Observe that $\mathcal{D}_\sigma$ is a diagonal matrix consisting of elements of D_σ, and $\mathcal{Q}$ is symmetric positive definite due to Assumption 2. Hence the eigenvalues of $\mathcal{D}_\sigma\mathcal{Q}$ are all positive, and

$$\rho(\mathscr{L}_\sigma) = \max\{|\lambda_{\max}(\mathcal{D}_\sigma\mathcal{Q}) - 1|, |\lambda_{\min}(\mathcal{D}_\sigma\mathcal{Q}) - 1|\}. \tag{5.12}$$

Since the eigenvalues of $\mathcal{D}_\sigma\mathcal{Q}$ are the same as those of $\mathcal{Q}^{1/2}\mathcal{D}_\sigma\mathcal{Q}^{1/2}$ and

$$\lambda_{\max}(\mathcal{D}_\sigma)\mathcal{Q} \succeq \mathcal{Q}^{1/2}\mathcal{D}_\sigma\mathcal{Q}^{1/2} \succeq \lambda_{\min}(\mathcal{D}_\sigma)\mathcal{Q},$$

we have

$$\lambda_{\max}(\mathcal{D}_\sigma\mathcal{Q}) \leqslant \mu_{\max}/(\sigma+\delta_*) \quad \text{and} \quad \lambda_{\min}(\mathcal{D}_\sigma\mathcal{Q}) \geqslant \mu_{\min}/(\sigma+s_*). \tag{5.13}$$

Inequality (5.9) is now a simple consequence of (5.12).

It follows immediately from (5.9) that

$$\rho(\mathscr{L}_\sigma) < 1 \quad \text{if} \quad 0 < \frac{\mu_{\min}}{\sigma+s_*} \leqslant \frac{\mu_{\max}}{\sigma+\delta_*} < 2,$$

i.e., $\sigma > \mu_{\max}/2 - \delta_*$. ∎

As an immediate consequence of Theorem 5.1, the level-shifted SCF is locally convergent for a sufficiently large σ! In fact, $\sigma > \mu_{\max}/2 - \delta_*$ guarantees $\rho_\sigma(\mathscr{L}) < 1$, although the latter may hold for much smaller σ than $\mu_{\max}/2 - \delta_*$. This is what we can prove without further detailed information. On the other hand, it follows from (5.13) that $\mathcal{D}_\sigma \mathcal{Q} \to 0$ as $\sigma \to +\infty$. Hence by (5.12) we have $\rho_\sigma(\mathscr{L}) \to 1$ as $\sigma \to +\infty$, implying slow convergence of the level-shifted SCF for σ that is too large. The question is how to pick a decent σ with fairly small $\rho_\sigma(\mathscr{L})$. In general, this is not an easy task because $\rho_\sigma(\mathscr{L})$ is usually unknown. One possible compromise is to minimize the upper bound of $\rho_\sigma(\mathscr{L})$ in (5.9), assuming good estimates to $\mu_{\min}$, $\mu_{\max}$, δ_*, and s_* are available. In fact, the minimizer of the upper bound is achieved when the two terms in the right-hand side of (5.9) coincide, which happens only if

$$\frac{\mu_{\max}}{\sigma + \delta_*} - 1 = 1 - \frac{\mu_{\min}}{\sigma + s_*},$$

due to $\sigma \in (-\delta_*, +\infty)$. This equation has a unique solution $\sigma_* \in (\mu_{\max}/2 - \delta_*, +\infty)$. We caution the reader that this σ_* can be far from the one that minimizes the actual $\rho_\sigma(\mathscr{L})$. It is just the best possible choice we can get with the limited information at hand. The true optimal σ, however, can be even smaller than $\mu_{\max}/2 - \delta_*$, as will be illustrated by numerical examples in section 6. In any case, the operator $\mathscr{L}_\sigma$ and its spectral radius provide us with a deep understanding of level-shifting strategy and an approach to seek a decent choice of the level-shifting parameter σ.

To end this section, we note that the results in this section are consistent with, and also complement, the convergence analysis of the level-shifted methods applied to Hartree–Fock equations [7]. Using optimization approaches, the authors showed that a sufficiently large shift σ can lead to global convergence. The condition (5.10), on the other hand, provided a closed-form lower bound on the size of σ needed to achieve local convergence. The bound of (5.10) involves the quantities δ_* and $\mu_{\max}$ defined by the exact solution V_* and is mostly of theoretical interest. For some applications, it is possible to have a priori estimates of δ_* and $\mu_{\max}$, as demonstrated in the examples in the next section.

6. Numerical examples. In this section, we present numerical examples to demonstrate the sharpness of the convergence rate estimates established in the previous sections. Specifically, the purpose of the examples is twofold: Firstly, to illustrate how these convergence results are manifested in practice, where various convergence rate estimates are compared and their sharpness in estimating the actual convergence rate is demonstrated. Secondly, to investigate and gain insight into the influence of the level-shifting parameter σ on the convergence rate of SCF (5.1).

6.1. Experiment setup. We will perform two case studies: one is a discrete Kohn–Sham equation with real coefficient matrices $H(V)$, and the other is a discrete Gross–Pitaevskii equation with complex matrices.

All our experiments are implemented and conducted in MATLAB 2019. In each simulation, the "exact" solution V_* is computed by the plain SCF (1.3), when it is convergent, to achieve a residual tolerance $\|H(V_*)V_* - V_*\Lambda_*\|_2 \leqslant 10^{-14}$. When the plain SCF failed to converge, V_* is computed by the level-shifted SCF (5.1) with a properly chosen shift σ, also to the same level of accuracy.

The convergence rate estimates to be investigated include

(i) η_{czbl} by [5], computed as (4.13) in the Frobenius norm,
(ii) $\eta_{\sup} = \|\!|\mathscr{L}|\!\|_{\text{F}}$ in (4.9) in the Frobenius norm, and
(iii) $\eta_{\sup,\infty} = \rho(\mathscr{L})$ in (4.9).

These convergence rate estimates will be compared against the *observed convergence rate* of SCF, computed from the convergence history of residual norms $\|H(V_i)V_i - V_i\Lambda_i\|_2$ of SCF iteration by the least squares approximation on the last few iterations (with residual norms below 10^{-8}).

Evaluation of $\eta_{\sup,\infty}(=\rho(\mathscr{L}))$. Although a matrix representation $\mathbf{L}$ is involved in the definition (2.11), its explicit formulation is not needed for computing $\rho(\mathscr{L})$. Recall that $\mathscr{L}\colon \mathbb{C}^{p\times k} \to \mathbb{C}^{p\times k}$ is an $\mathbb{R}$-linear operator. By viewing a complex matrix $X = X_{\mathbf{r}} + \imath X_{\mathbf{i}} \in \mathbb{C}^{p\times k}$ as a pair of real matrices $(X_{\mathbf{r}}, X_{\mathbf{i}})$ consisting of the real and imaginary parts, we express $\mathscr{L}$ as a linear operator $\widehat{\mathscr{L}}\colon \mathbb{R}^{p\times k}\times\mathbb{R}^{p\times k} \to \mathbb{R}^{p\times k}\times\mathbb{R}^{p\times k}$,

$$\widehat{\mathscr{L}}(X_{\mathbf{r}}, X_{\mathbf{i}}) = (\Re(\mathscr{L}(X)), \Im(\mathscr{L}(X))). \tag{6.1}$$

The input (as well as the output) matrix pair $(X_{\mathbf{r}}, X_{\mathbf{i}})$ can be regarded as a real "vector" of length $2N$. The largest eigenvalue in magnitude of the linear operator $\widehat{\mathscr{L}}$ can be computed conveniently by the MATLAB `eigs` function as follows:

```
v2m = @(x) reshape(x(1:N)+1i*x(N+1:end), p, []); % real vec x -> mat X
m2v = @(X) [real(X(:)); imag(X(:))];             % mat X -> real vec x
hatL = @(x) m2v(L(v2m(x)));                      % operator hat L
lam_max = eigs(hatL, 2*N, 1);                    % largest eigenval.
```

Evaluation of $\eta_{\sup}$ *and* η_{czbl}. The induced norm $|||\mathscr{L}|||_{\mathrm{F}}$ in (4.16) is defined as the square root of the largest eigenvalue of $\mathscr{L}^*\circ\mathscr{L}$, which is also an $\mathbb{R}$-linear operator. We can use exactly the same approach above to obtain $\lambda_{\max}(\mathscr{L}^*\circ\mathscr{L})$. Since the operator $\mathscr{L}^*\circ\mathscr{L}$ is self-adjoint, the largest eigenvalue is always a real number. In analogy, for η_{czbl} in (4.13), $|||\mathscr{L}_{\mathrm{czbl}}|||_{\mathrm{F}}$ can be computed as the square root of $\lambda_{\max}(\mathscr{L}^*_{\mathrm{czbl}}\circ\mathscr{L}_{\mathrm{czbl}})$.

6.2. Single particle Hamiltonian. Let us consider an NEPv (1.1) with a real coefficient matrix-valued function

$$H(V) = L + \alpha\ \mathrm{Diag}(L^{-1}\,\mathrm{diag}(VV^{\mathrm{T}})), \tag{6.2}$$

where tridiagonal matrix $L = \mathrm{tridiag}(-1, 2, -1) \in \mathbb{R}^{n\times n}$ is a discrete 1D Laplacian, $\alpha > 0$ is a given parameter, and $V \in \mathbb{O}^{n\times k} := \{X \in \mathbb{R}^{n\times k} : X^{\mathrm{T}}X = I_k\}$. $H(V)$ is known as the single-particle Hamiltonian arising from discretizing a 1D Kohn–Sham equation in electronic structure calculations and has become a standard testing problem for investigating the convergence of SCF due to its simplicity; see, e.g., [5, 17, 40, 44]. $H(V)$ is differentiable. By a straightforward calculation, the directional derivative operator $\mathbf{D}H(V_*)$ defined in (2.12) is given by

$$\mathbf{D}H(V)[X] = 2\alpha\ \mathrm{Diag}(L^{-1}\,\mathrm{diag}(XV^{\mathrm{T}})),$$

which is linear in X.

The local $\mathbb{R}$-linear operator $\mathscr{L}$ in (3.8) of the plain SCF (1.3) is given by

$$\mathscr{L}(Z) = 2\alpha\, D(V_*)\odot\left(V_{*\perp}^{\mathrm{T}}\ \mathrm{Diag}(L^{-1}\,\mathrm{diag}(V_{*\perp}ZV_*^{\mathrm{T}}))\,V_*\right). \tag{6.3}$$

The adjoint operator $\mathscr{L}^*$ is given by

$$\mathscr{L}^*(Y) = 2\alpha\, V_{*\perp}^{\mathrm{T}}\ \mathrm{Diag}\left(L^{-\mathrm{T}}\,\mathrm{diag}\left(V_{*\perp}(D(V_*)\odot Y)V_*^{\mathrm{T}}\right)\right)\ V_*; \tag{6.4}$$

see Appendix A for the derivation.

The local $\mathbb{R}$-linear operator $\mathscr{L}_\sigma$ (5.7) of the level-shifted SCF (5.1) is given by

$$\mathscr{L}_\sigma(Z) = D_\sigma(V_*)\odot\mathscr{Q}(Z) - I_{\mathrm{id}}, \tag{6.5}$$

where $\mathscr{Q}$ is the restricted derivative operator, which is defined in (5.6) and is given by

$$\mathscr{Q}(Z) = 2\alpha\, V_{*\perp}^{\mathrm{T}}\, \mathrm{Diag}(L^{-1}\,\mathrm{diag}(V_{*\perp} Z V_*^{\mathrm{T}}))\, V_* + (\Lambda_{*\perp} Z - Z\Lambda_*). \tag{6.6}$$

The largest eigenvalue $\mu_{\max}$ of $\mathscr{Q}$ can be bounded as follows: let $Z \in \mathbb{R}^{(n-k)\times k}$ be the corresponding eigenvector of $\mu_{\max}$; then

$$\begin{aligned}\mu_{\max} = \frac{\|\mathscr{Q}(Z)\|_{\mathrm{F}}}{\|Z\|_{\mathrm{F}}} &\leqslant 2\alpha\, \frac{\|\,\mathrm{Diag}(L^{-1}\,\mathrm{diag}(V_{*\perp} Z V_*^{\mathrm{T}}))\|_{\mathrm{F}}}{\|Z\|_{\mathrm{F}}} + s_* \\ &\leqslant 2\alpha\,\|L^{-1}\|_2 + s_* \leqslant 3\alpha\,\|L^{-1}\|_2 + 4,\end{aligned}$$

where s_* is the spectral span of $H(V_*)$, and for the last inequality we have used the inequalities $s_* \leqslant \lambda_n(H(V)) \leqslant \|L\|_2 + \alpha\|L^{-1}\|_2$ due to (6.2), and $\|L\|_2 \leqslant 4$.

Recalling the lower bound in (5.10) for the level-shifting parameter σ, we find

$$\sigma \geqslant \frac{3}{2}\,\alpha\,\|L^{-1}\|_2 + 2 \geqslant \frac{\mu_{\max}}{2} > \frac{\mu_{\max}}{2} - \delta_* \tag{6.7}$$

is sufficient to ensure local convergence of SCF (5.1). The first inequality provides an a priori lower bound on the shift σ. In practice, this crude bound is a bit pessimistic though. But it does reveal two key contributing factors—the parameter α and size n of the problem due to the fact that $\|L^{-1}\|_2 = 2^{-1}(1-\cos(\frac{\pi}{n+1}))^{-1} = \mathcal{O}(n^2)$ for the 1D Laplacian [9, Lemma 6.1]—that tend to negatively affect the size of shift.

Example 6.1. In this example, we compare the sharpness of the three convergence rate estimates of the plain SCF. We take $n = 10$ and $k = 2$ and use different α ranging from 0 to 1 in the Hamiltonian (6.2). For each run of SCF, the starting vectors are set to be the basis of the k smallest eigenvalues of L. The results are shown in Figure 1. A few observations are summarized as follows:

(a) For $\alpha = 0$, the NEPv reduces to a standard eigenvalue problem $LV = V\Lambda$, for which SCF converges in one iterative step. As α increases, SCF faces increasing challenges to converge. In particular, for α larger than 0.85 (e.g., $\alpha = 0.9$ in Figure 1), the plain SCF becomes divergent. For those α, the "exact" solutions V_* used to calculate convergence factors are computed by the level-shifted SCF.

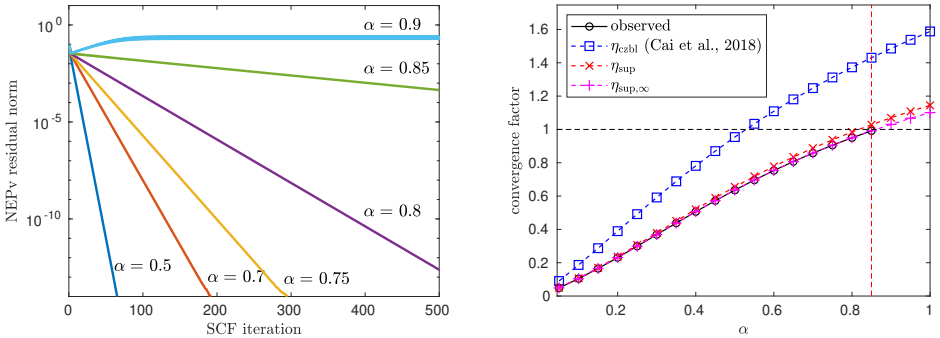

FIG. 1. *Example* 6.1*: convergence history of residual norm* $\|H(V_i)V_i - V_i\Lambda_i\|_2$ *by the plain SCF* (1.3) *for selected* α *(left plot) and convergence rate estimates as* α *varies (right plot). The observed rate marked by "circle" and the theoretic rate* $\eta_{\mathrm{sup},\infty}$ *marked by "+" coincide perfectly.*

(b) The right plot in Figure 1 shows that the asymptotic average contraction factor $\eta_{\text{sup},\infty}(=\rho(\mathscr{L}))$ successfully predicts the convergence behavior of SCF in all cases, as $\eta_{\text{sup},\infty}$ marked by "+" in the plot perfectly coincides with the observed rate marked by "circle." It can also be observed from the left plot that SCF iterations quickly enter into the region of linear convergence, and the factor $\eta_{\text{sup},\infty}$ yields excellent estimation after only a small number of iterative steps.

(c) The contraction factor estimate η_{sup} is an overestimate and provides a good prediction of local convergence for small α. It fails slightly at $\alpha = 0.85$, where, up to 10 digits,

$$\begin{aligned} \text{observed} &= 0.9913931781, & \eta_{\text{sup},\infty} &= 0.9913931591, \\ \eta_{\text{sup}} &= 1.028434776, & \eta_{\text{czbl}} &= 1.430511920. \end{aligned}$$

The difference between $\eta_{\text{sup},\infty}$ and η_{sup} implies $\mathscr{L}$ is a nonnormal operator as discussed in subsection 4.2.

(d) In comparison, the estimate η_{czbl} by [5] is less accurate. In particular, it fails to correctly indicate the convergence of the plain SCF starting at $\alpha = 0.55$, which is in contrast to η_{sup} starting at 0.85.

Example 6.2. In this example, we examine the convergence of the level-shifted SCF (5.1) with respect to the shift σ. The testing problem is the same as Example 6.1 but with a fixed $\alpha = 1$, for which the plain SCF (1.3) is divergent. We apply the level-shifted SCF with various choices of σ. The convergence history and the corresponding spectral radius of the operator $\mathscr{L}_\sigma$ in (5.7) are depicted in Figure 2.

From the spectral radius plot on the right side of Figure 2, we observe that $\rho(\mathscr{L}_\sigma)$ dropped quickly below 1. The minimal value $\rho(\mathscr{L}_\sigma) \approx 0.33$ at $\sigma \approx 0.36$ and leads to rapid convergence of SCF as shown in the left plot. As σ grows, $\rho(\mathscr{L}_\sigma)$ monotonically increases towards 1. Such a behavior of $\rho(\mathscr{L}_\sigma)$ is consistent with the bound obtained in Theorem 5.1, governed by rational functions in the form of $|1 - a/(\sigma + b)|$ with $a, b > 0$.

The sharp turning of the curve of $\rho(\mathscr{L}_\sigma)$ reveals the challenge in finding the optimal σ. The values of spectral radius grows quickly as α moves away from the

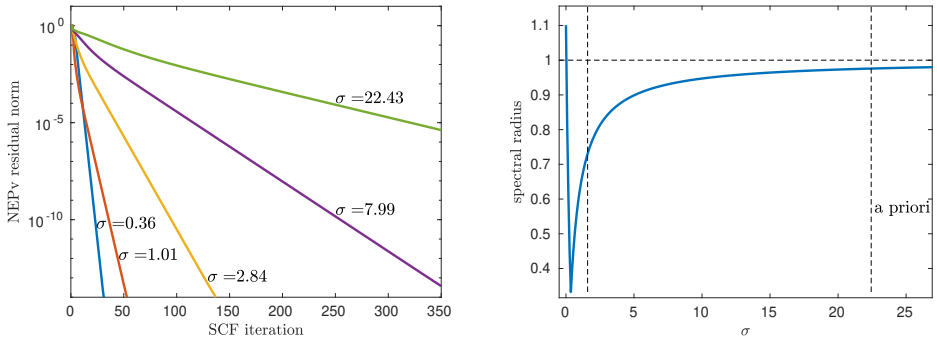

FIG. 2. *Example* 6.2*: convergence history of residual norm* $\|H(V_i)V_i - V_i\Lambda_i\|_2$ *by the level-shifted SCF* (5.1) *with selected* σ *(left plot) and spectral radius* $\rho(\mathscr{L}_\sigma)$ *as shift* σ *varies (right plot). The first vertical dash line is* $\sigma = \mu_{\max}/2 - \delta_*$ *as in* (5.10)*, the theoretical bound beyond which provably* $\rho(\mathscr{L}_\sigma) < 1$*, and the second one is a priori* $\sigma = \frac{3}{2}\alpha\|L^{-1}\|_2 + 2$ *suggested by* (6.7)*, while the optimal shift is* $\sigma \approx 0.36$*, which is smaller than* $\mu_{\max}/2 - \delta_*$*, the theoretical bound.* $H(V)$ *is given by* (6.2) *with* $\alpha = 1$.

optimal shift. We note that both the theoretic lower bound in (5.10) and a priori estimate (6.7) fall correctly into the convergence region. The a priori bound provided a pessimistic estimate of σ that leads to a less satisfactory convergence rate of the level-shifted SCF (5.1) than others.

6.3. Gross–Pitaevskii equation. In this experiment, we consider NEPvs with complex coefficient matrices $H(V)$ given by

$$H(V) = A_f + \beta\ \mathrm{Diag}(|V|)^2, \tag{6.8}$$

where $A_f \in \mathbb{C}^{n\times n}$ is a Hermitian matrix and positive definite, $\beta > 0$ is a parameter, $V \in \mathbb{C}^n$ is a complex vector, and $|\cdot|$ takes elementwise absolute value. Such an NEPv arises from discretizing the Gross–Pitaevskii equation (GPE) for modeling the physical phenomenon of Bose–Einstein condensation [4, 11, 12, 16].

The matrix A_f in (6.8) is dependent on a potential function f. For illustration, we will discuss a model 2D GPE studied in [11], where for a given potential function $f(x, y)$ over a two-dimensional domain $[-\ell, \ell] \times [-\ell, \ell]$, the corresponding matrix is

$$A_f = \mathrm{Diag}(\widetilde{f}) - \frac{1}{2}M - \imath\omega M_\phi, \tag{6.9}$$

where

$$\widetilde{f} = h^2\ [f(x_1, y_1), \ldots, f(x_N, y_1), f(x_1, y_2), \ldots, f(x_N, y_2), \ldots, f(x_N, y_N)]^{\mathrm{T}} \in \mathbb{R}^{N^2}$$

with $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$ being interior points of the interval $[-\ell, \ell]$ from the $N + 2$ equidistant discretization with spacing $h = \frac{2\ell}{N+1}$. The matrices M, M_ϕ are given by

$$M = D_{2,N}\otimes I + I\otimes D_{2,N},\ M_\phi = h\ \mathrm{Diag}(y_1, \ldots, y_N)\otimes D_N - D_N\otimes(h\ \mathrm{Diag}(x_1, \ldots, x_N))$$

with $N\times N$ tridiagonal matrices $D_N = \mathrm{tridiag}(-\frac{1}{2}, 0, \frac{1}{2})$ and $D_{2,N} = \mathrm{tridiag}(1, -2, 1)$.

Since V is a vector, by definition (2.12) the directional derivative operator of $H(V)$ is given by

$$\mathbf{D}H(V)[X] = 2\beta\ \mathrm{Diag}(\Re(\overline{V}\odot X)).$$

The local $\mathbb{R}$-linear operator of the plain SCF $\mathscr{L} : \mathbb{C}^{n-1} \to \mathbb{C}^n$ in (3.8) is

$$\mathscr{L}(Z) = 2\beta\, D(V_*)\odot(V_{*\perp}^{\mathrm{H}}\ \mathrm{Diag}(\Re(\overline{V}_*\odot(V_{*\perp}Z)))\, V_*), \tag{6.10}$$

and its adjoint operator $\mathscr{L}^*$, with respect to the standard inner product in $\mathbb{C}^{(n-k)\times k}$ ($k = 1$), i.e., $\langle\mathscr{L}(Z), Y\rangle \equiv \Re(\mathrm{tr}(Y^{\mathrm{H}}\mathscr{L}(Z))) = \langle Z, \mathscr{L}^*(Y)\rangle \equiv \Re(\mathrm{tr}([\mathscr{L}^*(Y)]^{\mathrm{H}}Z))$ for any Y, $Z \in \mathbb{C}^{(n-k)\times k}$, is given by

$$\mathscr{L}^*(Y) = 2\beta\, V_{*\perp}^{\mathrm{H}}\ \left(\Re\left(\mathrm{diag}(V_{*\perp}(D(V_*)\odot Y)V_*^{\mathrm{H}})\right)\odot V_*\right); \tag{6.11}$$

see Appendix A for the derivation.

For the level-shifted SCF, the local $\mathbb{R}$-linear operator $\mathscr{L}_\sigma$ in (5.7) is given by

$$\mathscr{L}_\sigma(Z) = D_\sigma(V_*)\odot\mathscr{Q}(Z) - I_{\mathrm{id}}, \tag{6.12}$$

where the restricted derivative operator $\mathscr{Q}(Z)$ is given by

$$\mathscr{Q}(Z) = 2\beta\, V_{*\perp}^{\mathrm{H}}\ \mathrm{Diag}(\Re(\overline{V_*}\odot(V_{*\perp}Z))\, V_* + (\Lambda_{*\perp}Z - Z\Lambda_*). \tag{6.13}$$

The largest eigenvalue $\mu_{\max}$ of $\mathscr{Q}$ can be bounded as follows. Let $Z \in \mathbb{C}^{n-1}$ be the eigenvector associated with $\mu_{\max}$. Then

$$\mu_{\max} = \frac{\|\mathscr{Q}(Z)\|_{\mathrm{F}}}{\|Z\|_{\mathrm{F}}} \leqslant 2\beta \frac{\|\operatorname{Diag}(\Re(\overline{V} \odot (V_{*\perp} Z))\|_{\mathrm{F}}}{\|Z\|_{\mathrm{F}}} + s_* \leqslant 2\beta + s_* \leqslant 3\beta + \|A_f\|_2,$$

where $s_* = \lambda_n(H(V_*)) - \lambda_1(H(V_*))$ is the spectral span, and for the last inequality we have used the inequalities $s_* \leqslant \lambda_n(H(V_*)) \leqslant \beta + \|A_f\|_2$ due to $H(V)$ in (6.8) being positive definite. Consequently, the lower bound on σ in (5.10) yields

$$\sigma \geqslant \frac{1}{2}(3\beta + \|A_f\|_2) \geqslant \frac{\mu_{\max}}{2} > \frac{\mu_{\max}}{2} - \delta_* \tag{6.14}$$

to ensure the local convergence of the level-shifted SCF.

Example 6.3. In this example, we select the parameters $\ell = 1$, $\omega = 0.85$, and $N = 10$ (hence $n = 100$). We use a radial symmetric potential $f(x, y) = (x^2 + y^2)/2$. Various values of β ranging from 0.5 to 5 have been tried. The simulation results are shown in Figure 3.

It is observed that the plain SCF becomes slower and slower and eventually divergent as β increases. Again, the spectral radius $\rho(\mathscr{L}_\sigma)$ and $\eta_{\sup}$ can well capture true convergence behavior. In particular, at $\beta = 3.5$, we find that up to 7 digits,

$$\text{observed} = 0.9136140,\ \eta_{\sup,\infty} = 0.9136173,\ \eta_{\sup} = 1.019727,\ \eta_{\mathrm{czbl}} = 2.342686.$$

Again, we see the sharpness of the estimate $\eta_{\sup,\infty}$.

The performance of the level-shifted SCF with respect to different shifts σ is shown in Figure 4, where we observe a similar convergence behavior to Figure 2 for Example 6.2 on the impact of the choice of shift σ.

Example 6.4. Exploiting symmetry in the potential function can be important for the numerical solution of a GPE [3]. This example is a repeat of Example 6.3 using an asymmetric potential function $f(x, y) = (x^2 + 100y^2)/2$. The plots in Figure 5 show only a slightly different performance of the plain SCF (1.3) compared to the radial

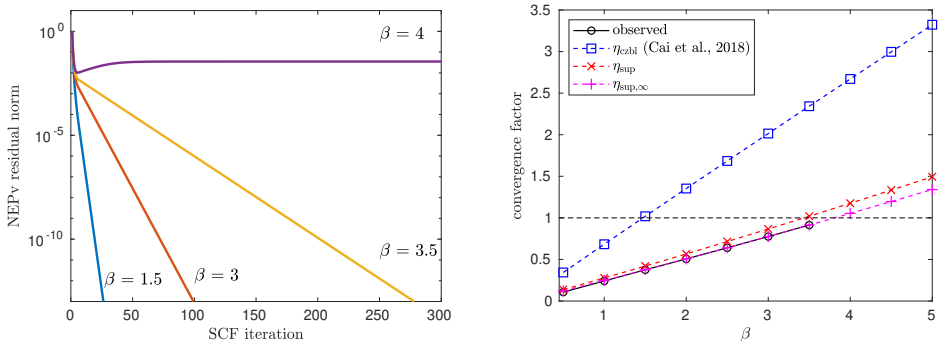

FIG. 3. *Example* 6.3*: convergence history of residual norm* $\|H(V_i)V_i - V_i\Lambda_i\|_2$ *by the plain SCF* (1.3) *for selected* β *(left plot) and convergence rate estimates as* β *varies (right plot). The observed rate marked by "circle" and the theoretic rate* $\eta_{\sup,\infty}$ *marked by "+" coincide perfectly.*

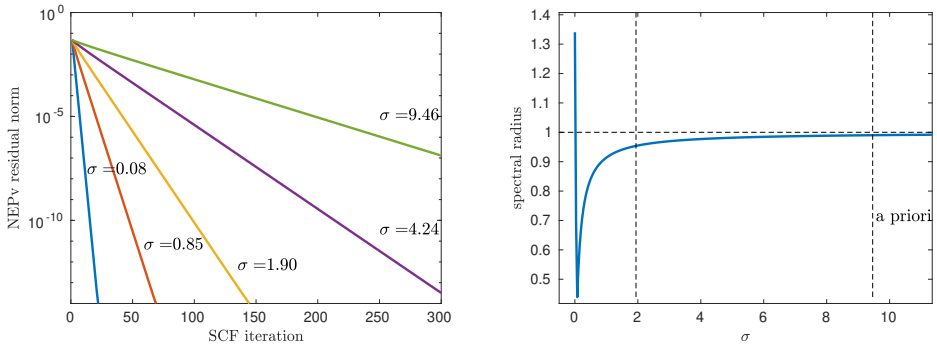

FIG. 4. *Example* 6.3*: convergence history of residual norm* $\|H(V_i)V_i - V_i\Lambda_i\|_2$ *by the level-shifted SCF* (5.1) *with selected* σ *(left plot) and spectral radius* $\rho(\mathscr{L}_\sigma)$ *as shift* σ *varies (right plot). The first vertical dash line is* $\sigma = \mu_{\max}/2 - \delta_*$ *as in* (5.10)*, the theoretical bound beyond which provably* $\rho(\mathscr{L}_\sigma) < 1$*, and the second one is a priori* $\sigma = \frac{1}{2}(3\beta + \|A_f\|_2)$ *suggested by* (6.14)*, while the optimal shift is* $\sigma \approx 0.08$*, which is smaller than* $\mu_{\max}/2 - \delta_*$*, the theoretical bound.* $H(V)$ *is given by* (6.8) *with* $\beta = 5$.

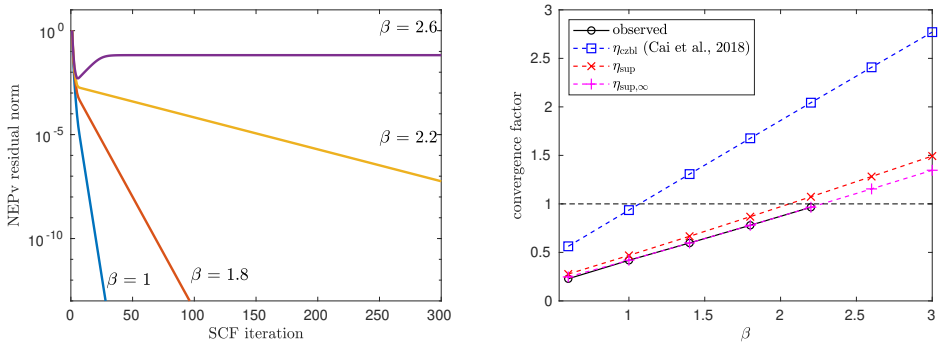

FIG. 5. *Example* 6.4*: convergence history of residual norm* $\|H(V_i)V_i - V_i\Lambda_i\|_2$ *by the plain SCF* (1.3) *for selected* β *(left plot) and convergence rate estimates as* β *varies (right plot). The observed rate marked by "circle" and the theoretic rate* $\eta_{\sup,\infty}$ *marked by "+" coincide perfectly.*

symmetric case of Example 6.3. The sharpness of the estimate $\eta_{\sup,\infty}$ on the local convergence rate can be best seen at $\beta = 2.2$, where, up to 7 digits,

$$\text{observed} = 0.9652599,\ \eta_{\sup,\infty} = 0.9652614,\ \eta_{\sup} = 1.073434,\ \eta_{\text{czbl}} = 2.043247.$$

The performance of the level-shifted SCF is depicted in Figure 6. Again we observe a similar convergence behavior to Example 6.3 with repect to the choice of shift σ.

7. Concluding remarks. We have presented a comprehensive local convergence analysis of the plain SCF iteration and its level-shifted variant for solving NEPvs. The exact convergence rate and its estimates are established. Our analysis is in terms of the tangent-angle matrix to measure the approximation error between consecutive SCF iterates and the intended target. We first established a relation between the tangent-angle matrices associated with any two consecutive SCF approximates, and with it we developed new formulas for the local error contraction factor and the asymptotic average contraction factor of SCF. The new formulas are sharper and

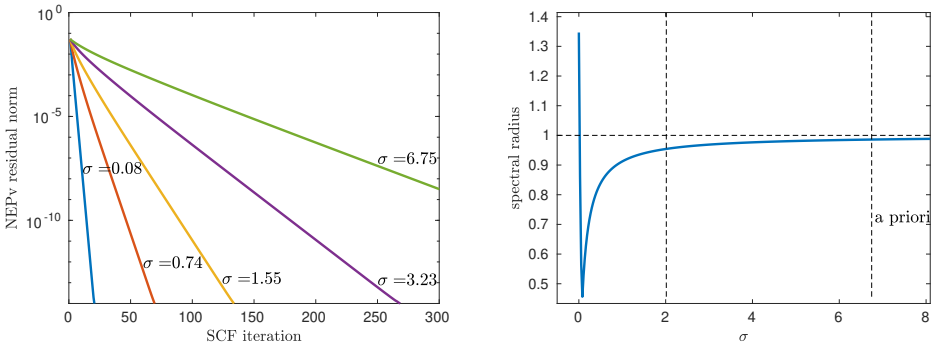

FIG. 6. *Example 6.4: convergence history of residual norm* $\|H(V_i)V_i - V_i\Lambda_i\|_2$ *by the level-shifted SCF* (5.1) *with selected* σ *(left plot) and spectral radius of* $\rho(\mathscr{L}_\sigma)$ *as shift* σ *varies (right plot). The first vertical dash line is* $\sigma = \mu_{\max}/2 - \delta_*$ *as in* (5.10)*, the theoretical bound beyond which provably* $\rho(\mathscr{L}_\sigma) < 1$*, and the second one is a priori* $\sigma = \frac{1}{2}(3\beta + \|A_f\|_2)$ *suggested by* (6.14)*, and the optimal shift is* $\sigma \approx 0.08$*, which is smaller than* $\mu_{\max}/2 - \delta_*$*, the theoretical bound. The* $H(V)$ *is given by* (6.8) *with* $\beta = 3$.

complement existing local convergence results. With the help of new convergence rate estimates, we derive an explicit lower bound on the shifting parameter to guarantee local convergence of the level-shifted SCF. These results are numerically confirmed by examples from applications in computational physics and chemistry.

Our analysis does not cover other sophisticated variants of SCF such as the damped SCF [6] and the direct inversion of iterative subspace [25, 26]. It is conceivable that by the tangent-angle matrix and the eigenspace perturbation theory, one may work out the local convergence analysis of those variants.

Finally, we note that we focused on NEPv (1.1) satisfying the invariant property (1.2). While this property is formulated as a result of some practically important applications, there are recent emerging NEPvs (1.1) that do not have this property, such as the one in [43], and yet similar SCF iterations can be used. It would be interesting to find out what now determines the local convergence rate. This will be a future project to pursue.

Appendix A. Adjoint operators.

The adjoint operator $\mathscr{L}^*$ in (6.4) is derived as follows:

$$
\begin{aligned}
&\langle Y, \mathscr{L}(Z)\rangle \\
&\quad = 2\alpha\langle Y,\ D(V_*)\odot(V_{*\perp}^{\mathrm{T}}\ \mathrm{Diag}(L^{-1}\,\mathrm{diag}(V_{*\perp}ZV_*^{\mathrm{T}}))\,V_*)\rangle \\
&\quad \overset{(1)}{=} 2\alpha\langle D(V_*)\odot Y,\ V_{*\perp}^{\mathrm{T}}\ \mathrm{Diag}(L^{-1}\,\mathrm{diag}(V_{*\perp}ZV_*^{\mathrm{T}}))\,V_*\rangle \ \text{by } \langle Y, D\odot X\rangle = \langle D\odot Y, X\rangle \\
&\quad \overset{(2)}{=} 2\alpha\langle V_{*\perp}[D(V_*)\odot Y]V_*^{\mathrm{T}},\ \mathrm{Diag}(L^{-1}\,\mathrm{diag}(V_{*\perp}ZV_*^{\mathrm{T}}))\rangle \ \text{by } \langle Y, AXB\rangle = \langle A^{\mathrm{T}}YB^{\mathrm{T}}, X\rangle \\
&\quad \overset{(3)}{=} 2\alpha\langle \mathrm{diag}(V_{*\perp}[D(V_*)\odot Y]V_*^{\mathrm{T}}),\ L^{-1}\,\mathrm{diag}(V_{*\perp}ZV_*^{\mathrm{T}})\rangle \ \text{by } \langle Y, \mathrm{Diag}(b)\rangle = \langle \mathrm{diag}(Y), b\rangle \\
&\quad \overset{(4)}{=} 2\alpha\langle L^{-1}\,\mathrm{diag}(V_{*\perp}[D(V_*)\odot Y]V_*^{\mathrm{T}}),\ \mathrm{diag}(V_{*\perp}ZV_*^{\mathrm{T}})\rangle \ \text{by moving } L \text{ to the left} \\
&\quad \overset{(5)}{=} 2\alpha\langle \mathrm{Diag}(L^{-1}\,\mathrm{diag}(V_{*\perp}[D(V_*)\odot Y]V_*^{\mathrm{T}})),\ V_{*\perp}ZV_*^{\mathrm{T}}\rangle \ \text{by } \langle b, \mathrm{diag}(Y)\rangle = \langle \mathrm{Diag}(b), Y\rangle.
\end{aligned}
$$

Finally, moving the last $V_{*\perp}$ and V_* to the left we obtain the formula (6.4).

The adjoint operator $\mathscr{L}^*$ in (6.11) is derived analogously. The first three steps are exactly the same as above, and so we continue with

$$\begin{aligned}
\langle Y, \mathscr{L}(Z)\rangle &= 2\beta\langle Y,\ D(V_*)\odot(V_{*\perp}^{\mathrm{H}}\ \mathrm{Diag}(\Re(\overline{V}_*\odot(V_{*\perp}Z)))\,V_*)\rangle \\
&= 2\beta\langle \mathrm{diag}(V_{*\perp}[D(V_*)\odot Y]V_*^{\mathrm{H}}), \Re(\overline{V}_*\odot(V_{*\perp}Z))\rangle \text{ by identities (1)–(3)} \\
&= 2\beta\langle \Re(\mathrm{diag}(V_{*\perp}[D(V_*)\odot Y]V_*^{\mathrm{H}})),\ \overline{V}_*\odot(V_{*\perp}Z)\rangle \text{ by vector inner product} \\
&= 2\beta\langle \Re(\mathrm{diag}(V_{*\perp}[D(V_*)\odot Y]V_*^{\mathrm{H}}))\odot V_*,\ V_{*\perp}Z\rangle \text{ by } \langle a, b\odot c\rangle = \langle a\odot\overline{b}, c\rangle.
\end{aligned}$$

Finally, moving the last $V_{*\perp}$ to the left, we obtain the formula (6.11).

Acknowledgments. We would like to thank anonymous referees for their careful reading and constructive comments and suggestions that have significantly improved our presentation.

REFERENCES

[1] P. A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princetin, NJ, 2009.
[2] Z. Bai, D. Lu, and B. Vandereycken, *Robust Rayleigh quotient minimization and nonlinear eigenvalue problems*, SIAM J. Sci. Comput., 40 (2018), pp. A3495–A3522.
[3] W. Bao and Y. Cai, *Mathematical theory and numerical methods for Bose-Einstein condensation*, Kinet. Relat. Models, 6 (2013), pp. 1–135.
[4] W. Bao and Q. Du, *Computing the ground state solution of Bose–Einstein condensates by a normalized gradient flow*, SIAM J. Sci. Comput., 25 (2004), pp. 1674–1697.
[5] Y. Cai, L.-H. Zhang, Z. Bai, and R.-C. Li, *On an eigenvector-dependent nonlinear eigenvalue problem*, SIAM J. Matrix Anal. Appl., 39 (2018), pp. 1360–1382.
[6] E. Cancès and C. Le Bris, *Can we outperform the DIIS approach for electronic structure calculations?*, Int. J. Quantum Chem., 79 (2000), pp. 82–90.
[7] E. Cancès and C. Le Bris, *On the convergence of SCF algorithms for the Hartree–Fock equations*, ESAIM Math. Model. Numer. Anal., 34 (2000), pp. 749–774.
[8] C. Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation.* III, SIAM J. Numer. Anal., 7 (1970), pp. 1–46.
[9] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.
[10] A. Isaev, *Twenty-One Lectures on Complex Analysis*, Springer, Cham, 2017.
[11] E. Jarlebring, S. Kvaal, and W. Michiels, *An inverse iteration method for eigenvalue problems with eigenvector nonlinearities*, SIAM J. Sci. Comput., 36 (2014), pp. A1978–A2001.
[12] S. Jia, H. Xie, M. Xie, and F. Xu, *A full multigrid method for nonlinear eigenvalue problems*, Sci. China Math., 59 (2016), pp. 2037–2048.
[13] 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, Springer, Cham, 2014, pp. 263–279.
[14] J. Koutecký and V. Bonačić, *On convergence difficulties in the iterative Hartree–Fock procedure*, J. Chem. Phys., 55 (1971), pp. 2408–2413.
[15] P. Lax, *Functional Analysis*, Wiley, New York, 2002.
[16] X.-G. Li, Y. Cai, and P. Wang, *Operator-compensation methods with mass and energy conservation for solving the Gross-Pitaevskii equation*, Appl. Numer. Math., 151 (2020), pp. 337–353.
[17] 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.
[18] X. Liu, Z. Wen, X. Wang, and Y. Ulbrich, and M. Yuan, *On the analysis of the discretized Kohn–Sham density functional theory*, SIAM J. Numer. Anal., 53 (2015), pp. 1758–1785.
[19] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods*, Cambridge University Press, Cambridge, UK, 2004.
[20] R. McWeeny, *Some recent advances in density matrix theory*, Rev. Modern Phys., 32 (1960), pp. 335–369.
[21] R. Meyer, *Nonlinear eigenvector algorithms for local optimization in multivariate data analysis*, Linear Algebra Appl., 264 (1997), pp. 225–246.

[22] T. T. Ngo, M. Bellalij, and Y. Saad, *The trace ratio optimization problem*, SIAM Rev., 54 (2012), pp. 545–569.
[23] J. Nocedal and S. Wright, *Numerical Optimization*, Springer Science & Business Media, New York, 2006.
[24] B. N. Parlett, *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, 1998.
[25] P. Pulay, *Convergence acceleration of iterative sequences. The case of SCF iteration*, Chem. Phys. Lett., 73 (1980), pp. 393–398.
[26] P. Pulay, *Improved SCF convergence acceleration*, J. Comput. Chem., 3 (1982), pp. 556–560.
[27] L. Qiu, Y. Zhang, and C.-K. Li, *Unitarily invariant metrics on the Grassmann space*, SIAM J Matrix Anal. Appl., 27 (2005), pp. 507–531.
[28] R. Remmert, *Theory of Complex Functions*, Springer Science & Business Media, New York, 1991.
[29] C. C. J. Roothaan, *New developments in molecular orbital theory*, Rev. Modern Phys., 23 (1951), pp. 69–89.
[30] V. Saunders and I. Hillier, *A Level–Shifting method for converging closed shell Hartree–Fock wave functions*, Int. J. Quantum Chem., 7 (1973), pp. 699–705.
[31] R. E. Stanton, *The existence and cure of intrinsic divergence in closed shell SCF calculations*, J. Chem. Phys., 75 (1981), pp. 3426–3432.
[32] R. E. Stanton, *Intrinsic convergence in closed-shell SCF calculations. A general criterion*, J. Chem. Phys., 75 (1981), pp. 5416–5422.
[33] G. W. Stewart, *Matrix Algorithms: Volume* II: *Eigensystems*, SIAM, Philadelphia, 2001.
[34] G. W. Stewart and J. G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.
[35] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, Courier Corporation, New York, 2012.
[36] L. Thøgersen, J. Olsen, D. Yeager, P. Jørgensen, P. Sałek, and T. Helgaker, *The trust-region self-consistent field method: Towards a black-box optimization in Hartree–Fock and Kohn–Sham theories*, J. Chem. Phys., 121 (2004), pp. 16–27.
[37] F. Tudisco and D. J. Higham, *A nonlinear spectral method for core-periphery detection in networks*, SIAM J. Math. Data Science, 1 (2019), pp. 269–292.
[38] P. Upadhyaya, E. Jarlebring, and E. H. Rubensson, *A density matrix approach to the convergence of the self-consistent field iteration*, Numer. Algebra Control Optim., 11 (2021), pp. 99–115.
[39] R. S. Varga, *Matrix Iterative Analysis*, Springer-Verlag, Berlin, 2000.
[40] 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.
[41] 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.
[42] L. Zhang and R.-C. Li, *Maximization of the sum of the trace ratio on the Stiefel manifold,* I*: Theory*, Sci. China Math., 57 (2014), pp. 2495–2508.
[43] L. Zhang, L. Wang, Z. Bai, and R.-C. Li, *A self-consistent-field iteration for orthogonal canonical correlation analysis*, IEEE Trans. Pattern Anal. Mach. Intell., 44 (2022), pp. 890–904. https://doi.org/10.1109/TPAMI.2020.3012541.
[44] 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.
[45] P. Zhu and A. V. Knyazev, *Angles between subspaces and their tangents*, J. Numer. Math., 21 (2013), pp. 325–340.