SIAM J. MATRIX ANAL. APPL.
Vol. 45, No. 1, pp. 84–111

VARIATIONAL CHARACTERIZATION OF MONOTONE NONLINEAR EIGENVECTOR PROBLEMS AND GEOMETRY OF SELF-CONSISTENT FIELD ITERATION*

ZHAOJUN BAI[†] AND DING LU[‡]

Abstract. This paper concerns a class of monotone eigenvalue problems with eigenvector nonlinearities (mNEPv). The mNEPv is encountered in applications such as the computation of joint numerical radius of matrices, best rank-one approximation of third-order partial-symmetric tensors, and distance to singularity for dissipative Hamiltonian differential-algebraic equations. We first present a variational characterization of the mNEPv. Based on the variational characterization, we provide a geometric interpretation of the self-consistent field (SCF) iterations for solving the mNEPv, prove the global convergence of the SCF, and devise an accelerated SCF. Numerical examples demonstrate theoretical properties and computational efficiency of the SCF and its acceleration.

Key words. nonlinear eigenvalue problem, self-consistent field iteration, variational characterization, geometry of SCF, convergence analysis

MSC codes. 65F15, 65H17

DOI. 10.1137/22M1525326

1. Introduction.

We consider the following eigenvector-dependent nonlinear eigenvalue problem:

$$H(x)\,x = \lambda\, x, \tag{1.1}$$

where $H(x)$ is a Hermitian matrix-valued function of the form

$$H(x) := \sum_{i=1}^{m} h_i(x^H A_i x)\, A_i, \tag{1.2}$$

$\{A_i\}$ are n-by-n Hermitian matrices, and $\{h_i\}$ are differentiable and nondecreasing functions over $\mathbb{R}$. The goal is to find a unit-length vector $x \in \mathbb{C}^n$ and a scalar $\lambda \in \mathbb{R}$ satisfying (1.1), and, furthermore, $\lambda\,(= x^H H(x) x)$ is the largest eigenvalue of $H(x)$. The solution vector x is called an eigenvector of the eigenvalue problem (1.1), and λ is the corresponding eigenvalue. Since $H(\gamma x) \equiv H(x)$ for any $\gamma \in \mathbb{C}$ with $|\gamma| = 1$, if x is an eigenvector, then so is γx.

The matrix-valued function $H(x)$ in (1.2) is a linear combination of constant matrices $\{A_i\}$ with monotonic functions $\{h_i\}$. We say $H(x)$ is of a monotone affine-linear structure and, for simplicity, call the eigenvalue problem (1.1) a monotone NEPv, or mNEPv. For the case $m = 1$, the mNEPv simplifies to $h(x^H Ax) \cdot Ax = \lambda x$, so its eigenvector x must also be an eigenvector of the Hermitian matrix A, and by the monotonicity of h, x corresponds to the largest eigenvalue of A.

[*] Received by the editors September 28, 2022; accepted for publication (in revised form) by K. Meerbergen August 22, 2023; published electronically January 11, 2024.
https://doi.org/10.1137/22M1525326

Funding: The work of the first author was supported by NSF DMS-1913364. The work of the second author was supported by NSF DMS-2110731.

[†] Department of Computer Science, University of California at Davis, Davis, CA 95616 USA (zbai@ucdavis.edu).

[‡] Department of Mathematics, University of Kentucky, Lexington, KY 40506 USA (Ding.Lu@uky.edu).

In section 2, we will see that the mNEPv (1.1) is intrinsically related to the following maximization problem:

$$\max_{x\in\mathbb{C}^n,\,\|x\|=1}\left\{F(x):=\sum_{i=1}^{m}\phi_i\left(x^HA_ix\right)\right\}, \tag{1.3}$$

where $\{\phi_i\}$ are antiderivatives of $\{h_i\}$, i.e., $\phi_i'(t)=h_i(t)$ for $i=1,\dots,m$. Since $\{h_i\}$ are differentiable and nondecreasing, $\{\phi_i\}$ are twice-differentiable and convex functions. We call (1.3) an associated maximization of the mNEPv (1.1), or aMax.

The mNEPv (1.1) is a class of the eigenvalue problems with eigenvector nonlinearities (NEPv). NEPv have been extensively studied in the Kohn–Sham density functional theory for electronic structure calculations [42] and the Gross–Pitaevskii eigenvalue problem, a nonlinear Schrödinger equation to describe the ground states of ultracold bosonic gases [9, 31]. NEPv have also been found in a variety of computational problems in data science, e.g., Fisher's linear discriminant analysis [47, 66, 67] and its robust version [8], spectral clustering using the graph p-Laplacian [16], core-periphery detection in networks [57], and orthogonal canonical correlation analysis [68].

Self-consistent field (SCF) iteration is a gateway algorithm to solve NEPv, much like the power method for solving linear eigenvalue problems. The SCF was introduced back in the 1950s [54]. Since then, the convergence analysis of the SCF has long been an active research topic in the study of NEPv; see [7, 17, 18, 40, 55, 59].

Although the underlying structure of the mNEPv (1.1) is commonly found in NEPv, it has been largely unexploited. In this paper, we will conduct a systematical study of the mNEPv and exploit its underlying structure. Theoretically, we will reveal a variational characterization of the mNEPv (1.1) by maximizers of the aMax (1.3). Using the variational characterization, we will provide a geometric interpretation of the SCF for solving the mNEPv (1.1), which reveals the global convergence of the algorithm. We will then prove the global monotonic convergence of the SCF. Finally, we will present an accelerated SCF by exploiting the underlying structure of $H(x)$ and demonstrate its efficiency with examples from a variety of applications.

The aMax (1.3) is interesting in its own right and finds numerous applications. One important source of the problems is a quartic maximization over the Euclidean ball, where $\phi_i(t)=t^2$ [46]. In section 5, we will discuss such quartic maximization problems arising from the joint numerical radius computation and the rank-one approximation of partial-symmetric tensors. Another application of the aMax (1.3) is from computing the distance to singularity for dissipative Hamiltonian differential-algebraic equation (dHDAE) systems [43]. The aMax (1.3) also arises in robust optimization with ellipsoid uncertainty; see e.g., [12]. By the intrinsic connection between the mNEPv and the aMax, we will devise an eigenvalue-based approach for solving the aMax that can exploit state-of-the-art eigensolvers from numerical linear algebra.

Optimizations of the form (1.3) have been investigated in the literature, but they are often formulated as the *minimization* of $F(x)$ over the vector space $\mathbb{R}^n$ or $\mathbb{C}^n$. Examples of recent studies include the quartic-quadratic optimization with $\phi_i(t)=t^2$ or t [29, 65] and the Crawford number computation with $\phi_i(t)=t^2$ [41]. For these minimization problems, eigenvalue-based approaches have been developed which lead to NEPv $H(x)x=\lambda x$ with $H(x)$ given by (1.2) and λ corresponding to the *smallest* eigenvalue of $H(x)$; see [29, 41]. However, as the target eigenvalue is the smallest rather than the largest, the solution and analyses of those NEPv differ fundamentally from those of the mNEPv (1.1). For example, the SCF is no longer globally convergent for computing the smallest eigenvalue.

The rest of this paper is organized as follows. Section 2 presents a variational characterization of the mNEPv (1.1) through maximizers of the aMax (1.3). Section 3 provides a geometric interpretation of the SCF and proves its global convergence. Section 4 focuses on the practical aspects of the SCF. Section 5 discusses the applications of the mNEPv (1.1). Numerical experiments are presented in section 6, and concluding remarks are provided in section 7.

We follow standard notation in matrix computations. $\mathbb{R}^{m\times n}$ and $\mathbb{C}^{m\times n}$ are the sets of m-by-n real and complex matrices, respectively. $\mathrm{Re}(\cdot)$ extracts the real part of a complex matrix or a number. For a matrix (or a vector) X, X^T stands for transpose, X^H for conjugate transpose, and $\|X\|$ for the matrix 2-norm. We use $\lambda_{\min}(X)$ and $\lambda_{\max}(X)$ for the smallest and largest eigenvalues of a Hermitian X. The spectral radius (i.e., the largest absolute value of eigenvalues) of a matrix or linear operator is denoted by $\rho(\cdot)$. Standard little-o and big-O notations are used: $f(x)=\mathbf{o}(g(x))$ means that $f(x)/g(x)\to 0$ as $x\to 0$, while $f(x)=\mathcal{O}(g(x))$ means that $f(x)/g(x)\le c$ for some constant c as $x\to 0$. Other notations will be explained as used.

2. Variational characterization.

Variational characterizations provide powerful tools to the study of eigenvalue problems, facilitating both theoretical analysis and numerical computations. A prominent example is the Hermitian linear eigenvalue problem of the form $Ax=\lambda x$, where the Courant–Fischer principle uses optimizers of the Rayleigh quotient x^HAx/x^Hx to form variational characterizations of the eigenvalues of A; see, e.g., [14]. With this characterization, bounds for eigenvalues, and interlacing, monotonicity of eigenvalues can be proved quickly. Variational characterizations have also been developed for eigenvalue-dependent nonlinear eigenvalue problems of the form $T(\lambda)x=0$ [34]. It is also well known that the NEPv in Kohn–Sham density functional theory is derived from the minimization of an energy function in electronic structure calculations; see, e.g., [42, 18]. In this section, we provide a variational characterization of the mNEPv (1.1) by exploring its relation to the aMax (1.3).

2.1. Stability of eigenvectors.

We start with the following NEPv without assuming the structure of $H(x)$ and the order of the eigenvalue λ:

$$H(x)x=\lambda x \quad \text{with} \quad \|x\|=1, \tag{2.1}$$

where $H(x)$ is Hermitian, differentiable (w.r.t. both real and imaginary parts of x), and unitarily scaling invariant (i.e., $H(\gamma x)=H(x)$ for any $\gamma\in\mathbb{C}$ with $|\gamma|=1$). Due to scaling invariance, we can view an eigenvector x of the NEPv (2.1) as an equivalent class $[x]:=\{\,\gamma x\mid \gamma\in\mathbb{C},\,|\gamma|=1\,\}$, i.e., a point in the Grassmannian $\mathrm{Gr}(1,\mathbb{C}^n)$.

Let x_* be an eigenvector of the NEPv (2.1) and the corresponding λ_* be the pth largest eigenvalue of $H(x_*)$. Assume that λ_* is a simple eigenvalue. Then $[x_*]$ can be interpreted as a solution to the fixed-point equation over $\mathrm{Gr}(1,\mathbb{C}^n)$,

$$[x]=\Pi([x]), \tag{2.2}$$

where the mapping $\Pi:\mathrm{Gr}(1,\mathbb{C}^n)\to\mathrm{Gr}(1,\mathbb{C}^n)$ is defined by $\Pi([x]):=[u(x)]$ and $u(x)$ is an (arbitrary) unit eigenvector for the pth largest eigenvalue of $H(x)$. The *attractiveness* of the fixed point $[x_*]$ for the mapping Π in (2.2) can be determined by the spectral radius of a related linear operator, as established in [7]. To introduce this linear operator, we first denote the eigenvalue decomposition of $H(x_*)$ as

$$H(x_*)\begin{bmatrix}x_* & X_{*\perp}\end{bmatrix}=\begin{bmatrix}x_* & X_{*\perp}\end{bmatrix}\begin{bmatrix}\lambda_* & \\ & \Lambda_{*\perp}\end{bmatrix}, \tag{2.3}$$

where $\begin{bmatrix} x_* & X_{*\perp} \end{bmatrix} \in \mathbb{C}^{n\times n}$ is unitary and $\Lambda_{*\perp} \in \mathbb{R}^{(n-1)\times(n-1)}$ is a diagonal matrix. We then define an $\mathbb{R}$-linear operator[1]

$$\mathcal{L} : \mathbb{C}^{n-1} \to \mathbb{C}^{n-1} \quad \text{with} \quad \mathcal{L}(z) = D_*^{-1} X_{*\perp}^H \left(\mathbf{D}H(x_*)[\,X_{*\perp} z\,] \right) x_*, \tag{2.4}$$

where $D_* = \lambda_* I_{n-1} - \Lambda_{*\perp}$ is diagonal and nonsingular since λ_* is a simple eigenvalue and $\mathbf{D}H(x)[\,d\,]$ is the derivative of H at x along the direction of d:

$$\mathbf{D}H(x)[\,d\,] := \lim_{\alpha\in\mathbb{R},\ \alpha\to 0} \frac{H(x+\alpha d) - H(x)}{\alpha}. \tag{2.5}$$

Let $\rho(\mathcal{L})$ be the spectral radius of $\mathcal{L}$ (i.e., the largest absolute value of the eigenvalues). Then by [7, Thm. 4.2], we know that if $\rho(\mathcal{L}) < 1$, then $[x_*]$ is an attractive fixed point of the mapping Π (2.2); if $\rho(\mathcal{L}) > 1$, then $[x_*]$ is a repulsive fixed point; and if $\rho(\mathcal{L}) = 1$, then no immediate conclusion can be drawn for the attractiveness of $[x_*]$. It is worth noting that although the theorem [7, Thm. 4.2] is stated for the case $\lambda_* = \lambda_n$ being the smallest eigenvalue of $H(x_*)$, the result holds for a general pth eigenvalue.

Returning to the mNEPv (1.1), in the following lemma, we can show that the operator $\mathcal{L}$ in (2.4) is both self-adjoint and positive semidefinite. Consequently, the conditions $\rho(\mathcal{L}) < 1$ or $\rho(\mathcal{L}) \le 1$ can be characterized using the definiteness of a characteristic function. To facilitate the analysis, we denote the vector space $\mathbb{C}^{n-1}$ over the field of real numbers $\mathbb{R}$ as $\mathbb{C}^{n-1}(\mathbb{R})$ and introduce an inner product over $\mathbb{C}^{n-1}(\mathbb{R})$ as

$$\langle\, y, z\,\rangle_D := \mathrm{Re}(\, y^H D z\,), \tag{2.6}$$

where D is a given Hermitian positive definite matrix of size $n-1$.

Lemma 2.1. *Let $x_* \in \mathbb{C}^n$ be an eigenvector of the mNEPv* (1.1) *with a simple eigenvalue λ_*. Then the $\mathbb{R}$-linear operator $\mathcal{L}$ in* (2.4) *is self-adjoint and positive semidefinite over $\mathbb{C}^{n-1}(\mathbb{R})$ in the inner product* (2.6) *with $D_* = \lambda_* I_{n-1} - \Lambda_{*\perp}$. Moreover,*

(a) $\rho(\mathcal{L}) < 1$ *if and only if $\varphi(d\,; x_*) < 0$ for all $d \neq 0$ and $d^H x_* = 0$;*

(b) $\rho(\mathcal{L}) \le 1$ *if and only if $\varphi(d\,; x_*) \le 0$ for all $d \neq 0$ and $d^H x_* = 0$.*

Here, $\varphi(d\,; x_*)$ is a quadratic function in $d \in \mathbb{C}^n$ and is parameterized by x_* as

$$\varphi(d\,; x_*) := d^H \Big(H(x_*) - (x_*^H H(x_*) x_*)\, I \Big) d + 2 \sum_{i=1}^m h_i'(x_*^H A_i x_*) \cdot (\mathrm{Re}(d^H A_i x_*))^2. \tag{2.7}$$

Proof. To show that $\mathcal{L}$ is self-adjoint and positive semidefinite, we first derive from the definition (1.2) of $H(x)$ that the directional derivative (2.5) is given by

$$\mathbf{D}H(x)[\,d\,] = 2 \sum_{i=1}^m \mathrm{Re}\left(x^H A_i d \right) \cdot h_i'(x^H A_i x) \cdot A_i.$$

Therefore, the $\mathbb{R}$-linear operator $\mathcal{L}$ in (2.4) takes the form of

$$\mathcal{L}(z) = 2 D_*^{-1} \sum_{i=1}^m \mathrm{Re}(x_*^H A_i X_{*\perp} z) \cdot h_i'(x_*^H A_i x_*) \cdot X_{*\perp}^H A_i x_*. \tag{2.8}$$

Since λ_* is a simple largest eigenvalue, $D_* = \lambda_* I_{n-1} - \Lambda_{*\perp}$ is a diagonal and positive definite matrix. A quick verification shows that

$$\langle\, \mathcal{L}(y), z\,\rangle_{D_*} = 2 \textstyle\sum_{i=1}^m h_i'(x_*^H A_i x_*) \cdot \mathrm{Re}(x_*^H A_i X_{*\perp} z) \cdot \mathrm{Re}(x_*^H A_i X_{*\perp} y) = \langle\, y, \mathcal{L}(z)\,\rangle_{D_*};$$

[1] $\mathcal{L} \colon \mathbb{C}^m \to \mathbb{C}^m$ is called $\mathbb{R}$-linear if $\mathcal{L}(\alpha x + \beta y) = \alpha \mathcal{L}(x) + \beta \mathcal{L}(y)$ for all $\alpha, \beta \in \mathbb{R}$ and $x, y \in \mathbb{C}^m$.

i.e., $\mathcal{L}$ is self-adjoint w.r.t. the inner product $\langle\cdot,\cdot\rangle_{D_*}$ over $\mathbb{C}^{n-1}(\mathbb{R})$. Letting $y=z$, we can also show that $\mathcal{L}$ is positive semidefinite:

$$\langle\,\mathcal{L}(z),z\,\rangle_{D_*}=2\sum_{i=1}^{m}h_i'(x_*^HA_ix_*)\cdot\mathrm{Re}(x_*^HA_iX_{*\perp}z)^2\geq 0, \tag{2.9}$$

where we used the assumption that h_i is nondecreasing (so h_i' is nonnegative).

Now by the variational principle for the eigenvalues of self-adjoint operators (see, e.g., [63, Chap. 1]), the spectral radius

$$\rho(\mathcal{L})=\lambda_{\max}(\mathcal{L})=\max_{z\neq 0}\frac{\langle\,\mathcal{L}(z),z\,\rangle_{D_*}}{\langle\,z,z\,\rangle_{D_*}}. \tag{2.10}$$

Let $d=X_{*\perp}z$. Then we have

$$\langle\,z,z\,\rangle_{D_*}\equiv z^H(\lambda_*I_{n-1}-\Lambda_{*\perp})z=d^H(x_*^HH(x_*)x_*\cdot I_n-H(x_*))d, \tag{2.11}$$

where we used the identities $\lambda_*=x_*^HH(x_*)x_*$ and $H(x_*)X_{*\perp}=X_{*\perp}\Lambda_{*\perp}$. Therefore,

$$\rho(\mathcal{L})-1=\max_{z\neq 0}\frac{\langle\,\mathcal{L}(z),z\,\rangle_{D_*}-\langle\,z,z\,\rangle_{D_*}}{\langle\,z,z\,\rangle_{D_*}}\equiv\max_{z\neq 0,\ d=X_{*\perp}z}\frac{\varphi(d;x_*)}{\langle\,z,z\,\rangle_{D_*}},$$

where φ is from (2.7), and we used (2.9) for $\langle\,\mathcal{L}(z),z\,\rangle_{D_*}$ and (2.11) for $\langle\,z,z\,\rangle_{D_*}$. Consequently, $\rho(\mathcal{L})<1$ (or $\rho(\mathcal{L})\leq 1$) if and only if $\varphi(d;x_*)<0$ (or $\varphi(d;x_*)\leq 0$) for all $d=X_{*\perp}z$ with $z\neq 0$. Since $[X_{*\perp},x_*]$ is unitary, a vector $d=X_{*\perp}z$ for some $z\neq 0$ if and only if $d^Hx_*=0$ with $d\neq 0$. Results in items (a) and (b) follow. □

By the standard notion of the stability of fixed points of a mapping in the fixed-point analysis (see, e.g., [2, 13]), we can classify the stability of the eigenvectors of the mNEPv (1.1) using the spectral radius $\rho(\mathcal{L})$ and, alternatively, the characterization function φ in Lemma 2.1.

Definition 2.2. *Let $x_*\in\mathbb{C}^n$ be an eigenvector of the mNEPv* (1.1) *and φ be as defined in* (2.7). *Then x_* is a stable eigenvector if $\varphi(d;x_*)<0$ for all $d\neq 0$ and $d^Hx_*=0$, and x_* is a weakly stable eigenvector if $\varphi(d;x_*)\leq 0$ for all $d\neq 0$ and $d^Hx_*=0$. Otherwise, x_* is called a nonstable eigenvector.*

Note that Definition 2.2 does not explicitly require that $\lambda_*(H(x_*))$ is a simple eigenvalue, as the characteristic function φ (2.7) is still well-defined for nonsimple eigenvalues. In addition, we note that for a *stable eigenvector* x_*, the corresponding λ_* must be a simple eigenvalue of $H(x_*)$. Otherwise, there would exist another eigenvector $\widetilde{x}$ of $\lambda_*=\lambda_{\max}(H(x_*))$ orthogonal to x_*. By letting $d=\widetilde{x}$ and recalling $h_i'(t)\geq 0$, we derive from (2.7) that $\varphi(d;x_*)\geq 0$, which contradicts the condition for a stable eigenvector that $\varphi(d;x_*)<0$ for all $d\neq 0$ and $d^Hx_*=0$.

2.2. Characterization of mNEPv via aMax.

The following theorem provides a variational characterization of the mNEPv (1.1) through the aMax (1.3). Before stating the theorem, let us recall a standard optimization concept (see, e.g., [48, sect. 2.1]): A unit vector x is called a *local maximizer* of the aMax (1.3) if there exists $\varepsilon>0$ s.t.

$$F(x)\geq F\left(\frac{x+d}{\|x+d\|}\right)\quad\text{for all } d\in\mathbb{C}^n \text{ with } d^Hx=0 \text{ and } \|d\|\leq\varepsilon, \tag{2.12}$$

and x is a *strict local maximizer* if the inequality for F in (2.12) holds strictly.

THEOREM 2.3. *Let $x \in \mathbb{C}^n$ be a unit vector:*

(a) *If x is a stable eigenvector of the mNEPv* (1.1)*, then x is a strict local maximizer of the aMax* (1.3)*.*

(b) *If x is a local maximizer of the aMax* (1.3)*, then x is a weakly stable eigenvector of the mNEPv* (1.1)*.*

Proof. Let $\widehat{x} = (x+d)/\|x+d\|$. Then we have $\widehat{x}^H A_i \widehat{x} = x^H A_i x + \delta_i$ for $i = 1, 2, \ldots, m$, where

$$\delta_i := 2 \cdot \mathrm{Re}(d^H A_i x) + d^H \left(A_i - (x^H A_i x) I\right) d + \mathcal{O}(\|d\|^3). \tag{2.13}$$

Hence, by (1.3), the ith term of $F(\widehat{x})$ satisfies

$$\phi_i\left(\widehat{x}^H A_i \widehat{x}\right) = \phi_i(g_i(x) + \delta_i) = \phi_i(g_i(x)) + h_i(g_i(x)) \cdot \delta_i + \frac{1}{2} h_i'(g_i(x)) \cdot \delta_i^2 + \mathbf{o}(\delta_i^2),$$

where $g_i(x) := x^H A_i x$. Summing over all ϕ_i from $i = 1$ to m, we obtain

$$\begin{aligned} F(\widehat{x}) &\equiv \sum_{i=1}^{m} \left[\phi_i(g_i(x)) + h_i(g_i(x)) \cdot \delta_i + \frac{1}{2} h_i'(g_i(x)) \cdot \delta_i^2 + \mathbf{o}(\delta_i^2) \right] \\ &= F(x) + 2\,\mathrm{Re}(d^H H(x) x) + d^H \left(H(x) - s(x) I\right) d \\ &\quad + 2 \sum_{i=1}^{m} h_i'(g_i(x)) \cdot \left(\mathrm{Re}(d^H A_i x)\right)^2 + \mathbf{o}(\|d\|^2) \\ &= F(x) + 2\,\mathrm{Re}(d^H H(x) x) + \varphi(d; x) + \mathbf{o}(\,\|d\|^2), \end{aligned} \tag{2.14}$$

where the second equality is by (2.13) and $s(x) := x^H H(x) x$.

For item (a): We need to show that the inequality (2.12) holds strictly. By the NEPv $H(x)x = \lambda x$ and the orthogonality $d^H x = 0$, we have $d^H H(x) x = 0$. So (2.14) implies that

$$F(\widehat{x}) = F(x) + \varphi(d; x) + \mathbf{o}(\|d\|^2). \tag{2.15}$$

Since the stability of x (Definition 2.2) implies that $\varphi(d; x) < 0$ and we can drop $\mathbf{o}(\|d\|^2)$ (which is negligible to $\varphi(d; x) = \mathcal{O}(\|d\|^2)$), (2.15) leads to $F(x) > F(\widehat{x})$ as $\|d\| \to 0$.

For item (b): Let d be sufficiently tiny and $d^H x = 0$. It follows from the local maximality (2.12) and the expansion (2.14) that

$$0 \geq F(\widehat{x}) - F(x) = 2 \cdot \mathrm{Re}(d^H H(x) x) + \varphi(d; x) + \mathbf{o}(\|d\|^2). \tag{2.16}$$

Therefore, the leading first-order term must vanish, that is, $\mathrm{Re}(d^H H(x) x) = 0$ for all d with $d^H x = 0$. This implies that $H(x)x$ and x have common null spaces, i.e.,

$$H(x) x = \lambda x \quad \text{for some scalar } \lambda. \tag{2.17}$$

To show that x is a weakly stable eigenvector (Definition 2.2), we still need to prove that (i) λ in (2.17) is the largest eigenvalue of $H(x)$ and that (ii) $\varphi(d; x) \leq 0$ for all d with $d^H x = 0$. Condition (ii) follows from (2.16) by noticing that the first term on the right side vanishes due to (2.17) and that $\mathbf{o}(\|d\|^2)$ is negligible to the quadratic function $\varphi(d; x)$ as $\|d\| \to 0$. Condition (ii), in turn, also implies that λ is the largest eigenvalue of $H(x)$. Otherwise, there is a $\widetilde{\lambda} > \lambda$ with $H(x)\widetilde{x} = \widetilde{\lambda}\widetilde{x}$ and

$\widetilde{x}^H x = 0$. Recall (2.7) that $\varphi(d;x) \geq d^H(H(x) - (x^H H(x)x)I)d$. Letting $d = \widetilde{x}$, we have $\varphi(d;x) \geq \widetilde{\lambda} - \lambda > 0$, contradicting $\varphi(d;x) \leq 0$. □

Results from Theorem 2.3 can be regarded as second-order sufficient and necessary conditions for the aMax (1.3). They are stated in a way to highlight the connections between the local maximizers of the aMax and the stable eigenvectors of the mNEPv, which benefits the analysis of the SCF to be discussed in section 3. We note that the objective function $F(x)$ of the aMax is not holomorphic (i.e., complex differentiable in $x \in \mathbb{C}^n$). Therefore, second-order KKT conditions (see, e.g., [48, sect. 12.5]) are not immediately applicable. Note that turning the problem to a real variable optimization (in the real and imaginary parts of $x \in \mathbb{C}^n$) and then applying the KKT condition will not lead to Theorem 2.3 since there would be no strict local maximizers for the real problem due to the unitary invariance of $F(x)$.

To end this section, let us discuss three immediate implications of the variational characterization in Theorem 2.3:

(1) Given the intrinsic connection between the mNEPv (1.1) and the aMax (1.3), stable and weakly stable eigenvectors of the NEPv are of particular interest. Since the aMax always has a global (hence local) maximizer, Theorem 2.3(b) guarantees the existence of weakly stable eigenvectors. Although such eigenvectors may not be unique and may correspond to local but nonglobal maximizers of the aMax (see Example 6.1), the connection to the aMax greatly facilitates the design and analysis of algorithms for the mNEPv (1.1), such as a geometric interpretation of the SCF in section 3.

(2) Theorem 2.3 is a generalization of the well-known variational characterization of Hermitian eigenvalue problems. Consider the case of the mNEPv (1.1) with $m = 1$ and $h_1(t) = 1$, i.e., $A_1 x = \lambda x$. Let $\lambda \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues of A_1 with eigenvectors $[x, x_2, \ldots, x_n]$. Since any nonzero d orthogonal to $[x]$ can be written as $d = \alpha_2 x_2 + \cdots + \alpha_n x_n$ for some $\{\alpha_i\}_{i=2}^n$, the function φ defined in (2.7) becomes $\varphi(d;x) = d^H(A_1 - \lambda I)d = \sum_{i=2}^n \alpha_i^2(\lambda_i - \lambda)$. Hence, $\varphi(d;x)$ is nonpositive and is strictly negative if λ is simple. Then Theorem 2.3 can be paraphrased to the well-known variational characterization of Hermitian eigenvalue problems: Eigenvectors of the largest eigenvalue of A_1 are global maximizers of $(x^H A_1 x)/(x^H x)$. If the largest eigenvalue is simple, then its eigenvector (up to scaling) is the only maximizer; see, e.g., [1, sect. 4.6.2].

(3) If the matrices $\{A_i\}$ of the mNEPv (1.1) are real symmetric, then $H(x)$ is real symmetric, and the eigenvectors of the mNEPv are all real vectors (up to a unitary scaling). Theorem 2.3(b) implies that the global maximum of the aMax (1.3) is always achieved at a real vector $x \in \mathbb{R}^n$, namely,

$$\max_{x\in\mathbb{C}^n,\, x^H x=1} F(x) \quad = \max_{x\in\mathbb{R}^n,\, x^T x=1} F(x). \tag{2.18}$$

The two maximizations above are fundamentally different in nature. The identity holds only due to the specific formulation of F, as demonstrated by Theorem 2.3. We highlight the identity (2.18) because many practical optimization problems come in the form of the right-hand side with $x \in \mathbb{R}^n$. We can nevertheless view such a problem as an aMax (1.3) with $x \in \mathbb{C}^n$. This allows us to develop a unified treatment for both real and complex variables, which is highly beneficial, as shown in the case of numerical radius computation in subsection 5.1.

3. Geometry and global convergence of the SCF. Much like the power method for solving linear eigenvalue problems, SCF iteration is a gateway method for NEPv; see [42, 17] and references therein. For the mNEPv (1.1), the SCF starts from an initial unit vector $x_0 \in \mathbb{C}^n$ and generates a sequence of approximate eigenvectors $x_1, x_2 \ldots$, via sequentially solving the linear eigenvalue problems

$$H(x_k)x_{k+1} = \lambda_{k+1}\, x_{k+1} \quad \text{for } k = 0, 1, \ldots, \tag{3.1}$$

where λ_{k+1} is the largest eigenvalue of $H(x_k)$ and x_{k+1} is a unit eigenvector. In the following, we first present a geometric interpretation of the SCF (3.1) and then provide a proof of the global convergence of the SCF based on the geometric observation.

3.1. Geometry of the SCF. In subsection 2.2, we discussed the variational characterization of the mNEPv (1.1) via the aMax (1.3). Now consider the change of variables

$$y = g(x) \quad \text{with} \quad g(x) := \left[x^H A_1 x, \ldots, x^H A_m x\right]^T \in \mathbb{R}^m. \tag{3.2}$$

The aMax (1.3) is then recast as an optimization over the joint numerical range

$$\max_{y \in W(\mathcal{A})} \left\{ \phi(y) := \sum_{i=1}^{m} \phi_i(\, y(i)) \right\}, \tag{3.3}$$

where $y(i)$ is the ith entry of y and $W(\mathcal{A}) \subset \mathbb{R}^m$ is a (first) *joint numerical range* of an m-tuple $\mathcal{A} := (A_1, \ldots, A_m)$ of Hermitian matrices $A_1, \ldots, A_m$ defined as

$$W(\mathcal{A}) = \left\{ y \in \mathbb{R}^m \mid y = g(x),\, x \in \mathbb{C}^m,\, \|x\| = 1 \right\}. \tag{3.4}$$

By definition, $W(\mathcal{A})$ is the range of the vector-valued function g over the unit sphere $\{x \in \mathbb{C}^n \mid \|x\| = 1\}$. Since g is a continuous and bounded function, $W(\mathcal{A})$ is a connected and bounded subset of $\mathbb{R}^m$. Moreover, it is known that the set of $W(\mathcal{A})$ is convex in cases such as $m = 1, 2$ for any matrix size n, $m = 3$ for $n \geq 3$ [3, 4], and other cases under proper conditions [36].

Before we proceed, let us first revisit the notion of supporting hyperplane for a general bounded and closed subset Ω of $\mathbb{R}^m$. To this end, we can define a hyperplane

$$\mathcal{P}_v := \left\{ y \in \mathbb{R}^m \mid v^T(y - y_v) = 0 \right\}, \tag{3.5}$$

where v is a given nonzero vector in $\mathbb{R}^m$ and y_v satisfies

$$y_v \in \operatorname*{argmax}_{y \in \Omega} v^T y. \tag{3.6}$$

The hyperplane $\mathcal{P}_v$ contains in one of its half-spaces the entire Ω, and it also passes through at least one point in Ω because

$$\text{(i)} \quad v^T y \leq v^T y_v \text{ for all } y \in \Omega \quad \text{and} \quad \text{(ii)} \quad y_v \in \Omega. \tag{3.7}$$

We will refer to $\mathcal{P}_v$ as a *supporting hyperplanes* of Ω with an outer normal vector v (pointing outward from Ω) and a supporting point y_v. Supporting hyperplanes are commonly used for studying convex sets; see, e.g., [15, sect. 2.5].

Finding the global optimizer in (3.6) for a general set Ω is hard. Fortunately, if the set $\Omega = W(\mathcal{A})$, then the following lemma shows that the supporting point y_v in (3.6) can be obtained by solving a Hermitian eigenvalue problem.

LEMMA 3.1. *Let $v \in \mathbb{R}^m$ be a nonzero vector. Then*

$$y_v \in \operatorname*{argmax}_{y \in W(\mathcal{A})} v^T y \quad \textit{if and only if} \quad y_v = g(x_v), \tag{3.8}$$

where x_v is an eigenvector for the largest eigenvalue λ_v of the Hermitian matrix

$$H_v := \sum_{i=1}^{m} v(i) \cdot A_i \tag{3.9}$$

and $v(i)$ is the ith entry of v.

Proof. Observe that

$$v^T g(x) = \sum_{i=1}^{m} (x^H A_i x) \cdot v(i) = x^H H_v x. \tag{3.10}$$

The maximization from (3.8) leads to

$$\max_{y \in W(\mathcal{A})} v^T y = \max_{\|x\|=1} v^T g(x) = \max_{\|x\|=1} x^H H_v x = x_v^H H_v x_v = v^T g(x_v),$$

where the second and the last equalities are due to (3.10) and the third equality is by the eigenvalue maximization principle of Hermitian matrices; namely, the maximizer of $x^H H_v x$ is achieved at any eigenvector x_v of the largest eigenvalue of H_v. □

Lemma 3.1 suggests a close relation between the SCF (3.1) and the search for supporting points of $W(\mathcal{A})$. Such relation is called a geometric interpretation of the SCF and is formally stated in the following theorem.

THEOREM 3.2. *Let $\{x_k\}$ be a sequence of unit vectors generated by the SCF* (3.1), *and let $y_k := g(x_k)$, where g is defined in* (3.2). *Then it holds that*

$$y_{k+1} \in \operatorname*{argmax}_{y \in W(\mathcal{A})} \nabla\phi(y_k)^T y. \tag{3.11}$$

Therefore, geometrically,

$$y_{k+1} \textit{ is a supporting point of } W(\mathcal{A}) \textit{ for the outer normal vector } \nabla\phi(y_k). \tag{3.12}$$

Proof. The coefficient matrix $H(x_k)$ by (1.2) is an H_v matrix in Lemma 3.1:

$$H(x_k) \equiv H_{v_k} \qquad \text{with} v_k = \nabla\phi(y_k) \text{ and } y_k = g(x_k) \in W(\mathcal{A}). \tag{3.13}$$

Hence, the kth SCF iteration (3.1) is to solve the eigenproblem $H_{v_k} x_{k+1} = \lambda_{k+1} x_{k+1}$. It follows from Lemma 3.1 that $y_{k+1} = g(x_{k+1})$ is a solution of (3.8) for $v_k = \nabla\phi(y_k)$. Therefore, y_{k+1} is a supporting point of $W(\mathcal{A})$ for the outer normal direction $\nabla\phi(y_k)$. □

By Theorem 3.2, the SCF iteration (3.1) can be visualized as searching the solution of the mNEPv (1.1) on the boundary of the joint numerical range $W(\mathcal{A})$. Moreover, at a solution x_* of the mNEPv (1.1), the geometric interpretation (3.12) is equivalent to the following geometric first-order optimality condition for the constrained optimization (3.3):

$$\nabla\phi(y_*) \text{ is an outer normal vector of } W(\mathcal{A}) \text{ at } y_*, \tag{3.14}$$

where $y_* = g(x_*)$. These concepts are illustrated by the example below.

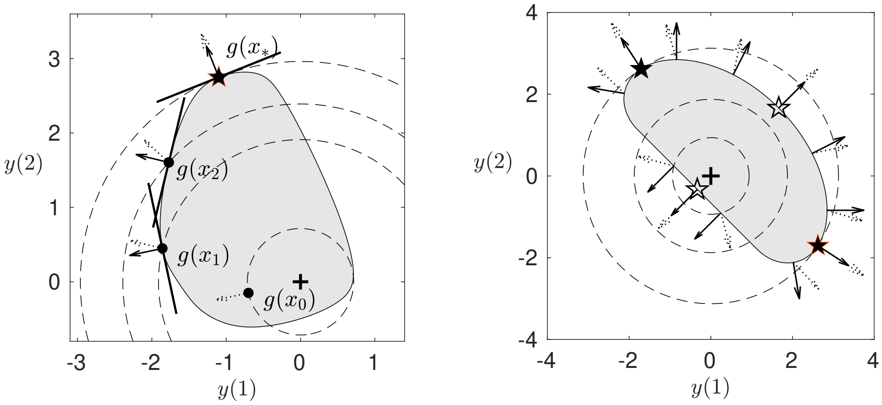

Fig. 1. *Left: Illustration of Example* 3.3 *for the first three iterates* x_0, x_1, x_2 *by the SCF* (3.1) *for the mNEPv* (3.15). *The shaded region is the joint numerical range* $W(A_1, A_2)$*; dashed lines are contours of* $\phi(y) = \|y\|^2/2$ *with dashed arrows the gradient directions* $\nabla\phi$*; solid tangent lines are "supporting hyperplanes" at* $y_i = g(x_i)$ *with solid arrows the normal direction* $\nabla\phi(y_{i-1})$*; the maximizer of* (3.16) *is marked as* ★. *Right: Illustration of Example* 3.6 *for stable eigenvectors, marked as solid stars* ★*, and nonstable eigenvectors, marked as hollow stars* ☆*: Close to a nonstable eigenvector, the gradients* $\nabla\phi$ *(dashed arrows) point away from the normal vectors (solid arrow), leading to divergence of the SCF from* ☆.

Example 3.3. Let us consider the mNEPv (1.1) of the form

$$H(x)x = \lambda x \quad \text{with} \quad H(x) = (x^H A_1 x)\cdot A_1 + (x^H A_2 x)\cdot A_2, \tag{3.15}$$

where A_1 and A_2 are Hermitian matrices. The mNEPv (3.15) arises from numerical radius computation and will be further discussed in subsection 5.1. By Theorem 2.3 and (3.3), the mNEPv (3.15) can be characterized by the optimization problems

$$\max_{\|x\|=1} \left\{ F(x) := [(x^H A_1 x)^2 + (x^H A_2 x)^2]/2 \right\} = \max_{y\in W(A_1,A_2)} \left\{ \phi(y) := \|y\|^2/2 \right\}, \tag{3.16}$$

where $W(A_1, A_2)$ is a joint numerical range of A_1 and A_2. The left plot in Figure 1 depicts the SCF as a search process for solving the mNEPv (3.15) with randomly generated Hermitian matrices A_1 and A_2 of size 10. Given the initial $y_0 = g(x_0)$, the SCF first searches in the gradient direction $v_0 = \nabla\phi(y_0)$ to obtain a supporting point $y_1 = g(x_1)$ and then searches in the gradient direction $\nabla\phi(y_1)$ to obtain the second supporting point $y_2 = g(x_2)$ and so on. When this process converges to $y_* = g(x_*)$, the gradient $\nabla\phi(y_*)$ overlaps the outer normal vector of $W(\mathcal{A})$ at y_*; i.e., the optimality condition (3.14) is achieved.

Another key indication of (3.11) is that the SCF is a *successive local linearization* for the optimization (3.3): At iteration k, it approximates $\phi(y)$ by its first-order expansion

$$\ell_k(y) := \phi(y_k) + \nabla\phi(y_k)^T(y - y_k) \tag{3.17}$$

and solves the optimization of the linear function over the joint numerical range

$$\max_{y\in W(\mathcal{A})} \ell_k(y). \tag{3.18}$$

By dropping the constant terms in $\ell_k(y)$, the maximizers of (3.18) satisfy

$$\operatorname*{argmax}_{y\in W(\mathcal{A})} \ \ell_k(y) \equiv \operatorname*{argmax}_{y\in W(\mathcal{A})} \ \nabla\phi(y_k)^T y.$$

Hence, the solution to (3.18) is exactly y_{k+1} in (3.11), and we have

$$\ell_k(y_{k+1}) = \max_{y\in W(\mathcal{A})} \ell_k(y). \tag{3.19}$$

These observations are helpful to the proof of the global convergence of the SCF to be presented in subsection 3.2.

3.2. Convergence analysis of the SCF. In this section, we show that the SCF iteration is globally convergent to an eigenvector of the mNEPv (1.1), as indicated by the visualization of the SCF in subsection 3.1. Moreover, the converged eigenvector is typically a stable one, and the rate of convergence is at least linear.

We begin with the following theorem on the global convergence of the SCF (3.1). Here for a sequence of unit vectors $\{x_k\}$, we call x_* an (entrywise) limit point if

$$x_* = \lim_{j\to\infty} x_{k_j} \text{ for some subsequence } \{x_{k_j}\} \text{ indexed by } k_1 < k_2 < \cdots. \tag{3.20}$$

By the Bolzano–Weierstrass theorem, a bounded sequence in $\mathbb{C}^n$ has a convergent subsequence, so the sequence $\{x_k\}$ of unit vectors has at least one limit point x_*.

THEOREM 3.4. *Let $\{x_k\}$ be a sequence of unit vectors from the SCF* (3.1) *for the mNEPv* (1.1) *and $F(x)$ be the objective function of the aMax* (1.3)*. Then*

(a) *$F(x_{k+1}) \geq F(x_k)$ for $k = 0,1,\ldots,$ with equality holding only if x_k is an eigenvector of the mNEPv* (1.1)*;*

(b) *each limit point x_* of $\{x_k\}$ must be an eigenvector of the mNEPv* (1.1)*, and it holds that $F(x_*) \geq F(x_k)$ for all $k \geq 0$.*

Proof. For item (a), recall that the linearization ℓ_k in (3.17) is a lower supporting function for the convex function ϕ, i.e., $\ell_k(y) \leq \phi(y)$ for $y \in W(\mathcal{A})$. Consequently,

$$F(x_{k+1}) \equiv \phi(y_{k+1}) \geq \ell_k(y_{k+1}) = \max_{y\in W(\mathcal{A})} \ell_k(y) \geq \ell_k(y_k) = \phi(y_k) \equiv F(x_k), \tag{3.21}$$

where the third equality is by (3.19). Moreover, if the equality $F(x_{k+1}) = F(x_k)$ holds, then (3.21) implies that

$$\ell_k(y_k) = \max_{y\in W(\mathcal{A})} \ell_k(y), \tag{3.22}$$

namely,

$$y_k \in \operatorname*{argmax}_{y\in W(\mathcal{A})} \ell_k(y) \equiv \operatorname*{argmax}_{y\in W(\mathcal{A})} \nabla\phi(y_k)^T y.$$

According to Lemma 3.1, $y_k = g(x_k)$, and x_k is an eigenvector for the largest eigenvalue of H_{v_k} with $v_k = \nabla\phi(y_k)$. Since $H_{v_k} \equiv H(x_k)$, we have $H(x_k)x_k = \lambda x_k$, and λ is the largest eigenvalue; i.e., x_k is an eigenvector of the mNEPv (1.1).

For item (b), let $\{x_{k_j}\}$ be a subsequence of $\{x_k\}$ convergent to x_*. The monotonicity from item (a) implies that $F(x_*) \geq F(x_k)$ for all $k \geq 0$. To show that x_* is an eigenvector, we denote by $y_{k_j} = g(x_{k_j})$ and $y_* = g(x_*)$. The linearization of ϕ at y_* satisfies

$$\ell_*(y) := \phi(y_*) + \nabla\phi(y_*)^T(y - y_*) = \lim_{j\to\infty} \ell_{k_j}(y), \tag{3.23}$$

where the last equality is due to (3.17), $y_* = \lim_{j\to\infty} y_{k_j}$, and the continuity of ϕ and $\nabla\phi$.

We first show that

$$\nabla\phi(y_*)^T(y-y_*)\leq 0 \quad \text{for all } y\in W(\mathcal{A}). \tag{3.24}$$

Otherwise, there exists a $\widetilde{y}\in W(\mathcal{A})$ with

$$\varepsilon := \nabla\phi(y_*)^T(\widetilde{y}-y_*)>0. \tag{3.25}$$

By the convergence of $\ell_{k_j}\to\ell_*$ in (3.23), there exists $N\geq 0$ such that for all $j\geq N$,

$$\ell_{k_j}(\widetilde{y})\geq \ell_*(\widetilde{y})-\varepsilon/2. \tag{3.26}$$

It then follows from (3.21) (with $k=k_j$) that for all $j\geq N$,

$$\phi(y_{k_j+1})\geq \max_{y\in W(\mathcal{A})}\ell_{k_j}(y) \;\geq\; \ell_{k_j}(\widetilde{y}) \;\geq\; \ell_*(\widetilde{y})-\frac{\varepsilon}{2}=\phi(y_*)+\frac{\varepsilon}{2},$$

where the last two equations are due to (3.26) and (3.25). The equation above implies that $F(x_{k_j+1})\geq F(x_*)+\varepsilon/2$, contradicting $F(x_*)\geq F(x_k)$ for all k.

It follows from (3.23) and (3.24) that

$$\ell_*(y_*)=\max_{y\in W(\mathcal{A})}\ell_*(y)=\phi(y_*).$$

Then by the same arguments as for the y_k in (3.22), we have that x_* is an eigenvector of the mNEPv (1.1). □

In section 5, we will discuss the mNEPv (1.1) arising from optimization of the form (1.3), for which the monotonicity of the objective function is highly desirable. Starting from any x_0, the SCF will find an eigenvector x_* that has an increased function value $F(x_*)\geq F(x_0)$.

Let's now consider the local convergence properties of the SCF. Theorem 3.4 guarantees that the SCF will converge globally to some eigenvector of the mNEPv (1.1) from any initial guess x_0. In theory, the SCF may terminate at a nonstable eigenvector x_* of the mNEPv, if it exists. In practice, however, convergence to a nonstable eigenvector is unlikely to happen because such eigenvectors are repulsive fixed points of the mapping Π (2.2), as explained in subsection 2.1. Therefore, the SCF (3.1), which is a fixed point iteration with Π, will diverge from a nonstable x_* when x_k is in a neighborhood of x_*. More rigorously, by the local convergence analysis of the SCF for a general unitarily invariant NEPv (see [7, Thm. 1]), we can draw the local convergence of the SCF (3.1) for the mNEPv (1.1), as stated in the following theorem.

THEOREM 3.5. *Let x_* be an eigenvector of the mNEPv* (1.1) *with a simple eigenvalue λ_*, $\mathcal{L}$ be the $\mathbb{R}$-linear operator* (2.4) *for x_*, and $\rho(\mathcal{L})$ be the spectral radius:*

(a) *If $\rho(\mathcal{L})<1$ (i.e., x_* is a stable eigenvector by Definition* 2.2*), then the SCF* (3.1) *is locally convergent to x_*, with an asymptotic convergence rate bounded by $\rho(\mathcal{L})$.*

(b) *If $\rho(\mathcal{L})>1$ (i.e., x_* is a nonstable eigenvector by Definition* 2.2*), then the SCF is locally divergent from x_*.*

Here we recall that an iterate x_k by the SCF (3.1) is understood as a one-dimensional subspace spanned by x_k. The local convergence and divergence of x_k in Theorem 3.5 is measured by the vector angle $\angle(x_*,x_k):=\cos^{-1}\left(|x_*^Hx_k|\right)$.

Example 3.6. By the geometric interpretation of the SCF from Theorem 3.2, we can visualize its local convergence behavior revealed in Theorem 3.5. The right plot in Figure 1 depicts the search directions of the SCF for a numerical radius problem described in (3.16), with the corresponding mNEPv (3.15). There are four eigenvectors (marked as stars, where the solid and dashed arrows overlap). Two solid stars are stable eigenvectors (i.e., local maximizers of (3.16)), and two hollow stars are nonstable eigenvectors (nonmaximizers). The reason why the SCF is locally convergent to stable eigenvectors is now clear: Close to a solid star, the search directions $\nabla\phi(y)$ by (3.12) (dashed arrow) bring the next iteration closer to the solid star. In contrast, close to a hollow star, the search directions lead away from the hollow star. This observation also justifies the name of *nonstable eigenvector* since a slight perturbation will lead the SCF to diverge from those solutions.

Combining the properties of global and local convergence in Theorems 3.4 and 3.5, we can summarize the overall convergence of the SCF (3.1) as follows:

1. Let x_* be an (entrywise) limit point of $\{x_k\}$ by the SCF. Then x_* is an eigenvector of the mNEPv (1.1); see Theorem 3.4(b).
2. The limit point x_* is *unlikely* a nonstable eigenvector since the SCF is locally divergent from nonstable eigenvectors; see Theorem 3.5(b).[2] Consequently, the SCF is expected to converge to (at least) a weakly stable eigenvector x_*.
3. If the limit point x_* is a stable eigenvector, then the SCF is at least locally linearly convergent to x_*; see Theorem 3.5(a).

4. The SCF in practice.

In this section, we will introduce an acceleration technique and discuss related implementation details of the SCF iteration.

4.1. Accelerated SCF.

The iterative process (3.1) is an SCF in its simplest form, also known as the plain SCF. There are a number of ways to accelerate the plain SCF, such as the damping scheme [19], level-shifting [64], direct inversion of iterative subspace with Anderson acceleration [51], and the preconditioned fixed-point iteration [39]. Most of these schemes are designed for solving NEPv from electronic structure calculations. In this section, we present an acceleration scheme of the SCF (3.1) for the mNEPv (1.1) based on the inverse iteration.

Inverse iterations are a commonly used technique for solving linear eigenvalue problems [30] and eigenvalue-dependent nonlinear eigenvalue problems [25]. Moreover, there is also an inverse iteration available for NEPv in the form

$$H(x/\|x\|)\cdot x = \lambda x, \tag{4.1}$$

where $H(x)$ is a real symmetric matrix that is differentiable in $x \in \mathbb{R}^n$ [31].[3] For normalized x, we have $H(x/\|x\|) \equiv H(x)$, so that the mNEPv (1.1) can be equivalently written to an NEPv (4.1). In the following, we will first revisit the inverse iteration scheme in [31] and then propose an improved scheme for solving the mNEPv (1.1) by exploiting its underlying structure.

Let x_k be a unit approximate eigenvector of the NEPv (4.1) and σ_k be a given shift close to a target eigenvalue. The following inversion step is proposed in [31] to improve x_k:

$$\widetilde{x}_k = \alpha_k\,(J(x_k) - \sigma_k I)^{-1}\,x_k \quad \text{with} \quad J(x) := \frac{\partial}{\partial x}(H(x/\|x\|)x), \tag{4.2}$$

[2] One exceptional but rare case is that some x_k coincides with a nonstable x_* and that the SCF terminates.

[3] The authors in [31] considered scaling invariant NEPv $H(x)\cdot x = \lambda x$ with $H(x) \equiv H(\alpha x)$ for all $\alpha \neq 0$, and they pointed out such NEPv cover (4.1) as a special case.

where α_k is a normalization factor. The formula (4.2) can be derived from Newton's method applied to the nonlinear equations $H(x/\|x\|)x - \lambda x = 0$ and $x^T x = 1$. Iteratively applying (4.2) with a fixed shift σ has been proven to converge linearly with a convergence factor proportional to $|\sigma - \lambda_*|$, whereas using dynamic Rayleigh shifts $\sigma_k = x_k^T H(x_k) x_k$ is expected to yield quadratic convergence [31]. However, directly applying the inverse iteration (4.2) may lead to convergence to an eigenvalue that is not the largest one. Hence, we will only use it as a local acceleration scheme for the SCF.

We first note that despite the matrix $H(x)$ of the mNEPv (1.1) being symmetric when all coefficient matrices $A_1, \ldots, A_m$ are real symmetric, the corresponding Jacobian $J(x)$ in (4.2) is generally not. Specifically, the Jacobian $J(x)$ is given by

$$J(x) \equiv \frac{\partial}{\partial x}\left(H(x/\|x\|)x \right) = H(x) + 2M(x)C(x)M(x)^T P(x), \tag{4.3}$$

where $M(x) = [A_1 x, \cdots, A_m x]$ and $C(x) = \mathrm{Diag}\left(h_1'\left(x^T A_1 x\right), \ldots, h_m'\left(x^T A_m x\right)\right)$ and $P(x) = I - xx^T$ is a projection matrix. To symmetrize $J(x)$, we introduce

$$J_{\mathrm{s}}(x) := J(x) + x \cdot q(x)^T = H(x) + 2\,P(x)M(x)C(x)M(x)^T P(x), \tag{4.4}$$

where $q(x) = 2P(x)M(x)C(x)M(x)^T x \in \mathbb{R}^n$. Since the new matrix J_s is a rank-one modification of J, by the Sherman–Morrison–Woodbury formula [28], we have

$$\left(J_{\mathrm{s}}(x_k) - \sigma_k I\right)^{-1} x_k = c \cdot \left(J(x_k) - \sigma_k I\right)^{-1} x_k$$

for some constant c. Therefore, we can reformulate the inversion step (4.2) to

$$\widetilde{x}_k = \widetilde{\alpha}_k \cdot \left(J_{\mathrm{s}}(x_k) - \sigma_k I\right)^{-1} x_k, \tag{4.5}$$

where $\widetilde{\alpha}_k$ normalizes $\widetilde{x}_k$ to a unit vector; i.e., we can replace J by the symmetric J_s.

If the coefficient matrices $\{A_i\}$ are complex Hermitian, then $H(x)$ is not holomorphically differentiable since its diagonal entries are always real and cannot be analytic functions. Consequently, the (holomorphic) Jacobian of $H(x/\|x\|)x$ does not exist. Nevertheless, the matrix $J_{\mathrm{s}}(x)$ by (4.4) is well-defined and Hermitian (with transpose $\cdot^T$ replaced by conjugate transpose $\cdot^H$), so it can still be used for the inversion (4.5).

4.2. Implementation issues.

The SCF with an optional acceleration for solving the mNEPv (1.1) is summarized in Algorithm 4.1. A few remarks on the implementation detail are in order:

(1) The initial x_0, in view of the geometry of the SCF discussed in subsection 3.1, can be chosen from sampled supporting points of $W(\mathcal{A})$. To do this, we randomly choose ℓ search directions $v_i \in \mathbb{R}^m$ for $i = 1, \ldots, \ell$ and then find the supporting points $y_{v_i} = g(x_{v_i})$ of $W(\mathcal{A})$ along each direction. Among $x_{v_1}, \ldots, x_{v_\ell}$, we choose the one with the largest value $F(x_{v_i})$ as x_0. This greedy sampling scheme increases the chance for the SCF to find the global maximizer of the aMax (1.3).

To compute the supporting points, Lemma 3.1 tells us that x_{v_i} is an eigenvector to the largest eigenvalue of the Hermitian matrix H_{v_i} in (3.9). Thus, we need to solve ℓ Hermitian eigenvalue problems to obtain ℓ supporting points. For efficiency, we can exploit the fact that $H_{-v_i} \equiv -H_{v_i}$, so we can compute two supporting points in both directions $\pm v_i$ by solving a single eigenvalue problem of H_{v_i}.

Algorithm 4.1 The SCF with optional acceleration

```
Input:  Starting $x_0 \in \mathbb{C}^n$, residual tolerance tol, and acceleration threshold $\mathrm{tol}_{\mathrm{acc}}$.
Output:  Approximate eigenpair $(\lambda_k, x_k)$ of the mNEPv (1.1).
1: for $k = 1, 2, \ldots$ do
2:     $H(x_{k-1})\, x_k = \lambda_k \cdot x_k$ with $\lambda_k = \lambda_{\max}(H(x_{k-1}))$;            % SCF
3:     if $\mathrm{res}(x_k) \leq \mathrm{tol}$, then return $(\lambda_k, x_k)$;            % test for convergence
4:     if $\mathrm{res}(x_k) \leq \mathrm{tol}_{\mathrm{acc}}$ then            % acceleration if activated
5:         compute $\widetilde{x}_k$ by (4.5) with the shift $\sigma_k = x_k^H H(x_k) x_k$.
6:         if $F(\widetilde{x}_k) > F(x_k)$, then update $x_k = \widetilde{x}_k$;
7:     end if
8: end for
```

(2) Algorithm 4.1 requires finding the eigenvector corresponding to the largest eigenvalue of the matrix $H(x_{k-1})$ in line 2. Additionally, when we apply acceleration, we need to solve a linear system with coefficient matrix $J_{\mathrm{s}}(x_k) - \sigma_k I$ in line 5. For the mNEPv of small to medium sizes, direct solvers can be applied, such as the QR algorithm for Hermitian eigenproblems and LU factorization for linear systems (e.g., MATLAB's `eig` and backslash, respectively). For large sparse problems, iterative solvers are applied, such as the Lanczos-type methods for Hermitian eigenproblems (e.g., MATLAB's `eigs`) and MINRES and SYMMLQ for linear systems; see, e.g., [6, 10].

(3) The acceleration with the inverse iteration is expected to work well for x_k close to a solution. A threshold $\mathrm{tol}_{\mathrm{acc}}$ is introduced to control the activation of inverse iteration in line 4. If $\mathrm{tol}_{\mathrm{acc}} = 0$, Algorithm 4.1 runs the plain SCF. If $\mathrm{tol}_{\mathrm{acc}} = \infty$, Algorithm 4.1 applies acceleration at each step. We observe that the choice of $\mathrm{tol}_{\mathrm{acc}}$ is not critical, and $\mathrm{tol}_{\mathrm{acc}} = 0.1$ is used in our numerical experiments.

(4) To maintain the monotonicity of $F(x_k)$, as in the SCF, the accelerated eigenvector $\widetilde{x}_k$ is accepted only if $F(\widetilde{x}_k) \geq F(x_k)$ in line 6.

(5) To assess the accuracy of iteration k in line 3, we use the relative residual norm

$$\mathrm{res}(\widehat{x}) := \|H(\widehat{x})\widehat{x} - (\widehat{x}^H H(\widehat{x})\widehat{x}) \cdot \widehat{x}\| / \|H(\widehat{x})\|, \tag{4.6}$$

where $\|H(\widehat{x})\|$ is some convenient-to-evaluate matrix norm, e.g., the matrix 1-norm as we used in the experiments.

5. Applications. The mNEPv (1.1) and the aMax (1.3) can be found in numerous applications. In this section, we discuss three of them. The first one is on the quartic maximization over the Euclidean sphere and its application for computing numerical radius. The second is on the best rank-one approximation of third-order partial-symmetric tensors. The third is from the study of the distance to singularity of dHADE systems.

5.1. Quartic maximization and numerical radius. A *(homogeneous) quartic maximization* over the Euclidean sphere is of the form

$$\max_{x \in \mathbb{C}^n,\, \|x\|=1} \left\{ F(x) := \frac{1}{2} \sum_{i=1}^{m} \left(x^H A_i x \right)^2 \right\}, \tag{5.1}$$

where $\{A_i\}$ are n-by-n Hermitian matrices. The optimization (5.1) is a classical problem in the field of polynomial optimization, although in the literature, it is usually formulated in real variables, i.e., $x \in \mathbb{R}^n$ with symmetric $\{A_i\}$ [27, 46, 70]. In addition, it also arises in the study of robust optimization with ellipsoid uncertainty [12]. Observe that the quartic maximization (5.1) is an aMax (1.3) with $\{\phi_i(t) = t^2/2\}$. Hence, the underlying mNEPv (1.1) is of the form

$$H(x)\,x = \lambda x \quad \text{with} \quad H(x) = \sum_{i=1}^{m} (x^H A_i x) \cdot A_i, \tag{5.2}$$

where the coefficient functions $h_i(t) = \phi_i'(t) = t$ are differentiable and nondecreasing.

The simplest nontrivial example of the quartic optimization (5.1) is when $m = 2$, which occurs in the well-known problem of computing the numerical radius of a square matrix. The *numerical radius* of a matrix $B \in \mathbb{C}^{n\times n}$ is defined as

$$r(B) := \max_{x\in\mathbb{C}^n,\,\|x\|=1} |x^H B x| = \max_{x\in\mathbb{C}^n,\,\|x\|=1} \Big((x^H A_1 x)^2 + (x^H A_2 x)^2 \Big)^{1/2}, \tag{5.3}$$

where $A_1 = \frac{1}{2}(B^H + B)$ and $A_2 = \frac{\imath}{2}(B^H - B)$ with $\imath = \sqrt{-1}$ are Hermitian matrices [28]. An extension of (5.3) is the *joint numerical radius* of an m-tuple of Hermitian matrices $\mathcal{A} = (A_1, \ldots, A_m)$ defined as

$$r(\mathcal{A}) := \max_{x\in\mathbb{C}^n,\,\|x\|=1} \left(\sum_{i=1}^{m} (x^H A_i x)^2 \right)^{1/2}; \tag{5.4}$$

see [22]. The (joint) numerical radius plays important roles in numerical analysis. For examples, the numerical radius of a matrix is applied to quantify the transient effects of discrete-time dynamical systems and analyze classical iterative methods [5, 56]. The joint numerical radius of a matrix tuple is used for studying the joint behavior of several operators; see [35] and references therein.

Numerical algorithms for computing the numerical radius of a single matrix have been extensively studied [26, 44, 45, 58, 62]. To find the global maximizer of (5.3), many methods adopt the scheme of local optimization followed by global certification. Most of those algorithms, however, do not immediately extend to computing the joint numerical radius with $m \geq 3$. A major benefit of the NEPv approach presented in this paper is to allow fast computation of the local maximizers to accelerate existing approaches. Moreover, the NEPv approach provides a unified treatment for matrix tuple $\mathcal{A}$ with m matrices and can serve as the basis for future development of algorithms toward the global solution of $r(\mathcal{A})$ with $m \geq 3$.

5.2. Best rank-one approximation of third-order partial-symmetric tensors. Let $T \in \mathbb{R}^{n\times n\times m}$ be a third-order partial-symmetric tensor; i.e., each slice $A_i := T(:,:,i) \in \mathbb{R}^{n\times n}$ is symmetric for $i = 1, \ldots, m$. The problem of the best rank-one partial-symmetric tensor approximation is defined by the minimization

$$\min_{\substack{\mu\in\mathbb{R},\,x\in\mathbb{R}^n,\,z\in\mathbb{R}^m\\ \|x\|=1,\|z\|=1}} \|T - \mu \cdot x \otimes x \otimes z\|_F^2, \tag{5.5}$$

where $\otimes$ is the Kronecker product. The solution of (5.5) provides a rank-one partial-symmetric tensor $\mu_* \cdot x_* \otimes x_* \otimes z_*$ that best approximates T in the Frobenius norm $\|\cdot\|_F$ and is also known as a truncated rank-one CP decomposition of T; see, e.g., [33, 70].

The best rank-one approximations (5.5) are often recast as quartic maximizations (5.1); see, e.g., [21, eq. (6)]. Let x_i denote the ith element of a vector x. Then

$$\|T - \mu \cdot x \otimes x \otimes z\|_F^2 = \|T\|_F^2 + \mu^2 - 2\mu \sum_{i,j,k} t_{ijk} x_i x_j z_k, \tag{5.6}$$

where the range of indices i, j, k is omitted in the summation for clarity. Since the minimum w.r.t. μ is achieved at $\mu = \sum_{i,j,k} t_{ijk} x_i x_j z_k$, the best rank-one approximation (5.5) becomes the maximization

$$\max_{\substack{\|x\|=1\\ \|z\|=1}} \left(\sum_{i,j,k} t_{ijk} x_i x_j z_k \right)^2 = \max_{\substack{\|x\|=1\\ \|z\|=1}} \left(\sum_k z_k \cdot x^T A_k x \right)^2 = \max_{\|x\|=1} \sum_k \left(x^T A_k x \right)^2, \tag{5.7}$$

where the first equality is by $A_i = T(:,:,i)$ and the second equality is due to the maximization w.r.t. z being solved at

$$z = \alpha \cdot g(x) \equiv \alpha \cdot [x^T A_1 x, \ldots, x^T A_m x]^T \tag{5.8}$$

with α being a normalization factor for $\|z\| = 1$ provided that $g(x) \neq 0$. The formula of z in (5.8) follows from $|z^T g(x)|^2 \leq \|g(x)\|^2$ with equality holding if $z = g(x)/\|g(x)\|$.

Problem (5.7) leads to a quartic maximization (5.1) with real symmetric matrices $\{A_i\}$ and real variables $x \in \mathbb{R}^n$, i.e., an aMax (1.3) with $\{\phi_i(t) = t^2/2\}$. By Theorem 2.3, the optimizer x_* is an eigenvector of the mNEPv (5.2) with $h_i(t) = \phi'(t) = t$, and the corresponding eigenvalue is

$$\lambda_* = x_*^T H(x_*) x_* = \sum_k \left(x_*^T A_k x_* \right)^2 = \mu_*^2. \tag{5.9}$$

Any other eigenvalue λ of (5.2) must satisfy $\lambda \equiv x^T H(x) x = \sum_k \left(x^T A_k x \right)^2 \leq \lambda_*$ due to (5.9) and maximization (5.7).

The best rank-one approximation is a fundamental problem in tensor analysis; see [23, 32, 69]. Third-order partial-symmetric tensors are intensively studied [20, 37, 53, 70] and found in applications such as crystal structure [21, 49] and social networks (Example 6.4). It is known that tensor rank-one approximation problems are closely related to tensor eigenvalue problems [53], such as the *Z-eigenvalue* [52] and ℓ^2-eigenvalue [38] for general supersymmetric tensors and the *C-eigenvalue* for third-order partial-symmetric tensors [21]. Tensor eigenvalue problems provide first-order optimality conditions for the best rank-one approximation. But those eigenvalue problems are neither formulated nor studied through the NEPv as presented in this paper. For a third-order partial-symmetric tensor, its largest C-eigenvalue μ_* and the corresponding C-eigenvectors (x_*, z_*) form the best rank-one approximation (5.5) [21]. However, solving the tensor C-eigenvalue problems, which involve two coupled nonlinear equations in (μ, x, z), are fundamentally different from solving the mNEPv (5.2). Efficient solutions to the nonlinear equations for the C-eigenvalue are still largely open.

5.3. Distance problem in dHDAE systems.

Consider the following dHDAE:

$$J \frac{d^j u}{dt^j} = B_0 + B_1 \frac{du}{dt} + \cdots + B_\ell \frac{d^\ell u}{dt^\ell}, \tag{5.10}$$

where $u\colon \mathbb{R} \to \mathbb{R}^n$ is a state function, j is an integer between 0 and ℓ, $J = -J^T$ is skew symmetric, and $B_i \succeq 0$ are symmetric positive semidefinite for $i = 0, \ldots, \ell$.

By convention, $\frac{d^0 u}{dt^0} = u$. The dHDAE (5.10) arises in energy-based modeling of dynamical systems [43, 60]. An important special case is with $j = 0$ and $\ell = 1$, known as the linear time-invariant dHDAE system [11, 60]. Another one is the second-order dHDAE (5.10) with $j = 1$ and $\ell = 2$ [11, 43].

To analyze the dynamical properties of a dHDAE system, one needs to know whether the system is close to a singular one. A dHDAE system (5.10) is called *singular* if $\det(P(\lambda)) \equiv 0$ for all $\lambda \in \mathbb{C}$, where

$$P(\lambda) = -\lambda^j J + B_0 + \lambda B_1 + \cdots + \lambda^\ell B_\ell \tag{5.11}$$

is the characteristic matrix polynomial. The distance of a dHDAE system to the closest singular dHDAE system is measured by the quantity $d_{\text{sing}}(P(\lambda))$:

$$d_{\text{sing}}(P(\lambda)) = \min_{\substack{x\in\mathbb{R}^n\\ \|x\|=1}} \left\{ 2\|Jx\|^2 + \sum_{i=0}^{\ell} \Big(2\|(I - xx^T)B_i x\|^2 + (x^T B_i x)^2 \Big) \right\}^{1/2}; \tag{5.12}$$

see [43, Thm. 16]. We can reformulate the optimization (5.12) to an aMax (1.3). First, by the skew-symmetry of J and the symmetry of B_i, we can write (5.12) as

$$\begin{aligned} \Big(d_{\text{sing}}(P(\lambda))\Big)^2 &= \min_{\substack{x\in\mathbb{R}^n\\ \|x\|=1}} \left\{ 2\cdot x^T(J^T J)x + \sum_{i=0}^{\ell} \left[2x^T(B_i^T B_i)x - (x^T B_i x)^2 \right] \right\} \\ &= -2\cdot \max_{\substack{x\in\mathbb{R}^n\\ \|x\|=1}} \left\{ x^T A_1 x + \frac{1}{2}\sum_{i=2}^{\ell+2} (x^T A_i x)^2 \right\}, \end{aligned} \tag{5.13}$$

where $A_1 \equiv J^2 - \sum_{i=0}^{\ell} B_i^2$ and $A_i \equiv B_{i-2}$ for $i = 2, \dots, \ell+2$. Consequently, (5.13) is of the form of the aMax (1.3),

$$\max_{x\in\mathbb{R}^n,\, \|x\|=1} \left\{ F(x) := x^T A_1 x + \frac{1}{2}\sum_{i=2}^{\ell+2} \left(x^T A_i x\right)^2 \right\}, \tag{5.14}$$

with $\phi_1(t) = t$ and $\phi_i(t) = t^2/2$ for $i = 2, \dots, \ell+2$. By Theorem 2.3, a local maximizer of (5.14) can be found by solving the following mNEPv of the form (1.1):

$$H(x)x = \lambda x \quad \text{with} \quad H(x) \equiv A_1 + \sum_{i=2}^{\ell+2} (x^T A_i x)\cdot A_i, \tag{5.15}$$

where $h_1(t) = 1$ and $h_i(t) = t$ for $i = 2, \dots, \ell+2$ are nondecreasing functions.

Computable upper and lower bounds of the quantity $d_{\text{sing}}(P(\lambda))$ have been studied in [43, 50], and a recent method using two-level minimization and gradient flow has been proposed for estimating $d_{\text{sing}}(P(\lambda))$ [24]. In comparison, the mNEPv approach provides a computationally efficient alternative for estimating $d_{\text{sing}}(P(\lambda))$ for dHDAE systems of any order; see Examples 6.2 and 6.3 in section 6.

6. Numerical examples.

In this section, we present numerical examples of Algorithm 4.1 for solving the mNEPv (1.1) arising from the applications described in section 5. The main purpose of the experiments is to illustrate the convergence behavior of the SCF (Algorithm 4.1 with $\text{tol}_{\text{acc}} = 0$) and the efficiency of the accelerated SCF (Algorithm 4.1 with $\text{tol}_{\text{acc}} = 0.1$). The error tolerance for both algorithms is set to tol $=10^{-13}$. All experiments are carried out in MATLAB and run on a Dell desktop with an Intel i9-9900K CPU at 3.6 GHZ and 16 GB core memory. In the spirit of reproducible research, we have made available the MATLAB scripts implementing the algorithms and the data used to generate the numerical results presented in this paper. They can be accessed at https://github.com/ddinglu/mnepv.

Example 6.1. In subsection 5.1, we discussed that the computation of the numerical radius of a matrix $B\in\mathbb{C}^{n\times n}$ is related to mNEPv (3.15) and the variational characterization (3.16) with Hermitian $A_1=(B^H+B)/2$ and $A_2=(B^H-B)\cdot\imath/2$. For the numerical experiment, let us consider the following matrix:

$$B=\begin{bmatrix}0.6 & -0.2 & -1.9 & -0.3\\ -0.1 & -0.3 & -1.3 & -1.2\\ -2.0 & -1.6 & -2.1 & 1.3\\ -0.1 & -1.6 & 1.5 & -0.1\end{bmatrix}+\imath\begin{bmatrix}0.6 & 2.5 & -0.2 & 2.5\\ 2.3 & -2.6 & 0.4 & 1.3\\ 0.0 & 0.6 & -0.4 & 1.2\\ 2.0 & 1.4 & 1.0 & -2.3\end{bmatrix}. \tag{6.1}$$

The corresponding numerical range $W(A_1,A_2)$ is depicted in Figure 2 as the shaded region. We sampled 100 different starting vectors x_0 to run the SCF, where each $y_0=g(x_0)$ is a supporting point of $W(A_1,A_2)$, depicted in Figure 2 as dots on the boundary of $W(A_1,A_2)$. By the discussion on the implementation of Algorithm 4.1, such initial x_0 are obtained from the eigenvectors x_v of the matrix H_v for sampled directions $v\in\mathbb{R}^2$ (see Lemma 3.1 and subsection 4.2). Since a unit direction $v\in\mathbb{R}^2$ can be represented by polar coordinates as $v=[\cos\theta,\sin\theta]^T$ with $\theta\in[0,2\pi)$, the initials x_0 are set as

$$x_0:=x_v \quad \text{with } v=[\cos\theta,\sin\theta]^T \tag{6.2}$$

using 100 equally distant θ between 0 and 2π. The sampled $g(x_0)$ are well distributed on the boundary of $W(A_1,A_2)$, as shown in Figure 2.

For 100 runs of the SCF, three different solutions are found. In Figure 2, they are labeled, respectively, with I, II, III, in descending order of their objective values of (3.16). The initial $g(x_0)$ on the boundary of $W(A_1,A_2)$ are colored the same

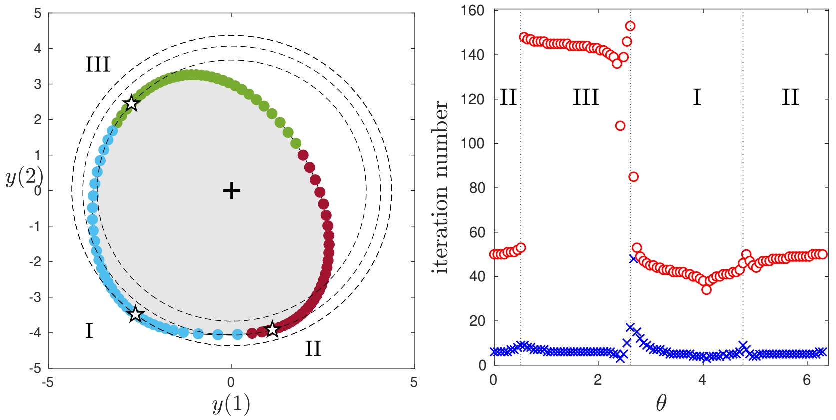

FIG. 2. *Left: Numerical range $W(A_1,A_2)$ of the matrix in* (6.1). *☆ represents the solution for the mNEPv and ● the starting $g(x_0)$ of the SCF. The ● are colored according to the solution they have computed (blue is for solution* I, *red for* II, *and green for* III*). The dashed lines are contours of $\phi(y)=\|y\|^2/2$; see* (3.16). *Right: Number of SCF iterations ("o") and the accelerated SCF ("×") for different x_0 parameterized by $\theta\in[0,2\pi)$, as in* (6.2).

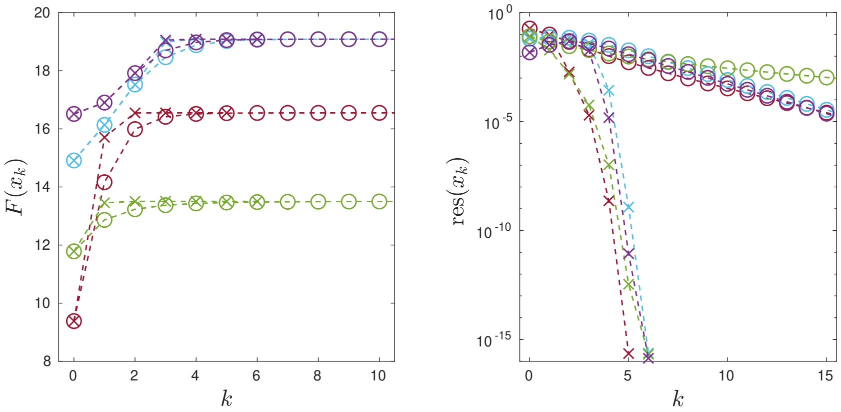

FIG. 3. *Left: Convergence history of $F(x_k)$ by the SCF ("o") and the accelerated SCF ("×"), where each colored curve is a run with a particular x_0 from four different starting vectors. Right: Relative residual norms* (4.6) *of the mNEPv.*

if the SCF will converge to the same solution, which, hence, reveals the region of convergence for the SCF. The numbers of SCF iterations with each x_0 are reported in Figure 2. For the SCF, the iteration numbers vary for different solutions, whereas the accelerated SCF are almost independent of the choice of the initial x_0 with only a moderate increase on the boundary of two convergence regions.

The left plot of Figure 3 depicts the convergence history of the objective function $F(x_k)$ for four different starting vectors x_0, corresponding to the equally distant $\theta \in \{0, \pi/2, \pi, 3\pi/2\}$ from Figure 2. As expected, the SCF demonstrates monotonic convergence. The right plot in Figure 3 shows the relative residual norms of x_k, as defined in (4.6). We can see that the SCF quickly enters the region of linear convergence in all cases (in about three iterations). The acceleration takes full advantage of the rapid initial convergence and speeds up the SCF significantly. We note that in this example, the matrices A_1 and A_2 are complex Hermitian, for which the inverse iteration (4.5) with Rayleigh shift σ_k is not guaranteed quadratically convergent.

Example 6.2. In this example, we consider the mNEPv (5.15) arising from the distance problem of dHDAE systems described in subsection 5.3. The characteristic polynomial of a linear dHDAE system is given by

$$P(\lambda) := -J + R + \lambda E, \tag{6.3}$$

where $J = -J^T$ is skew symmetric and E and R are symmetric positive definite matrices. As discussed in subsection 5.3, the computation of distance to singularity $d_{\mathrm{sing}}(P(\lambda))$ leads to the optimization (5.13) and the associated mNEPv (5.15), where

$$F(x) = x^T A_1 x + \frac{1}{2}\sum_{i=2}^{3}(x^T A_i x)^2 \quad \text{and} \quad H(x) = A_1 + \sum_{i=2}^{3}(x^T A_i x)\cdot A_i \tag{6.4}$$

and $A_1 = J^2 - E^2 - R^2$, $A_2 = E$, and $A_3 = R$.

For experiments, the matrices $\{J, R, E\}$ of order 30 are generated randomly.[4] Similar to Example 6.1, the initial x_0 of the SCF are computed from supporting points

[4] For J: `X=randn(n); X = X-X'; X = X/norm(X)`. For E and R: `X=randn(n); X = orth(X); X = X * diag(rand(n,1) + 1.6E-6)*X'`.

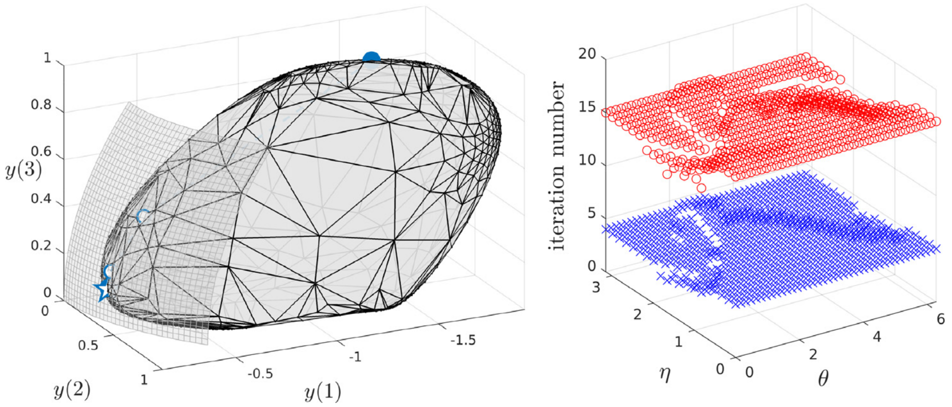

Fig. 4. *Left: Computed numerical range $W(A_1, A_2, A_3)$ based on* 800 *sample supporting points on the boundary (nodes of the mesh); ☆ represents the solution for the mNEPv, ● the starting $g(x_0)$, and "○" the first few supporting points $g(x_k)$ by the SCF. The smaller mesh that crosses ☆ is part of the level surface $\phi(y) = \phi(y_*)$ for $\phi(y) = y(1) + (y(2)^2 + y(3)^2)/2$ at the solution $y_* = g(\widehat{x}_*)$. Right: Number of SCF iterations ("o") and the accelerated SCF ("×") for different starting x_0 parameterized by $\theta \in [0, 2\pi)$ and $\eta \in [0, \pi)$, as in* (6.5).

of the joint numerical range $W(A_1, A_2, A_3) \subset \mathbb{R}^3$ along several sampled directions $v \in \mathbb{R}^3$. Recall that a unit $v \in \mathbb{R}^3$ can be represented by spherical coordinates as

$$v = [\sin\eta\cos\theta,\, \sin\eta\sin\theta,\, \cos\eta]^T \quad \text{with } \eta \in [0, 2) \text{ and } \theta \in [0, 2\pi). \tag{6.5}$$

We hence construct an equispaced grid of 20 by 40 points of $(\eta, \theta) \in [0, \pi] \times [0, 2\pi]$, yielding 800 supporting points of $W(A_1, A_2, A_3)$. They are depicted in Figure 4, together with the approximate joint numerical range they generate.[5]

From all 800 initial x_0, the SCF converge to the same solution, as marked in Figure 4. This solution appears to be the global optimizer of (5.13), as visually verified by the level surface of the objective function $\phi(y)$ for the corresponding optimization over the joint numerical range (3.3). From the numbers of iterations reported in Figure 4, we can see that both the SCF and the accelerated SCF converge rapidly to the solution. The numbers of SCF iterations are not sensitive to the choice of x_0. Figure 5 depicts the convergence history of $F(x_k)$ and the relative residual norms by the SCF from six different starting vectors x_0 (sampled supporting points of $W(A_1, A_2, A_2)$ along the three coordinate axes). We observe that the SCF converges monotonically to the same solution regardless of the starting vector used. The accelerated SCF greatly reduces the number of iterations and shows a quadratic convergence rate.

In general, a computed $\widehat{x}_*$ may not be a global maximizer of the aMax (5.13). But we have at least an upper bound of the distance:

$$d_{\text{sing}}(P(\lambda)) \equiv \big(-2 \cdot \max_{\|x\|=1} F(x)\big)^{1/2} \leq \big(-2 \cdot F(\widehat{x}_*)\big)^{1/2}. \tag{6.6}$$

If the initial vector x_0 of the SCF is especially set to be the eigenvector corresponding to the largest eigenvalue of A_1, then we have

$$\big(-2 \cdot F(\widehat{x}_*)\big)^{1/2} \leq \big(-2 \cdot F(x_0)\big)^{1/2} \leq \delta_M := \big(-2 \cdot \lambda_{\max}(A_1)\big)^{1/2}, \tag{6.7}$$

[5] Plot generated by MATLAB functions `trisurf` and `boundary` using the 800 supporting points.

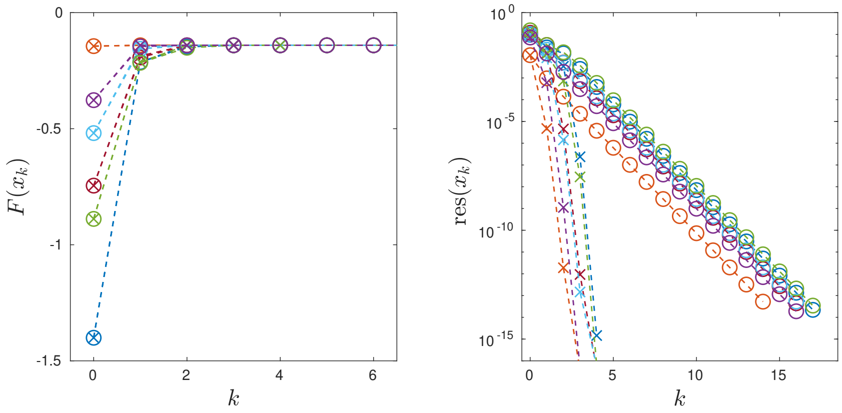

FIG. 5. *Left: Convergence history of $F(x_k)$ by the SCF ("o") and the accelerated SCF ("×"), where each colored curve is a run with a particular x_0 from six different starting vectors. Right: Relative residual norms (4.6) of the mNEPv.*

where the first inequality is by the monotonicity of the SCF (see Theorem 3.4) and the second inequality is by the definition of $F(x)$ (6.4). The quantity δ_M was introduced in [43] and used as an estimation of $d_{\mathrm{sing}}(P(\lambda))$. By the inequalities (6.6) and (6.7), the SCF always produces a sharper upper bound of $d_{\mathrm{sing}}(P(\lambda))$. In this example, the SCF provides a sharper estimation $\sqrt{-2\cdot F(\widehat{x}_*)} \approx 0.5989$, as opposed to $\delta_M \approx 0.6923$.

An alternative computable upper bound to the quantity δ_M has been recently proposed in [50], which involves an optimization of sum of Rayleigh quotients, but it does not ensure a better estimation than δ_M [50, Thm. 3.7 and Example (3)]. In another related work [24], the authors considered an approach to estimate the distance $d_{\mathrm{sing}}(P(\lambda))$, based on the observation that the distance is the smallest root of a monotonically decreasing function w. A root-finding method such as bisection can be applied. The difficulty there lies in the evaluation of the function w. For a given ϵ, evaluating $w(\epsilon)$ can be very expensive, as it requires an optimization by a gradient flow method, which involves repeated solution of Hermitian eigenvalue problems of size n.

Example 6.3. In this example, we consider a quadratic dHDAE system with the characteristic polynomial

$$P(\lambda) := -\lambda G + K + \lambda D + \lambda^2 M,$$

where $G = -G^T$ is skew symmetric and M, D, and K are symmetric positive definite. By subsection 5.3, the computation of distance to singularity $d_{\mathrm{sing}}(P(\lambda))$ leads to the optimization (5.13) and the mNEPv (5.15) with

$$F(x) = x^T A_1 x + \frac{1}{2}\sum_{i=2}^{4}(x^T A_i x)^2 \quad \text{and} \quad H(x) = A_1 + \sum_{i=2}^{4}(x^T A_i x)\cdot A_i,$$

where $A_1 = G^2 - M^2 - D^2 - K^2$, $A_2 = M$, $A_3 = D$, and $A_4 = K$.

For numerical experiments, we consider a lumped-parameter mass-spring-damper system $M\ddot{u} + D\dot{u} + Ku = f$ with n point masses and n spring-damper pairs. The matrices D and K are interchangeable with $DK = KD$ and are simultaneously diagonalizable [61]. We pick a random skew symmetric G to simulate the gyroscopic effect.

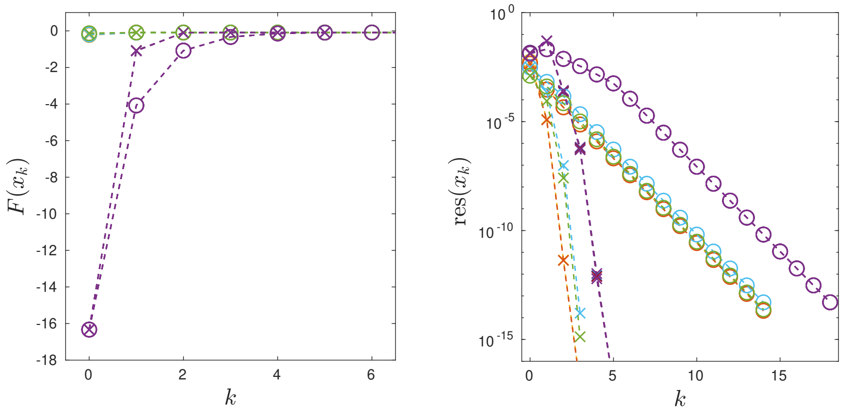

FIG. 6. *Left: Convergence history of $F(x_k)$ by the SCF ("o") and the accelerated SCF ("×"), where each colored curve is a run with a particular x_0 from eight different starting vectors (lines overlapped). Right: Relative residual norms* (4.6) *of the mNEPv.*

TABLE 1

Number of iterations and computation time (in seconds) for various problem sizes n. Reported are average results from 100 *runs with different starting vectors, with the largest deviations marked.*

n	algorithms	$F(x_*)$	iterations	timing
500	RTR	-0.094157045470939 $(\pm 7\cdot 10^{-17})$	24.2 (±10)	1.63 (±0.47)
	SCF	-0.094157045470939 $(\pm 7\cdot 10^{-17})$	17.0 (±4.0)	1.18 (±0.34)
	accel. SCF	-0.094157045470939 $(\pm 7\cdot 10^{-17})$	5.3 (±1.3)	0.34 (±0.14)
1000	RTR	-0.095120974693461 $(\pm 7\cdot 10^{-17})$	27.8 (±8.8)	8.98 (±1.33)
	SCF	-0.095120974693461 $(\pm 4\cdot 10^{-17})$	21.3 (±3.3)	6.54 (±1.15)
	accel. SCF	-0.095120974693461 $(\pm 6\cdot 10^{-17})$	4.7 (±1.3)	1.32 (±0.48)
2000	RTR	-0.090910959613593 $(\pm 6\cdot 10^{-17})$	27.6 (±12)	58.35 (±9.37)
	SCF	-0.090910959613593 $(\pm 4\cdot 10^{-17})$	17.0 (±4.0)	4.91 (±1.58)
	accel. SCF	-0.090910959613593 $(\pm 6\cdot 10^{-17})$	4.8 (±1.8)	1.52 (±0.73)
3000	RTR	-0.089186202007536 $(\pm 7\cdot 10^{-17})$	28.4 (±11)	181.65 (±20.6)
	SCF	-0.089186202007536 $(\pm 8\cdot 10^{-17})$	16.9 (±3.9)	20.51 (±5.33)
	accel. SCF	-0.089186202007536 $(\pm 7\cdot 10^{-17})$	5.2 (±1.2)	6.39 (±1.93)

The sizes n of the matrices are set ranging from 500 to 3000. For each set of testing matrices, we run the SCF with 100 different starting vectors x_0. Again, those x_0 are computed from supporting points of the joint numerical range $W(\mathcal{A})\subset\mathbb{R}^4$ along 100 randomly sampled directions $v\in\mathbb{R}^4$.

Similar to the linear system in Example 6.2, the SCF converge to the same solution from all 100 different starting vectors. Figure 6 depicts the convergence history of the SCF and the accelerated SCF for a case of $n = 1000$, with eight randomly selected starting vectors. It shows the same convergence behavior of the SCF and the accelerated SCF, as in the previous example. Table 1 summarizes the iteration number and computation time for the algorithms from all testing cases. We can see that the performance of both the SCF and the accelerated SCF are not much affected by the choice of initial vectors. Both algorithms converge rapidly, and the accelerated SCF speed up to a factor between 2.5 and 6.2. For comparison, we have included the results by the Riemannian trust region (RTR) method for solving the optimization problem (5.14). We used the `trustregions` function provided by `Manopt`, a MATLAB toolbox available at https://www.manopt.org. RTR is considered as a state-of-the-art approach for the optimization problems with spherical constraints of the form$\|x\|=1$. We observe that RTR finds the same solution as the proposed NEPv approach, but it takes significantly more running time.

Example 6.4. As discussed in subsection 5.2, the problem of best rank-one approximation for a partial-symmetric tensor $T \in \mathbb{R}^{n\times n\times m}$ leads to a quartic optimization (5.7) and the corresponding mNEPv (5.2), where the coefficient matrices are $A_i := T(:,:,i) \in \mathbb{R}^{n\times n}$ for $i = 1,\dots,m$. For nonnegative tensors, the objective function $F(x) = \frac{1}{2}\sum_i \left(x^T A_i x\right)^2$ of (5.7) satisfies $F(|x|) \geq F(x)$, where $|\cdot|$ denotes componentwise absolute value. Therefore, it is advisable to start the SCF (3.1) with a nonnegative initial x_0. Note that if $x_k \geq 0$, then $H(x_k) \geq 0$, so by the Perron–Frobenius theorem [28], the eigenvector x_{k+1} for the largest eigenvalue of $H(x_k)$ is also nonnegative. Consequently, the iterates x_k by the SCF will remain nonnegative.

We note that for a nonnegative tensor T and a nonnegative initial x_0, the SCF (3.1) is indeed equivalent to the alternating least squares (ALS) algorithm for finding the best rank-one approximation (5.5). Recall that in subsection 5.2, the best rank-one approximation (5.5) is turned into the maximization problem

$$\max_{\|x\|=1,\,\|z\|=1} \left(z^T\cdot g(x)\right)^2, \tag{6.8}$$

where $g(x) = [x^T A_1 x,\dots,x^T A_m x]^T$. Maximizing alternatively with respect to z and x leads to the alternating iteration

$$\begin{cases} z_{k+1} = \underset{\|z\|=1}{\operatorname{argmax}} \left(z^T\cdot g(x_k)\right)^2 = \alpha_k\cdot g(x_k), \\ x_{k+1} = \underset{\|x\|=1}{\operatorname{argmax}} \left(z_{k+1}^T\cdot g(x)\right)^2 = \underset{\|x\|=1}{\operatorname{argmax}} \left(x^T\cdot H(x_k)\cdot x\right)^2 \end{cases} \tag{6.9}$$

for $k = 1,2,\dots$, where $\alpha_k > 0$ is a normalization factor for z_{k+1}. Note that $H(x_k) \geq 0$ if $x_k \geq 0$. The maximizer x_{k+1} of (6.9) is the eigenvector corresponding to the largest eigenvalue of $H(x_k)$ by the Perron–Frobenius theorem. Therefore, the iteration (6.9) coincides with the SCF. The ALS algorithms are commonly used for low-rank approximations in tensor computations [33].

For numerical experiments, we use the following third-order partial-symmetric tensors: the *New Orleans tensor*[6] is created from a Facebook network and has size $63891 \times 63891 \times 20$ with 477778 nonzeros, the *Princeton tensor*[7] is from a Facebook "friendship" network and has size $6593 \times 6593 \times 6$ with 70248 nonzeros, and the *Reuters tensor*[8] is from a news network based on all stories released by the news agency Reuters concerning the September 11, 2011, attack during the 66 consecutive days beginning on September 11, and the size of the tensor T is $13332 \times 13332 \times 66$ with 486894 nonzeros. All three tensors are nonnegative and sparse (density $\approx 10^{-5}$), and so are the corresponding coefficient matrices $A_i = T(:,:,i)$ for $i = 1,\dots,m$.

In Algorithm 4.1, we use MATLAB `eigs` for the eigenvalue computation and `minres` for solving the linear system in the acceleration (4.5). We use an adaptive error tolerance $\mathrm{Tol} = \min\{10^{-3}, \mathrm{res}(x_k)^2\}$ for each call of `eigs` and `minres`. We use 100 randomly generated and nonnegative starting vectors x_0 to run the SCF (using `x0=abs(randn(n,1))`). The convergence history is reported in Figure 7. We observe

[6] Data available at http://socialnetworks.mpi-sws.org/data-wosn2009.html.

[7] Data available at https://archive.org/details/oxford-2005-facebook-matrix.

[8] Data available at http://vlado.fmf.uni-lj.si/pub/networks/data/CRA/terror.htm.

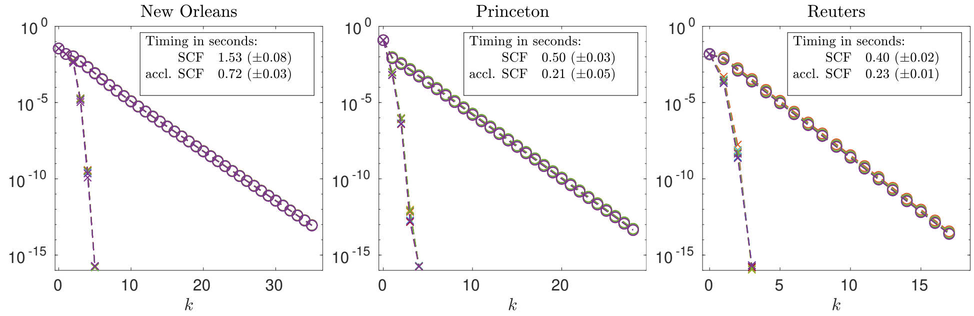

FIG. 7. *Convergence history of relative residual norms* $res(x_k)$ (4.6) *by the SCF (“*$\circ$*”) and the accelerated SCF (“*$\times$*”). Each colored curve represents a run with a different starting vector from* 100 *randomly generated* $x_0 \geq 0$ *(due to curve overlapping, eight selected curves are reported). The reported computational times are the average results from the* 100 *runs, with the largest deviations marked.*

that from different starting x_0, Algorithm 4.1 always converges to the same solution, and the convergence rate appears not to be affected by the choice of x_0. Also, the accelerated SCF significantly reduces the number of the SCF iterations and has a quadratic convergence rate. It is noteworthy that the SCF can find the solution in just about a fraction of a second. This is a surprising result given the large size of the Hermitian eigenvalue problem that is solved in each iteration.

7. Concluding remarks.

A variational characterization for the mNEPv (1.1) is revealed. Based on that, we provided a geometric interpretation of the SCF iterations for solving the mNEPv. The geometry of the SCF illustrates the global monotonic convergence of the algorithm and leads to a rigorous proof of its global convergence. In addition, we presented an inverse iteration–based scheme to accelerate the convergence of the SCF. Numerical examples demonstrated the effectiveness of the accelerated SCF for solving the mNEPv arising from different applications. By the intrinsic connection between the mNEPv (1.1) and the aMax (1.3), we developed an NEPv approach for solving the aMax. Algorithmically, it allows the use of state-of-the-art eigensolvers for fast solution.

Most results presented in this work can be extended to the case of NEPv (1.1) with h_i being nondecreasing and locally Lipschitz continuous functions. A variational characterization of such NEPv similar to Theorem 2.3 can be established. The present work also lays the groundwork for studying a more general class of NEPv in the form (1.1), where the coefficient of A_i is a composite function $h_i(g(x))$ with a given $h_i : \mathbb{R}^m \to \mathbb{R}$ and $g(x)$, as defined in (3.2). Expanding theoretical analysis and geometric interpretation of the SCF discussed in the present work to such NEPv is a topic for future study.

Acknowledgment. We thank the anonymous referees for their many helpful comments that substantially improved this manuscript.

REFERENCES

[1] P. A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, 2009.

[2] H. Amann, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Rev., 18 (1976), pp. 620–709.

[3] Y.-H. Au-Yeung and N.-K. Tsing, *An extension of the Hausdorff-Toeplitz theorem on the numerical range*, Proc. Amer. Math. Soc., 89 (1983), pp. 215–218.

[4] Y.-H. Au-yeung and N.-K. Tsing, *Some theorems on the generalized numerical ranges*, Linear Multilinear Algebra, 15 (1984), pp. 3–11.

[5] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, Cambridge, 1996.

[6] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, SIAM, Philadelphia, 2000.

[7] Z. Bai, R.-C. Li, and D. Lu, *Sharp estimation of convergence rate for self-consistent field iteration to solve eigenvector-dependent nonlinear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 43 (2022), pp. 301–327.

[8] Z. Bai, D. Lu, and B. Vandereycken, *Robust Rayleigh quotient minimization and nonlinear eigenvalue problems*, SIAM J. Sci. Comput., 40 (2018), pp. A3495–A3522.

[9] 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.

[10] R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. Van der Vorst, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, SIAM, Philadelphia, 1994.

[11] C. Beattie, V. Mehrmann, H. Xu, and H. Zwart, *Linear port-Hamiltonian descriptor systems*, Math. Control Signals Syst., 30 (2018), pp. 1–27.

[12] A. Ben-Tal and A. Nemirovski, *Robust convex optimization*, Math. Oper. Res., 23 (1998), pp. 769–805.

[13] V. Berinde, *Iterative Approximation of Fixed Points*, Springer, New York, 2007.

[14] R. Bhatia, *Matrix Analysis*, Springer, New York, 2013.

[15] S. P. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

[16] T. Bühler and M. Hein, *Spectral clustering based on the graph p-Laplacian*, in Proceedings of the 26th International Conference on Machine Learning, 2009, pp. 81–88.

[17] 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.

[18] E. Cancès, G. Kemlin, and A. Levitt, *Convergence analysis of direct minimization and self-consistent iterations*, SIAM J. Matrix Anal. Appl., 42 (2021), pp. 243–274.

[19] 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.

[20] J. D. Carroll and J.-J. Chang, *Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition*, Psychometrika, 35 (1970), pp. 283–319.

[21] Y. Chen, A. Jákli, and L. Qi, *The C-eigenvalue of third order tensors and its application in crystals*, J. Ind. Manag. Optim., 19 (2023), pp. 265–281.

[22] M. Chō and M. Takaguchi, *Boundary points of joint numerical ranges*, Pacific J. Math., 95 (1981), pp. 27–35.

[23] L. De Lathauwer, B. De Moor, and J. Vandewalle, *A multilinear singular value decomposition*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1253–1278.

[24] N. Guglielmi and V. Mehrmann, *Computation of the nearest structured matrix triplet with common null space*, Electron. Trans. Numer. Anal., 55 (2022), pp. 508–531.

[25] S. Güttel and F. Tisseur, *The nonlinear eigenvalue problem*, Acta Numer., 26 (2017), pp. 1–94.

[26] C. He and G. Watson, *An algorithm for computing the numerical radius*, IMA J. Numer. Anal., 17 (1997), pp. 329–342.

[27] S. He, Z. Li, and S. Zhang, *Approximation algorithms for homogeneous polynomial optimization with quadratic constraints*, Math. Program., 125 (2010), pp. 353–383.

[28] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 2012.

[29] P. Huang, Q. Yang, and Y. Yang, *Finding the global optimum of a class of quartic minimization problem*, Comput. Optim. Appl., 81 (2022), pp. 923–954.

[30] I. C. F. Ipsen, *Computing an eigenvector with inverse iteration*, SIAM Rev., 39 (1997), pp. 254–291.

[31] 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.

[32] E. Kofidis and P. A. Regalia, *On the best rank-*1 *approximation of higher-order supersymmetric tensors*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 863–884.

[33] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500.

[34] J. Lampe and H. Voss, *A survey on variational characterization for nonlinear eigenvalue problems*, Electron. Trans. Numer. Anal., 55 (2022), pp. 1–75.

[35] C.-K. Li and E. Poon, *Maps preserving the joint numerical radius distance of operators*, Linear Algebra Appl., 437 (2012), pp. 1194–1204.

[36] C.-K. Li and Y.-T. Poon, *Convexity of the joint numerical range*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 668–678.

[37] N. Li, C. Navasca, and C. Glenn, *Iterative methods for symmetric outer product tensor decomposition*, Electron. Trans. Numer. Anal., 44 (2015), pp. 124–139.

[38] L.-H. Lim, *Singular values and eigenvalues of tensors: A variational approach*, in 1st IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, IEEE, New York, 2005, pp. 129–132.

[39] L. Lin and C. Yang, *Elliptic preconditioner for accelerating the self-consistent field iteration in Kohn–Sham density functional theory*, SIAM J. Sci. Comput., 35 (2013), pp. S277–S298.

[40] X. Liu, Z. Wen, X. Wang, 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.

[41] D. Lu, *Nonlinear eigenvector methods for convex minimization over the numerical range*, SIAM J. Matrix Anal. Appl., 41 (2020), pp. 1771–1796.

[42] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods*, Cambridge University Press, Cambridge, 2004.

[43] C. Mehl, V. Mehrmann, and M. Wojtylak, *Distance problems for dissipative Hamiltonian systems and related matrix polynomials*, Linear Algebra Appl., 623 (2021), pp. 335–366.

[44] E. Mengi and M. L. Overton, *Algorithms for the computation of the pseudospectral radius and the numerical radius of a matrix*, IMA J. Numer. Anal., 25 (2005), pp. 648–669.

[45] T. Mitchell, *Convergence Rate Analysis and Improved Iterations for Numerical Radius Computation*, preprint, arXiv:2002.00080, 2020.

[46] Y. Nesterov, *Random Walk in a Simplex and Quadratic Optimization over Convex Polytopes*, Technical report, CORE Discussion Paper, UCL, Louvain-la-Neuve, Belgium, 2003.

[47] T. T. Ngo, M. Bellalij, and Y. Saad, *The trace ratio optimization problem*, SIAM Rev., 54 (2012), pp. 545–569.

[48] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, New York, 2006.

[49] J. F. Nye, *Physical Properties of Crystals: Their Representation by Tensors and Matrices*, Oxford University Press, Oxford, 1985.

[50] A. Prajapati and P. Sharma, *Estimation of structured distances to singularity for matrix pencils with symmetry structures: A linear algebra–based approach*, SIAM J. Matrix Anal. Appl., 43 (2022), pp. 740–763.

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

[52] L. Qi, *Eigenvalues of a real supersymmetric tensor*, J. Symbolic Comput., 40 (2005), pp. 1302–1324.

[53] L. Qi, H. Chen, and Y. Chen, *Tensor Eigenvalues and Their Applications*, Vol. 39, Springer, New York, 2018.

[54] C. C. J. Roothaan, *New developments in molecular orbital theory*, Rev. Mod. Phys., 23 (1951), p. 69.

[55] R. E. Stanton, *Intrinsic convergence in closed-shell SCF calculations. A general criterion*, J. Chem. Phys., 75 (1981), pp. 5416–5422.

[56] L. N. Trefethen and M. Embree, *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*, Princeton University Press, 2005.

[57] F. Tudisco and D. J. Higham, *A nonlinear spectreal method for core-periphery detection in networks*, SIAM J. Math. Data Sci., 1 (2019), pp. 269–292.

[58] F. Uhlig, *Geometric computation of the numerical radius of a matrix*, Numer. Algorithms, 52 (2009), p. 335.

[59] 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.

[60] A. Van Der Schaft and D. Jeltsema, *Port-Hamiltonian systems theory: An introductory overview*, Found. Trends Syst. Control, 1 (2014), pp. 173–378.

[61] K. Veselić, *Damped Oscillations of Linear Systems: A Mathematical Introduction*, Vol. 2023, Springer, Berlin, Heidelberg, 2011.

[62] G. A. Watson, *Computing the numerical radius*, Linear Algebra Appl., 234 (1996), pp. 163–172.

[63] A. Weinstein and W. Stenger, *Methods of Intermediate Problems for Eigenvalues: Theory and Ramifications*, Academic Press, Elsevier, 1972.

[64] 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.

[65] H. Zhang, A. Milzarek, Z. Wen, and W. Yin, *On the geometric analysis of a quartic-quadratic optimization problem under a spherical constriant*, Math. Program., (2021).

[66] L.-H. Zhang, *On optimizing the sum of the Rayleigh quotient and the generalized Rayleigh quotient on the unit sphere*, Comput. Optim. Appl., 54 (2013), pp. 111–139.

[67] L.-H. 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.

[68] L.-H. 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.

[69] T. Zhang and G. H. Golub, *Rank-one approximation to high order tensors*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 534–550.

[70] X. Zhang, L. Qi, and Y. Ye, *The cubic spherical optimization problems*, Math. Comput., 81 (2012), pp. 1513–1525.