Locally Unitarily Invariantizable NEPv and Convergence Analysis of SCF

Ding Lu* and Ren-Cang Li †

Abstract

We consider a class of eigenvector-dependent nonlinear eigenvalue problems (NEPv) without the unitary invariance property. Those NEPv commonly arise as the first-order optimality conditions of a particular type of optimization problems over the Stiefel manifold, and previously, special cases have been studied in the literature. Two necessary conditions, a definiteness condition and a rank-preserving condition, on an eigenbasis matrix of the NEPv that is a global optimizer of the associated optimization problem are revealed, where the definiteness condition has been known for the special cases previously investigated. We show that, locally close to the eigenbasis matrix satisfying both necessary conditions, the NEPv can be reformulated as a unitarily invariant NEPv, the so-called *aligned NEPv*, through a basis alignment operation — in other words, the NEPv is locally unitarily invariantizable. Numerically, the NEPv is naturally solved by an SCF-type iteration. By exploiting the differentiability of the coefficient matrix of the aligned NEPv, we establish a closed-form local convergence rate for the SCF-type iteration and analyze its level-shifted variant. Numerical experiments confirm our theoretical results.

Keywords. Optimization on Stiefel manifold, eigenvector-dependent nonlinear eigenvalue problems, self-consistent-field iteration, rate of convergence, unitary invariance, polar decomposition.

Mathematics Subject Classification (2010). 65F15, 65H17

1 Introduction

Consider the *eigenvector-dependent nonlinear eigenvalue problem* (NEPv): Find $X \in \mathbb{R}^{n\times k}$ that has orthonormal columns (i.e., $X^{\mathrm{T}}X = I_k$) and a square $\Lambda \in \mathbb{R}^{k\times k}$ satisfying

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

where $H(X) \in \mathbb{R}^{n\times n}$ is a symmetric matrix continuously dependent of X. When (1.1) holds, we call X an *eigenbasis matrix* and (X, Λ) an eigen-matrix pair of the NEPv. Necessarily, $k \le n$ (usually $k \ll n$), and the columns of X form an orthonormal basis of the eigenspace of $H(X)$ evaluated at X, and the corresponding eigenvalues are the k eigenvalues of $\Lambda = X^{\mathrm{T}}H(X)X \in \mathbb{R}^{k\times k}$, which is necessarily symmetric. Note that Λ may not be diagonal and individual columns of X may not be some eigenvectors of $H(X)$.

*Department of Mathematics, University of Kentucky, Lexington, KY 40506, USA, Ding.Lu@uky.edu. Supported in part by NSF DMS-2110731.

†Department of Mathematics, University of Texas at Arlington, Arlington, TX 76019-0408, USA, rcli@uta.edu. Supported in part by NSF DMS-2009689.

NEPv (1.1) commonly arises in important real-life applications. The two most prominent examples are the Kohn-Sham equation in the density functional theory [33, 42] and the Gross-Pitaevskii equation for the Bose-Einstein condensation [6, 24], both from computational physics and chemistry. In recent years, NEPv increasingly show up in the fields of data science and machine learning where various optimization problems with orthogonality constraints need to be solved; see, e.g., [5, 25, 34, 36, 45, 51, 52, 53, 54, 55].

NEPv is a term coined by the authors of [10], where it refers to a problem (1.1) with $H(\cdot)$ satisfying the so-called *unitary invariance property*, by which we mean[1]

$$H(XQ) = H(X) \quad \text{for all orthogonal } Q \in \mathbb{R}^{k\times k}. \tag{1.2}$$

Under condition (1.2), if X solves NEPv (1.1) then so does XQ, and, hence, a solution X is really a representative of a set of solutions that share the same column space, i.e., a point in the Grassmannian $\mathcal{G}_k^n$ (the collection of all k dimensional subspaces in $\mathbb{R}^n$). This unitary invariance property is possessed by some practical NEPv, such as those from computational physics and chemistry mentioned above, and it facilitates theoretical analysis. Particularly, it allows us to investigate the NEPv and related numerical algorithms by the eigenspace perturbation theory of the symmetric eigenvalue problem [29, 40]. By exploiting unitary invariance, sufficient conditions for the existence of solutions of NEPv have been established in [10], and estimations of the convergence rate for the often used self-consistent field (SCF) iteration have been obtained in [4, 31, 46]. With (1.2), we can also require that the columns of a solution representative X are eigenvectors of $H(X)$ and Λ is diagonal. Because otherwise we can diagonalize Λ by $Q^{\mathrm{T}}\Lambda Q$ with an orthogonal $Q \in \mathbb{R}^{k\times k}$, and take $(XQ, Q^{\mathrm{T}}\Lambda Q)$ to be the new eigen-matrix pair, since $H(XQ)XQ = XQ(Q^{\mathrm{T}}\Lambda Q)$ by (1.1) and (1.2).

Although most existing studies of NEPv (1.1) focus on the ones with the unitary invariance property (1.2), many recent applications also give rise to NEPv without this property. One important source of such problems is the multi-view subspace learning (see, e.g., [14, 41, 47, 48, 55, 56]), where the problems of interest appear in some forms of the trace-ratio maximization over the Stiefel manifold

$$\mathbb{O}^{n\times k} := \{X \in \mathbb{R}^{n\times k} \,:\, X^{\mathrm{T}}X = I_k\}, \tag{1.3}$$

the collection of matrices in $\mathbb{R}^{n\times k}$ that have orthonormal columns. For example, in [48] it is considered

$$\max_{X\in\mathbb{O}^{n\times k}} f_\theta(X) \quad \text{with} \quad f_\theta(X) := \frac{\mathrm{tr}(X^{\mathrm{T}}AX + X^{\mathrm{T}}D)}{[\mathrm{tr}(X^{\mathrm{T}}BX)]^\theta}, \tag{1.4}$$

where $\theta \in [0, 1]$ is a tunable parameter, A, $B \in \mathbb{R}^{n\times n}$ are symmetric with $B \succ 0$ (positive definite), and $D \in \mathbb{R}^{n\times k}$. Depending on the coefficient matrices and the parameter θ, the optimization (1.4) includes a wide variety of practical cases, and a few special ones have been playing important roles in numerical linear algebra, machine learning, and statistics:

- $D = 0$ and $\theta = 1$ appears in Fisher's linear discriminant analysis (LDA) [35, 53, 54] in the setting of supervised machine learning;
- $A = 0$ and $\theta = 1/2$ shows up in the orthogonal canonical correlation analysis (OCCA) [55];
- $B = I_n$ or $\theta = 0$ arises in the unbalanced orthogonal Procrustes problem, a fundamental problem in numerical linear algebra, optimization, and applied statistics [8, 13, 16, 18, 19, 21, 23, 32, 56, 57].

[1]The term 'unitary' is from the setting of complex matrices, and we stick to this name convention in our discussion. In fact, the results in this paper can be extended to cover the complex case, as we will comment later in Section 8.

The optimization problem (1.4) in its general form has been studied in [48], where it is shown that the first-order optimality condition of (1.4) is equivalent to an NEPv (1.1) with

$$H(X) = \frac{1}{[\mathrm{tr}(X^{\mathrm{T}}BX)]^{\theta}}\left[2\Big(A - \theta\frac{\mathrm{tr}(X^{\mathrm{T}}AX)}{\mathrm{tr}(X^{\mathrm{T}}BX)}\cdot B\Big) + \Big(DX^{\mathrm{T}} + XD^{\mathrm{T}} - 2\theta\frac{\mathrm{tr}(X^{\mathrm{T}}D)}{\mathrm{tr}(X^{\mathrm{T}}BX)}\cdot B\Big)\right]. \quad (1.5)$$

Due to the presences of DX^{T} and XD^{T}, this $H(X)$ does not satisfy the unitary invariance property (1.2). Hence, existing analyses and techniques [4, 10] developed for NEPv with property (1.2) do not directly apply. Especially, the plain SCF iteration does not work for those problems and needs a redesign. In [48, 55, 56], the authors developed SCF-type iterations to solve their particular NEPv and established global convergence for the algorithms, which is rather remarkable since the optimization problems (1.4) are non-convex where global convergence cannot be guaranteed in general. However, their analyses and proofs are problem-specific and cannot be easily extended to other cases, and they do not lead to useful quantitative estimates for rates of convergence of their SCF-type iterations. New convergence theories are needed to better understand and predict the convergence behaviors of such algorithms.

Contribution. The major goal of this paper is twofold: i) to develop a general theory for analyzing a class of NEPv (1.1) that violates the unitary invariance property (1.2) while includes the ones with (1.5) as special cases; ii) to extend the local convergence analysis of [4] to such a class of NEPv. Our analyses will apply to more general NEPv (1.1) than those such as the ones with (1.5), but for now and for clarity, let us first summarize our results for the case of NEPv (1.1) with (1.5) as follows.

- We show that any global optimizer X_* must be a *D-regular eigenbasis matrix*, i.e., satisfying

$$X_*^{\mathrm{T}}D \succeq 0 \qquad \text{and} \qquad \mathrm{rank}(X_*^{\mathrm{T}}D) = \mathrm{rank}(D). \quad (1.6)$$

  The first condition $X_*^{\mathrm{T}}D \succeq 0$ is not hard to establish and has been known to [48, 55, 56], whereas the second condition $\mathrm{rank}(X_*^{\mathrm{T}}D) = \mathrm{rank}(D)$ is new and is a critical one for our local convergence analysis to go through.

- Our analysis is made possible by a novel transformation of the NEPv with (1.5), through a *basis alignment* operation, to an equivalent one that does admit the unitary invariance property and has a differentiable coefficient matrix. The resulting NEPv, which we will call the *aligned NEPv*, is well-defined for all X close to a D-regular eigenbasis matrix. Namely, locally close to a D-regular eigenbasis matrix, the NEPv with (1.5) is unitarily invariantizable.

- We show that the SCF-type iteration for NEPv with (1.5), as developed in [48] and others in [55, 56], is equivalent to the plain SCF iteration for the aligned NEPv. By extending the local convergence analysis in [4], we establish a sharp estimation of the rate of convergence for the SCF-type iteration and build the theoretic foundation for a level-shifting scheme to fix the potential divergence issue.

Extensive numerical experiments are also provided to demonstrate our theoretical results.

Organization. The rest of this paper is organized as follows. In Section 2, we introduce a class of NEPv without unitary invariance as the KKT condition for an optimization problem over the Stiefel manifold. In Section 3, we discuss necessary conditions for an eigenbasis matrix to be a global maximizer of the optimization problem, where we will introduce the notions of basis alignment and D-regular eigenbasis matrix. Sections 4 and 5 are devoted to SCF, where we will propose an SCF-type iteration, establish its connection to the plain SCF for an aligned NEPv that

is unitarily invariant, and obtain its local convergence rate. Section 6 is on a level-shifting scheme with a theoretical foundation for fixing the potential divergence issue of the SCF-type iteration. Numerical experiments are presented in Section 7 and concluding remarks are made in Section 8.

Notation. $\mathbb{R}^{n\times m}$ is the set of n-by-m real matrices, and $\mathbb{O}^{n\times k}$ defined in (1.3) is the Stiefel manifold, where $k \le n$ (usually $k \ll n$) and I_k is the $k \times k$ identity matrix. For a vector or matrix $B \in \mathbb{R}^{m\times n}$, B^{T} stands for its transpose, $\mathcal{R}(B)$ for its column space, and $\mathcal{N}(B)$ for its null space. The singular values of B are denoted by $\sigma_i(B)$, for $i = 1, \ldots, \min\{m, n\}$, arranged in the nonincreasing order: $\sigma_1(B) \ge \sigma_2(B) \ge \cdots \ge \sigma_{\min\{m,n\}}(B)$. $\|B\|$ denotes some consistent matrix norm of B such as the spectral norm and the Frobenius norm. For a square matrix $A \in \mathbb{R}^{n\times n}$, $\mathrm{tr}(A)$ and $\rho(A)$ denote, respectively, its trace and spectral radius (i.e., the largest absolute value of the eigenvalues of A). If A is also symmetric, then its eigenvalues are enumerated from largest to smallest as $\lambda_1(A) \ge \lambda_2(A) \ge \cdots \ge \lambda_n(A)$, and, in particular, $\lambda_{\max}(A) := \lambda_1(A)$ and $\lambda_{\min}(X) := \lambda_n(A)$. $A \succeq 0$ ($A \succ 0$) means that A is symmetric and positive semidefinite (definite). Other notations will be explained at their first appearances.

2 A Class of NEPv Without Unitary Invariance Property

In this section, we introduce a class of NEPv as the first-order optimality condition, also known as the KKT condition, to a particular optimization problem over the Stiefel manifold.

Throughout this paper, for a scalar function f defined on $\mathbb{O}^{n\times k}$, its gradient at $X = [x_{ij}]$ as a matrix variable in $\mathbb{R}^{n\times k}$ is denoted as and defined by

$$\frac{\partial f(X)}{\partial X} \in \mathbb{R}^{n\times k} \quad \text{with} \quad \left[\frac{\partial f(X)}{\partial X}\right]_{ij} := \frac{\partial f(X)}{\partial x_{ij}}. \tag{2.1}$$

More generally, for a (Fréchet) differentiable function $F : \mathbb{R}^{n\times k} \to \mathbb{R}^{p\times q}$, its Fréchet derivative at $X \in \mathbb{R}^{n\times k}$ along direction $Y \in \mathbb{R}^{n\times k}$, denoted as $\mathbf{D}F(X)[Y]$, is defined by

$$\mathbf{D}F(X)[Y] := \lim_{t\to 0} \frac{1}{t}\big[F(X + tY) - F(X)\big] = \frac{d}{dt} F(\,X + tY\,)\Big|_{t=0}. \tag{2.2}$$

We can see that $\mathbf{D}F(X)[\,\cdot\,] : \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times q}$ is a linear operator.

For the gradient in (2.1), we emphasize it being defined at X as a matrix variable in $\mathbb{R}^{n\times k}$, i.e., all entries of X are treated as independent, although X lives on $\mathbb{O}^{n\times k}$. The reader should not confuse it with the notion of gradient over the Stiefel manifold $\mathbb{O}^{n\times k}$ [1, (3.37)].

For ease of presentation, we formally define the notion of *unitary invariance* of a scalar or matrix-valued function on $\mathbb{R}^{n\times k}$ as follows.

Definition 2.1. A function $F\colon \mathbb{R}^{n\times k} \to \mathbb{R}^{p\times q}$ is said *right unitarily invariant*, or *unitarily invariant* for short, if

$$F(XQ) \equiv F(X) \quad \text{for } X \in \mathbb{R}^{n\times k},\ Q \in \mathbb{O}^{k\times k}.$$

2.1 An Optimization Problem on the Stiefel Manifold

Consider the following maximization problem over the Stiefel manifold $\mathbb{O}^{n\times k}$:

$$\max_{X\in\mathbb{O}^{n\times k}} f(X) \qquad \text{with} \qquad f(X) := \phi(X) + \psi(X)\cdot \mathrm{tr}(X^{\mathrm{T}} D), \tag{2.3}$$

where $D \in \mathbb{R}^{n\times k}$, ϕ and ψ are continuously differentiable functions in $X \in \mathbb{R}^{n\times k}$ and unitarily invariant. We assume that ψ is a positive[2] function, i.e., $\psi(X) > 0$, for all $X \in \mathbb{O}^{n\times k}$. Problem (2.3) takes (1.4) as a special case with

$$\phi(X) = \frac{\operatorname{tr}(X^{\mathrm{T}}AX)}{[\operatorname{tr}(X^{\mathrm{T}}BX)]^{\theta}} \quad \text{and} \quad \psi(X) = \frac{1}{[\operatorname{tr}(X^{\mathrm{T}}BX)]^{\theta}}.$$

They are unitarily invariant in X because both $\operatorname{tr}(X^{\mathrm{T}}AX)$ and $\operatorname{tr}(X^{\mathrm{T}}BX)$ share this property (recall that matrix traces are invariant under similarity transformations and, in particular, we have $\operatorname{tr}(Q^{\mathrm{T}}MQ) = \operatorname{tr}(M)$ for all $M \in \mathbb{R}^{k\times k}$ and $Q \in \mathbb{O}^{k\times k}$). If $D = 0$ in (2.3), then the objective function f becomes unitarily invariant, and (2.3) is a maximization problem over the Grassmannian $\mathscr{G}_k^n$.

We will show that the first-order optimality condition, also known as the KKT condition, of the optimization problem (2.3) can be formulated as an NEPv and any optimizer is an orthonormal eigenbasis matrix of the NEPv. Further necessary conditions for the eigenbasis matrix to be a global maximizer will be derived in Section 3.

2.2 From the First-Order Optimality Condition to NEPv

Traditionally, problem (2.3) is viewed as an optimization problem with orthogonality constraints:

$$\max_{X\in\mathbb{R}^{n\times k}} f(X) \quad \text{s.t.} \quad X^{\mathrm{T}}X = I_k. \tag{2.4}$$

We will derive the optimality conditions of (2.4) by the standard method of Lagrange's multipliers.[3] Let us begin with a useful expression for the gradients of unitarily invariant functions.

Lemma 2.1. *Let $\phi\colon \mathbb{R}^{n\times k} \to \mathbb{R}$ be a differentiable function that is unitarily invariant. Then there exists a matrix-valued function $H_\phi(X) \in \mathbb{R}^{n\times n}$, which is continuous and unitarily invariant in $X \in \mathbb{R}^{n\times k}$, and symmetric, such that the gradient of ϕ admits*

$$\frac{\partial\phi(X)}{\partial X} = H_\phi(X)\cdot X \tag{2.5}$$

for $X \in \mathbb{R}^{n\times k}$ with orthonormal columns. Furthermore, H_ϕ can be made differentiable if ϕ is twice differentiable. In general, such $H_\phi(X)$ is not unique, and one of them is

$$H_\phi(X) = \frac{\partial\phi(X)}{\partial X}\cdot X^{\mathrm{T}} + X\cdot\left[\frac{\partial\phi(X)}{\partial X}\right]^{\mathrm{T}} - X\cdot\left(\left[\frac{\partial\phi(X)}{\partial X}\right]^{\mathrm{T}} X\right)\cdot X^{\mathrm{T}}. \tag{2.6}$$

Proof. We will show that the matrix in (2.6) satisfies all the requirement. The validity of the expression (2.5) with (2.6) is a consequence of the following more general fact: Given $X \in \mathbb{R}^{n\times k}$ with $X^{\mathrm{T}}X = I_k$, it holds for any matrix $C \in \mathbb{R}^{n\times k}$ that

$$C = CX^{\mathrm{T}}X = (CX^{\mathrm{T}} + XC^{\mathrm{T}} - XC^{\mathrm{T}})\cdot X = (CX^{\mathrm{T}} + XC^{\mathrm{T}} - XC^{\mathrm{T}}XX^{\mathrm{T}})\cdot X.$$

Letting $C = \frac{\partial\phi(X)}{\partial X}$ yields (2.5) with (2.6). It can been seen that $H_\phi(\,\cdot\,)$ in (2.6) is continuous and also differentiable if ϕ is twice differentiable.

[2]This condition may be relaxed to that ψ is non-negative, i.e., $\psi(X) \geq 0$, and $\psi(X) > 0$ at the KKT points of (2.3). But such conditions are hard to verify since the KKT points are in general unknown in the first place.

[3]Alternatively, we can use the Riemannian gradient of a scalar function over the Stiefel manifold $\mathbb{O}^{n\times k}$ to derive the KKT condition, as in [48] (see also [1, (3.37)]).

It remains to show that H_ϕ in (2.6) is symmetric and unitarily invariant. To that end, we will first establish the following identities for the derivative of unitarily invariant function ϕ: For $X \in \mathbb{R}^{n\times k}$, $Q \in \mathbb{O}^{k\times k}$, and $Y := XQ$, we have

$$\frac{\partial \phi(XQ)}{\partial Q} = X^{\mathrm{T}} \frac{\partial \phi(X)}{\partial X} Q \quad \text{and} \quad \frac{\partial \phi(Y)}{\partial Y} = \frac{\partial \phi(X)}{\partial X} \cdot Q\,. \tag{2.7}$$

These two identities in (2.7) follow directly from the following first-order expansions: Perturbing Q to $Q + \delta Q$ and $Y \equiv XQ$ to $Y + \delta Y$, we have

$$\begin{aligned}
\phi(X(Q+\delta Q)) &= \phi(X(I + \delta Q \cdot Q^{\mathrm{T}})Q) \\
&= \phi(X + X \cdot \delta Q \cdot Q^{\mathrm{T}}) \\
&= \phi(X) + \mathrm{tr}\left(\left[\frac{\partial \phi(X)}{\partial X} \right]^{\mathrm{T}} X \cdot \delta Q \cdot Q^{\mathrm{T}} \right) + \mathcal{O}(\|\delta Q\|^2) \\
&= \phi(X) + \mathrm{tr}\left(Q^{\mathrm{T}} \left[\frac{\partial \phi(X)}{\partial X} \right]^{\mathrm{T}} X \cdot \delta Q \right) + \mathcal{O}(\|\delta Q\|^2),
\end{aligned}$$

and

$$\begin{aligned}
\phi(Y+\delta Y) &= \phi((X + \delta Y \cdot Q^{\mathrm{T}})Q) \\
&= \phi(X + \delta Y \cdot Q^{\mathrm{T}}) \\
&= \phi(X) + \mathrm{tr}\left(\left[\frac{\partial \phi(X)}{\partial X} \right]^{\mathrm{T}} \cdot \delta Y \cdot Q^{\mathrm{T}} \right) + \mathcal{O}(\|\delta Y\|^2) \\
&= \phi(X) + \mathrm{tr}\left(Q^{\mathrm{T}} \left[\frac{\partial \phi(X)}{\partial X} \right]^{\mathrm{T}} \cdot \delta Y \right) + \mathcal{O}(\|\delta Y\|^2).
\end{aligned}$$

Next, we claim that $X^{\mathrm{T}} \frac{\partial \phi(X)}{\partial X}$ is always symmetric, from which the symmetry of $H_\phi(X)$ in (2.6) follows immediately. To verify this claim, we notice that $\phi(XQ)$ with a fixed $X \in \mathbb{R}^{n\times k}$ is a constant over $Q \in \mathbb{O}^{k\times k}$ by unitary invariance. Therefore, any $Q \in \mathbb{O}^{k\times k}$ is an optimal solution to

$$\max_{Q \in \mathbb{R}^{k\times k}} \phi(XQ) \quad \text{s.t.} \quad Q^{\mathrm{T}}Q = I_k,$$

because the objective function has a constant value. By the method of Lagrange's multipliers, we have for any $Q \in \mathbb{O}^{k\times k}$, there exists a symmetric multiplier $\Gamma \in \mathbb{R}^{k\times k}$ such that $\frac{\partial \mathcal{L}(Q,\Gamma)}{\partial Q} = 0$, where

$$\mathcal{L}(Q,\Gamma) := \phi(XQ) - \frac{1}{2}\mathrm{tr}\left([Q^{\mathrm{T}}Q - I_k] \cdot \Gamma \right)$$

is the associated Lagrangian function. By the first identity in (2.7), we obtain immediately

$$\frac{\partial \mathcal{L}(Q,\Gamma)}{\partial Q} = X^{\mathrm{T}} \frac{\partial \phi(X)}{\partial X} Q - Q\Gamma = 0.$$

Letting $Q = I_k$ yields $X^{\mathrm{T}} \cdot \frac{\partial \phi(X)}{\partial X} = \Gamma$, which is symmetric, as was to be shown.

Finally, by the second identity in (2.7), it can be verified that $H_\phi(XQ) = H_\phi(X)$ for $H_\phi(\,\cdot\,)$ given by (2.6). Namely, $H_\phi(\,\cdot\,)$ is unitarily invariant.

We defer addressing the non-uniqueness of $H_\phi(X)$ satisfying (2.5) to Remark 2.1 below. □

Remark 2.1. In general, the choice of continuous and unitarily invariant $H_\phi(X)$ to satisfy (2.5) is not unique. For a particular functions ϕ, one may select an H_ϕ that is more convenient and efficient to work with. For example, for $\phi(X) = h\left(\mathrm{tr}(X^{\mathrm{T}}AX)\right)$, a composition of a differentiable function h with a quadratic form of a symmetric matrix A, we have

$$\frac{\partial \phi(X)}{\partial X} = 2h'\left(\mathrm{tr}(X^{\mathrm{T}}AX)\right) \cdot AX,$$

where h' is the first-order derivative of h. In this case, we have immediately

$$\frac{\partial \phi(X)}{\partial X} = H_\phi(X) \cdot X \qquad \text{with} \qquad H_\phi(X) = 2h'\left(\mathrm{tr}(X^{\mathrm{T}}AX)\right) \cdot A. \tag{2.8}$$

In comparison, the general formula (2.6) leads to

$$H_\phi(X) = 2h'\left(\mathrm{tr}(X^{\mathrm{T}}AX)\right) \cdot \left[AXX^{\mathrm{T}} + XX^{\mathrm{T}}A - X(X^{\mathrm{T}}AX)X^{\mathrm{T}}\right],$$

a much more complicated one than that in (2.8).

By Lemma 2.1, we can write the gradients of unitarily invariant ϕ and ψ as

$$\frac{\partial \phi(X)}{\partial X} = H_\phi(X) \cdot X \qquad \text{and} \qquad \frac{\partial \psi(X)}{\partial X} = H_\psi(X) \cdot X, \tag{2.9}$$

where $H_\phi(X) \in \mathbb{R}^{n\times n}$ and $H_\psi(X) \in \mathbb{R}^{n\times n}$ are symmetric and unitarily invariant. In what follows, we assume some forms of $H_\phi(X)$ and $H_\psi(X)$ have been selected to fulfill (2.9), but we will not specify which particular ones are used. In fact, our development will work with any choice.

Theorem 2.1. *$X \in \mathbb{O}^{n\times k}$ is a KKT point of (2.3) if and only if $D^{\mathrm{T}}X$ is symmetric and X satisfies the NEPv*

$$H(X)X = X\Lambda, \tag{2.10a}$$

where $H(X)$ is symmetric and given by

$$H(X) = H_\phi(X) + \mathrm{tr}(X^{\mathrm{T}}D) \cdot H_\psi(X) + \psi(X) \cdot (DX^{\mathrm{T}} + XD^{\mathrm{T}}). \tag{2.10b}$$

Proof. We treat (2.3) as a constrained optimization problem subject to $X^{\mathrm{T}}X = I_k$ as in (2.4), for which the Lagrangian function of multipliers is written as

$$\mathcal{L}(X, \Gamma) := \phi(X) + \psi(X) \cdot \mathrm{tr}(X^{\mathrm{T}}D) - \frac{1}{2}\mathrm{tr}([X^{\mathrm{T}}X - I_k]\Gamma),$$

where $\Gamma \in \mathbb{R}^{k\times k}$ is a symmetric matrix of multipliers. Then $X \in \mathbb{O}^{n\times k}$ is a KKT point of (2.3) if and only if it satisfies the first-order optimality condition $\frac{\partial \mathcal{L}}{\partial X} = 0$ for some symmetric Γ, i.e.,

$$\frac{\partial \phi(X)}{\partial X} + \mathrm{tr}(X^{\mathrm{T}}D) \cdot \frac{\partial \psi(X)}{\partial X} + \psi(X) \cdot D = X\Gamma,$$

which, by (2.9), is equivalent to

$$H_\phi(X)X + \mathrm{tr}(X^{\mathrm{T}}D) \cdot H_\psi(X)X + \psi(X) \cdot D = X\Gamma. \tag{2.11}$$

If $X \in \mathbb{O}^{n\times k}$ is a KKT point of (2.3), then (2.11) yields (2.10) with $\Lambda = \Gamma + \psi(X) \cdot D^{\mathrm{T}}X$. Moreover, since $\Lambda \equiv X^{\mathrm{T}}H(X)\,X$ is symmetric, we have $D^{\mathrm{T}}X = (\Lambda - \Gamma)/\psi(X)$ is also symmetric.

Conversely, if $X \in \mathbb{O}^{n\times k}$ satisfies (2.10) and $D^{\mathrm{T}}X$ is symmetric, then we have that (2.11) holds with a symmetric $\Gamma = \Lambda - \psi(X) \cdot D^{\mathrm{T}}X$. Hence X is a KKT point. □

For the special case of $D = 0$, the optimization problem (2.3) and NEPv (2.10) become

$$\max_{X \in \mathbb{O}^{n\times k}} \phi(X) \qquad \text{and} \qquad H_\phi(X)X = X\Lambda, \tag{2.12}$$

respectively, where both ϕ and H_ϕ are unitarily invariant. So, optimizing a unitarily invariant function on the Stiefel manifold $\mathbb{O}^{n\times k}$ always leads to a unitarily invariant NEPv. For the general case of $D \neq 0$, due to the presences of DX^{T} and $X^{\mathrm{T}}D$, the coefficient matrix $H(\,\cdot\,)$ in (2.10b) is not unitarily invariant any more; see the particular example (1.5), obtained by an application of Theorem 2.1 to (1.4).

As a first-order optimality condition, Theorem 2.1 does not specify which k eigenvalues of $H(X)$ correspond to those of Λ for the purpose of solving the underlying optimization problem. For a local maximizer X of (2.3), the corresponding eigenvalues are typically the k largest ones. This has been proven for particular optimizations in the form of (2.3) (e.g., in [48, 51, 53, 54, 55, 56]), and it is also a common practice for handling the NEPv from related optimization problems in electronic structure calculation (e.g., in [11, 50]), where the eigenvalues are the k smallest ones because they are about minimization. But there exist cases for which not all of the k eigenvalues are the extreme ones, and when that happens, numerical difficulties arise. We will come back to this issue in Section 6.

3 Necessary Conditions for Global Maximizers

Ideally, we should seek those solutions of NEPv (2.10) that are the global maximizers of optimization problem (2.3). For that purpose, we will first establish two necessary conditions for a global maximizer of (2.3), beyond its KKT condition given as NEPv (2.10). Specifically, they are

$$\textit{definiteness:} \quad X^{\mathrm{T}}D \succeq 0, \tag{3.1}$$

$$\textit{rank-preserving:} \quad \operatorname{rank}(X^{\mathrm{T}}D) = \operatorname{rank}(D). \tag{3.2}$$

The two conditions above will serve as guides to what solutions to NEPv (2.10) we should look for. The definiteness condition (3.1) has been known and successfully exploited in existing works for a few special cases of (2.3); see, e.g., [48] and reference therein. In contrast, the rank-preserving condition (3.2) is a new discovery, which is also a crucially missing piece for analyzing the local convergence of SCF for solving NEPv (2.10). In fact, this new condition makes our systematic treatment in the rest of this paper possible.

Our main theorem in this section is Theorem 3.1 below, and its proof will be given towards the end of this section after we fully investigate the definiteness condition (3.1) and the rank-preserving condition (3.2) separately.

Theorem 3.1. *Let $X_* \in \mathbb{O}^{n\times k}$ be a global maximizer of* (2.3)*. Then X_* satisfies both the definiteness and rank-preserving conditions, i.e.,* (3.1) *and* (3.2) *hold with $X = X_*$.*

3.1 Definiteness and Basis Alignment

Let $X \in \mathbb{O}^{n\times k}$ be a given approximate solution to optimization problem (2.3). Since the objective function $f(X)$ is not unitarily invariant, it is possible to find a better approximate solution (in the sense of a larger objective value) in the form of XQ with $Q \in \mathbb{O}^{k\times k}$ determined by

$$\max_{Q \in \mathbb{O}^{k\times k}} f(XQ). \tag{3.3}$$

As we will show shortly, the construction of such best XQ has a lot to do with the definiteness condition (3.1). We will refer to finding the best XQ via (3.3) as the problem of *basis alignment* because essentially Q picks up a new orthonormal basis matrix $\widetilde{X} = XQ$ for the column space of X. In fact, any maximizer Q of (3.3) will yield an improved solution $\widetilde{X}$ for optimization problem (2.3), in that $f(\widetilde{X}) \geq f(X)$, and usually the inequality is strict unless $Q = I_k$ is a maximizer of (3.3). This idea can be regarded as seeking a best solution to (2.3) in the orbit $\{XQ : Q \in \mathbb{O}^{k\times k}\}$. For convenience of discussion, we introduce an *alignment function* for $X \in \mathbb{O}^{n\times k}$ as

$$[\![X]\!] := \left\{ XQ : \, Q \in \operatorname*{arg\,max}_{Q\in\mathbb{O}^{k\times k}} f(XQ) \right\}. \tag{3.4}$$

Note that $[\![X]\!]$ may be multi-valued.

From the definition of f in (2.3), since both ϕ and ψ are unitarily invariant, we have

$$f(XQ) = \phi(X) + \psi(X) \cdot \operatorname{tr}(Q^{\mathrm{T}} X^{\mathrm{T}} D). \tag{3.5}$$

By assumption $\psi(X) > 0$, problem (3.3) is therefore equivalent to

$$\max_{Q\in\mathbb{O}^{k\times k}} \operatorname{tr}\left(Q^{\mathrm{T}}(X^{\mathrm{T}}D)\right). \tag{3.6}$$

Problem (3.6) is a classical matrix optimization problem, which also arises in, e.g., the orthogonal Procrustes [20, Section 6.4.1], and it is known to have a closed-form solution via SVD, as shown in Lemma 3.1, but the latter reveals more structural properties in solution than those traditionally known. We notice that the ideas for the alignment (3.4) and its solution via (3.6) have occurred in previous works on special cases of (3.3); see, e.g., [48, 55, 56]. Recently, the authors of [43] viewed (3.6) as an optimization problem over the orthonormal basis matrices of the given subspace $\mathcal{R}(X)$ and investigated how the optimal basis matrices vary as the subspace varies.

Lemma 3.1 below is essentially [48, Lemma 3.2] but stated differently, and a proof is provided here for self-containedness and it is also simpler than the one in [48].

Lemma 3.1 ([48])**.** *Let the singular value decomposition (SVD) of $X^{\mathrm{T}}D \in \mathbb{R}^{k\times k}$ be*

$$X^{\mathrm{T}}D = U\Sigma V^{\mathrm{T}} \equiv \begin{array}{c} \begin{matrix} \ell & k-\ell \end{matrix} \\ \begin{bmatrix} U_1 & U_2 \end{bmatrix} \end{array} \times \begin{array}{c} \begin{matrix} \ell \\ k-\ell \end{matrix} \\ \begin{bmatrix} \Sigma_1 & \\ & 0 \end{bmatrix} \end{array} \times \begin{array}{c} \begin{matrix} \ell \\ k-\ell \end{matrix} \\ \begin{bmatrix} V_1^{\mathrm{T}} \\ V_2^{\mathrm{T}} \end{bmatrix} \end{array}, \tag{3.7}$$

where $\ell = \operatorname{rank}(X^{\mathrm{T}}D)$ and $\Sigma_1 = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_\ell) \succ 0$. Then,

(a) *$Q \in \mathbb{O}^{k\times k}$ is a global maximizer of (3.3) if and only if*

$$Q = U \begin{bmatrix} I_\ell & \\ & \Omega \end{bmatrix} V^{\mathrm{T}}, \tag{3.8}$$

where $\Omega \in \mathbb{O}^{(n-\ell)\times(n-\ell)}$ is arbitrary;

(b) *$Q \in \mathbb{O}^{k\times k}$ takes the form of (3.8) if and only if $Q^{\mathrm{T}}(X^{\mathrm{T}}D) \succeq 0$.*

Proof. For item (a), it follows from SVD (3.7) that

$$\operatorname{tr}\left(Q^{\mathrm{T}}[X^{\mathrm{T}}D]\right) = \operatorname{tr}\left(\widetilde{Q}^{\mathrm{T}} \begin{bmatrix} \Sigma_1 & \\ & 0 \end{bmatrix}\right) = \sum_{i=1}^{\ell} \sigma_i \cdot \widetilde{Q}_{ii} \leq \sum_{i=1}^{\ell} \sigma_i, \tag{3.9}$$

where $\widetilde{Q} = U^{\mathrm{T}}QV$ is orthogonal and $\widetilde{Q}_{ii}$ is the ith diagonal entry of $\widetilde{Q}$. The last equation in (3.9) is due to $|\widetilde{Q}_{ii}| \leq 1$ for all i. Clearly, the right hand side achieves the maximum $\sum_{i=1}^{\ell} \sigma_i$ if and only if $\widetilde{Q}_{ii} = 1$ for $1 \leq i \leq \ell$. Therefore, the optimal $\widetilde{Q} = \begin{bmatrix} I_\ell & \\ & \Omega \end{bmatrix}$, where $\Omega \in \mathbb{O}^{(n-\ell)\times(n-\ell)}$. Substituting it back to $Q = U\widetilde{Q}V^{\mathrm{T}}$, we obtain (3.8).

For item (b), it can be verified that $Q^{\mathrm{T}}(X^{\mathrm{T}}D) \succeq 0$ for Q given by (3.8). Conversely, if $Q \in \mathbb{O}^{k\times k}$ satisfies $Q^{\mathrm{T}}(X^{\mathrm{T}}D) \succeq 0$, then

$$\mathrm{tr}\left(Q^{\mathrm{T}}(X^{\mathrm{T}}D)\right) = \sum_{i=1}^{\ell} \sigma_i\left(Q^{\mathrm{T}}(X^{\mathrm{T}}D)\right) = \sum_{i=1}^{\ell} \sigma_i\left(X^{\mathrm{T}}D\right) = \sum_{i=1}^{\ell} \sigma_i,$$

where the first equality is because of the eigenvalues of $Q^{\mathrm{T}}(X^{\mathrm{T}}D)$ are the same as its singular values. By (3.9), Q is a global maximizer of (3.6) and must take the form of (3.8) by item (a). $\square$

The results in Lemma 3.1 can alternatively be interpreted using the polar decomposition of $X^{\mathrm{T}}D$, as done traditionally for the solution of the Procrustes problem. The polar decomposition of $X^{\mathrm{T}}D$ refers to

$$X^{\mathrm{T}}D = QM \quad \text{with } Q \in \mathbb{O}^{k\times k} \text{ and } M \succeq 0. \tag{3.10}$$

The polar factors Q and M in (3.10) can be expressed as Q in (3.8) and $M = V\Sigma V^{\mathrm{T}}$, via SVD (3.7). If $X^{\mathrm{T}}D$ is non-singular, then the polar decomposition (3.10) is unique, and thus the orthogonal polar factor $Q = UV^{\mathrm{T}}$ is also uniquely defined (i.e., independent of any inherent freedom in SVD (3.7)); see, e.g., [7, p.220], [22, Chapter 8], and also [27, 28, 29, 30]. Otherwise, when $X^{\mathrm{T}}D$ is singular, orthogonal polar factor Q is non-unique, and Lemma 3.1 reveals the inherent structure of optimal Q with the form of (3.8). The structure of this solution form is crucial in our subsequent analysis.

By Lemma 3.1, any orthogonal polar factor Q in (3.10) is a maximizer of (3.6), and vice versa. Moreover,

$$\max_{Q\in\mathbb{O}^{k\times k}} \mathrm{tr}\left(Q^{\mathrm{T}}(X^{\mathrm{T}}D)\right) = \mathrm{tr}\left(M\right). \tag{3.11}$$

The following results are direct consequences of Lemma 3.1.

Corollary 3.1. *Given $X \in \mathbb{O}^{n\times k}$, let the SVD of $X^{\mathrm{T}}D$ be given by (3.7) with $\ell = \mathrm{rank}(X^{\mathrm{T}}D)$, and define $[\![X]\!]$ as in (3.4).*

(a) *We have*

$$[\![X]\!] = \left\{ XQ :\ Q \in \underset{Q\in\mathbb{O}^{k\times k}}{\arg\max}\, \mathrm{tr}\left(Q^{\mathrm{T}}(X^{\mathrm{T}}D)\right) \right\} \tag{3.12a}$$

$$= \left\{ XQ :\ Q = U\begin{bmatrix} I_\ell & \\ & \Omega \end{bmatrix} V^{\mathrm{T}},\ \Omega \in \mathbb{O}^{(k-\ell)\times(k-\ell)} \right\}. \tag{3.12b}$$

Thus, $[\![X]\!] = \{X\left(UV^{\mathrm{T}}\right)\}$ contains just one element if $\ell = k$, but if $\ell < k$ then $[\![X]\!]$ contains infinitely many elements.

(b) *$\widetilde{X} \in [\![X]\!]$ if and only if $\widetilde{X} = XQ$ for some $Q \in \mathbb{O}^{k\times k}$ such that $\widetilde{X}^{\mathrm{T}}D \succeq 0$.*

(c) *$X \in [\![X]\!]$ if and only if $X^{\mathrm{T}}D \succeq 0$.*

(d) *$[\![X]\!] = [\![XQ]\!]$ for any $Q \in \mathbb{O}^{k\times k}$.*

(e) *$D^{\mathrm{T}}X_a = D^{\mathrm{T}}X_b$ and $X_aD^{\mathrm{T}}X = X_bD^{\mathrm{T}}X$ for any two elements $X_a, X_b \in [\![X]\!]$.*

Proof. For item (a), equation (3.12a) follows from the equivalency between optimization problems (3.3) and (3.6), while equation (3.12b) just restates Lemma 3.1(a). That $[\![X]\!]$ has just one element when $\ell = k$ is a consequence of the uniqueness of the corresponding polar decomposition (2.12) [29]. Item (b) restates Lemma 3.1(b), which then leads to item (c). For item (d), we notice that $[\![X]\!]$ consists of all orthonormal basis matrices of the subspace $\mathcal{R}(X)$, at which f achieves its maximum in the orbit $\{XQ : Q \in \mathbb{O}^{k\times k}\}$. Since $\mathcal{R}(X) = \mathcal{R}(XQ)$, we have $[\![X]\!] = [\![XQ]\!]$. Finally, the equations in item (e) can be verified by the general expression for Q in (3.8). □

3.2 Rank-preserving

We now establish several equivalent statements to rank-preserving condition (3.2).

Theorem 3.2. *Let $X \in \mathbb{O}^{n\times k}$ and $H(\,\cdot\,)$ be as in* (2.10b)*. The following statements are equivalent.*

(a) *X satisfies rank-preserving condition* (3.2)*, i.e.,* $\operatorname{rank}(X^{\mathrm{T}}D) = \operatorname{rank}(D)$.

(b) *$DV_2 = 0$, where V_2, a basis matrix of the null space $\mathcal{N}(X^{\mathrm{T}}D)$, is from SVD* (3.7).

(c) *$DX_a^{\mathrm{T}} = DX_b^{\mathrm{T}}$ for all $X_a, X_b \in [\![X]\!]$.*

(d) *$H(X_a) = H(X_b)$ for all $X_a, X_b \in [\![X]\!]$.*

(e) *$H(X_a)X = H(X_b)X$ for all $X_a, X_b \in [\![X]\!]$.*

Proof. (a) ⇔ (b): Apply the rank-nullity theorem to D and $X^{\mathrm{T}}D$, respectively, to get

$$\operatorname{rank}(D) + \dim(\mathcal{N}(D)) = k = \operatorname{rank}(X^{\mathrm{T}}D) + \dim(\mathcal{N}(X^{\mathrm{T}}D)).$$

Hence, $\operatorname{rank}(D) = \operatorname{rank}(X^{\mathrm{T}}D)$ if and only if $\dim(\mathcal{N}(D)) = \dim(\mathcal{N}(X^{\mathrm{T}}D))$. The latter, since $\mathcal{N}(D) \subset \mathcal{N}(X^{\mathrm{T}}D)$, holds if and only if $\mathcal{N}(X^{\mathrm{T}}D) \subset \mathcal{N}(D)$, which is equivalent to $DV_2 = 0$ with V_2 from SVD (3.7).

(b) ⇔ (c): It follows from (3.8) that there are Q_a and Q_b taking the form of (3.8) such that

$$X_a = XQ_a = X(U_1V_1^{\mathrm{T}} + U_2\Omega_aV_2^{\mathrm{T}}), \quad X_b = XQ_b = X(U_1V_1^{\mathrm{T}} + U_2\Omega_bV_2^{\mathrm{T}}), \tag{3.13}$$

for some $\Omega_a, \Omega_b \in \mathbb{O}^{(k-\ell)\times(k-\ell)}$. If $DV_2 = 0$, then

$$DX_a^{\mathrm{T}} = DV_1U_1^{\mathrm{T}}X^{\mathrm{T}} = DX_b^{\mathrm{T}},$$

which establishes item (c). Conversely, if item (c) holds, then by taking two particular X_a and X_b from (3.13) with $\Omega_a = I_{k-\ell}$ and $\Omega_b = -I_{k-\ell}$, we have

$$0 = DX_a^{\mathrm{T}} - DX_b^{\mathrm{T}} = 2 \cdot DV_2U_2^{\mathrm{T}}X^{\mathrm{T}}.$$

Post-multiplication by XU_2 yields $DV_2 = 0$, as expected.

(c) ⇒ (d): Recall the definition of $H(\,\cdot\,)$ in (2.10b):

$$H(X) = H_\phi(X) + \operatorname{tr}(X^{\mathrm{T}}D) \cdot H_\psi(X) + \psi(X) \cdot (DX^{\mathrm{T}} + XD^{\mathrm{T}}),$$

where H_ϕ, H_ψ, and $\psi(X) > 0$ are all unitarily invariant. For $X_a = XQ_a$ and $X_b = XQ_b$,

$$H(X_a) = H_\phi(X) + \operatorname{tr}(X_a^{\mathrm{T}}D) \cdot H_\psi(X) + \psi(X) \cdot (DX_a^{\mathrm{T}} + X_aD^{\mathrm{T}}), \tag{3.14a}$$
$$H(X_b) = H_\phi(X) + \operatorname{tr}(X_b^{\mathrm{T}}D) \cdot H_\psi(X) + \psi(X) \cdot (DX_b^{\mathrm{T}} + X_bD^{\mathrm{T}}). \tag{3.14b}$$

By the assumption of item (c): $DX_a^{\mathrm{T}} = DX_b^{\mathrm{T}}$, we get $X_aD^{\mathrm{T}} = X_bD^{\mathrm{T}}$ by taking transpose, and

$$\mathrm{tr}(X_a^{\mathrm{T}}D) = \mathrm{tr}(DX_a^{\mathrm{T}}) = \mathrm{tr}(DX_b^{\mathrm{T}}) = \mathrm{tr}(X_b^{\mathrm{T}}D),$$

immediately implying $H(X_a) = H(X_b)$ by (3.14).

(d) $\Rightarrow$ (e): This holds trivially.

(e) $\Rightarrow$ (c): It follows from (3.14) and Corollary 3.1(e) that

$$0 = H(X_a)X - H(X_b)X = \psi(X) \cdot (DX_a^{\mathrm{T}}X - DX_b^{\mathrm{T}}X).$$

Since $\psi(X) > 0$, we conclude $DX_a^{\mathrm{T}}X = DX_b^{\mathrm{T}}X$ and thus $D(X_a^{\mathrm{T}}X)X^{\mathrm{T}} = D(X_b^{\mathrm{T}}X)X^{\mathrm{T}}$. Plug in $X_a = XQ_a$ and $X_b = XQ_b$ to get $DX_a^{\mathrm{T}} = DX_b^{\mathrm{T}}$. The proof is completed. □

3.3 Proof of Theorem 3.1

Now we are ready to prove our main Theorem 3.1 of the section.

Proof of Theorem 3.1. Let $\widetilde{X} \in [\![X_*]\!]$. It follows from (3.4) that

$$f(X_*) \le f(\widetilde{X}) \le \max_{X \in \mathbb{O}^{n\times k}} f(X) = f(X_*), \tag{3.15}$$

where the equality is by the global maximality of f at X_*. So all equalities in (3.15) must hold, and thus we have

$$X_* \in [\![X_*]\!] \qquad \text{and} \qquad \widetilde{X} \in \operatorname*{arg\,max}_{X \in \mathbb{O}^{n\times k}} f(X), \tag{3.16}$$

recalling the definition in (3.4). By $X_* \in [\![X_*]\!]$ and Corollary 3.1(c), we have $X_*^{\mathrm{T}}D \succeq 0$.

It remains to show $\mathrm{rank}(X_*^{\mathrm{T}}D) = \mathrm{rank}(D)$. Equivalently, we will prove the statement in Theorem 3.2(e) for $X = X_*$. We first recall (3.16) that any $\widetilde{X} \in [\![X_*]\!]$ is a global maximizer of (2.3) and, therefore, an eigenbasis matrix of NEPv (2.10):

$$H(\widetilde{X})\widetilde{X} = \widetilde{X}\widetilde{\Lambda} \qquad \text{with} \qquad \widetilde{\Lambda} = \widetilde{X}^{\mathrm{T}}H(\widetilde{X})\widetilde{X}.$$

Since $\widetilde{X} = X_*Q$ for some $Q \in \mathbb{O}^{k\times k}$, we have

$$H(\widetilde{X})X_* = X_*Q\widetilde{\Lambda}Q^{\mathrm{T}} = X_* \cdot (X_*^{\mathrm{T}}H(\widetilde{X})X_*).$$

Hence, for any two $X_a, X_b \in [\![X_*]\!]$,

$$H(X_a)X_* = X_* \cdot \big(X_*^{\mathrm{T}}H(X_a)X_*\big) = X_* \cdot \big(X_*^{\mathrm{T}}H(X_b)X_*\big) = H(X_b)X_*,$$

where for obtaining the second equality, we have used $X_*^{\mathrm{T}}H(X_a)X_* = X_*^{\mathrm{T}}H(X_b)X_*$, which can be verified straightforwardly with the help of (3.14) and Corollary 3.1(e) with $X = X_*$. □

By now it is clear that, as far as the global maximality of optimization problem (2.3) is concerned, we should at least seek a solution X to NEPv (2.10) such that both definiteness condition (3.1) and rank-preserving condition (3.2) are satisfied (although such X is not guaranteed to be a global maximizer). Our developments in the rest of this paper focus on finding such a solution. Notice that, although the two conditions are consequences of studying optimization problem (2.3), they are really about the two matrices $X \in \mathbb{O}^{n\times k}$ and $D \in \mathbb{R}^{n\times k}$ only. Given the importance of both conditions (3.1) and (3.2) on X, we hence introduce the notion of *D-regularity* for the solution of NEPv (2.10).

Definition 3.1. An eigenbasis matrix $X \in \mathbb{O}^{n\times k}$ of NEPv (2.10) is called *D-regular* if it satisfies both definiteness condition (3.1) and rank-preserving condition (3.2).

In the case of rank-deficient D, a D-regular eigenbasis matrix X of NEPv (2.10) will have infinitely many equivalent solutions given by $\widetilde{X} = XQ$ with Q of form (3.8), i.e., $\widetilde{X} \in [\![X]\!]$ is an arbitrary alignment: Clearly, $\widetilde{X}$ so defined satisfies both conditions (3.1) and (3.2), and we also have $H(X) = H(\widetilde{X})$, due to the rank-preserving condition and Theorem 3.2(d), and so $H(\widetilde{X})\widetilde{X} = \widetilde{X}\widetilde{\Lambda}$ with $\widetilde{\Lambda} = Q^{\mathrm{T}}\Lambda Q$. Since any alignment $\widetilde{X} \in [\![X]\!]$ is still a D-regular eigenbasis matrix, we view them as equivalent solutions.

4 SCF Iteration

The self-consistent field (SCF) iteration is one of the most general and widely used methods for solving NEPv [10]. The term SCF comes from the community of computational physics and chemistry in solving the Kohn-Sham equation in the density functional theory [33, 42], and it has since been widely adopted to refer to similar ideas for solving any general NEPv (1.1). In this section, we propose an SCF-type iteration for NEPv (2.10), with the purpose of solving optimization problem (2.3) over the Stiefel manifold $\mathbb{O}^{n\times k}$ in mind. The SCF-type iteration is essentially due to [48, 55, 56], where special cases of (2.3) were considered.

From the perspective of the underlying optimization problem (2.3), at each step, our SCF-type iteration seeks a new approximate solution of NEPv (2.10) that fulfills definiteness condition (3.1) and meanwhile increases the objective value. It is rather easy to make $X_i^{\mathrm{T}}D \succeq 0$ for each iterate X_i by alignment, but monotonically increasing the objective value is highly nontrivial. Inspired by the special cases in [48, 55, 56], where using the eigenspace of $H(X_i)$ associated with its k largest eigenvalues as X_{i+1} can always increase the objective value, and also for the sake of explaining our convergence analysis, we will state our SCF-type iteration using the top eigenspace. But we also caution that such an approach may not always increase the objective value in general; when that happens, the so-called *level-shifting* can often be applied, and we will return to this in Section 6.

Our SCF-type iteration for NEPv (2.10) can be described as follows: Starting from an initial guess $X_0 \in \mathbb{O}^{n\times k}$, generate sequentially $X_1, X_2, \ldots$, all in $\mathbb{O}^{n\times k}$, by solving the symmetric eigenvalue problems

$$H(X_i)\widetilde{X}_{i+1} = \widetilde{X}_{i+1}\widetilde{\Lambda}_{i+1} \qquad \text{for } i = 0, 1, 2, \ldots, \tag{4.1}$$

where $\widetilde{X}_{i+1} \in \mathbb{O}^{n\times k}$ is an orthonormal eigenbasis matrix of $H(X_i)$ associated with its k largest eigenvalues. Here, the eigenbasis matrix $\widetilde{X}_{i+1}$ may not automatically satisfy definiteness condition (3.1). Since we are interested in D-regular eigenbasis matrices satisfying (3.1), we can align $\widetilde{X}_{i+1}$ after solving (4.1) to

$$X_{i+1} \in [\![\widetilde{X}_{i+1}]\!], \tag{4.2}$$

namely, X_{i+1} is taken from $[\![\widetilde{X}_{i+1}]\!]$ (and hence $X_{i+1}^{\mathrm{T}}D \succeq 0$), which can be obtained by the SVD (or polar decomposition) of $\widetilde{X}_{i+1}^{\mathrm{T}}D$, recalling (3.8). We therefore arrive at an SCF-type iteration summarized in Algorithm 4.1, which is essentially from [48, 55, 56] as mentioned.

In Algorithm 4.1, we assume that the initial basis matrix satisfies $X_0^{\mathrm{T}}D \succeq 0$; otherwise, it can be aligned, i.e., $X_0 \in [\![X_0]\!]$ as in (4.2). We have also left out the stopping criterion for the SCF-loop. As inspired by the common practice for linear eigenvalue problems [2, 3, 17], we introduce the normalized residual norm to gauge the accuracy of an approximate solution X of NEPv (2.10):

$$\mathrm{NRes}(X) := \frac{\|H(X)X - X[X^{\mathrm{T}}H(X)X]\|}{\|H(X)\|}, \tag{4.3}$$

Algorithm 4.1 An SCF-type iteration for NEPv (2.10)

Input: $X_0 \in \mathbb{O}^{n\times k}$ such that $X_0^{\mathrm{T}} D \succeq 0$;
Output: a solution to NEPv (2.10) for the purpose of solving optimization problem (2.3).
1: **for** $i = 0, 1, 2, \ldots$ until convergence **do**
2: solve the symmetric eigenproblem $H(X_i)\widetilde{X}_{i+1} = \widetilde{X}_{i+1}\widetilde{\Lambda}_{i+1}$, where $\widetilde{X}_{i+1} \in \mathbb{O}^{n\times k}$ contains the eigenvectors for the k largest eigenvalues of $H(X_i)$;
3: align $\widetilde{X}_{i+1}$ to get $X_{i+1} \in [\![\widetilde{X}_{i+1}]\!]$;
4: **end for**
5: **return** the last X_i as a solution to NEPv (2.10).

where $\|\cdot\|$ is some matrix norm that is convenient to evaluate, e.g., the matrix 1-norm (the maximum column sum in absolute value) or Frobenius norm (the square root of the sum of squares of all elements). Given a tolerance `tol`, a small number, the SCF-loop is considered converged if $\mathrm{NRes}(X_i) \le$ `tol`.

For a unitarily invariant NEPv, alignment (4.2) becomes unnecessary. For such NEPv, the local convergence behavior of the plain SCF, i.e., (4.1) with $X_{i+1} = \widetilde{X}_{i+1}$, has been extensively studied, and various improvements for SCF have been developed; see, e.g., [4, 10, 12, 31, 39, 46, 49]. In particular, the recent work [4] established a sharp estimation for the local convergence rate of SCF, and it also provided a theoretical guarantee for the effectiveness of a level-shifting technique to fix the potential divergence issue of SCF.

For Algorithm 4.1, due to the non-unitary invariance of $H(X)$ and the extra alignment step in line 3, most existing analysis and techniques for the plain SCF do not apply directly. Although several recent works [48, 55, 56] proved the convergence of such algorithms for their respective special cases, their analyses are specialized and do not apply to general NEPv (2.10), nor do they produce sharp estimation of convergence rate. New techniques are needed to better understand the convergence (or divergence) of Algorithm 4.1.

5 Local Convergence Analysis

We will start by showing that, locally near a D-regular solution of NEPv (2.10), the SCF-type iteration in Algorithm 4.1 is the plain SCF for an equivalent NEPv that is unitarily invariant. By this connection, we can conveniently perform its local convergence analysis along the line of [4]. In particular, a closed-form local convergence rate of Algorithm 4.1 will be established, and a theoretical foundation for a level-shifted variant of SCF will be built.

5.1 Aligned NEPv

To begin with, denote by $\mathbb{O}_D^{n\times k}$ the set of $X \in \mathbb{O}^{n\times k}$ satisfying rank-preserving condition (3.2):

$$\mathbb{O}_D^{n\times k} := \{X \in \mathbb{O}^{n\times k} \,:\, \mathrm{rank}(X^{\mathrm{T}} D) = \mathrm{rank}(D)\}. \tag{5.1}$$

Given $X \in \mathbb{O}_D^{n\times k}$, the matrix $H(\widetilde{X})$ will not change as $\widetilde{X}$ varies in $[\![X]\!]$, due to Theorem 3.2(d), and so $H([\![X]\!]) := H(\widetilde{X})$ for $\widetilde{X} \in [\![X]\!]$ is well-defined. We hence introduce

$$G(X) := H([\![X]\!]) \quad \text{for } X \in \mathbb{O}_D^{n\times k}, \tag{5.2}$$

and call G the *aligned* function of H in (2.10b). Accordingly, we introduce the *aligned NEPv* of NEPv (2.10) as

$$G(X)X = X\Lambda \quad \text{for } X \in \mathbb{O}_D^{n\times k}. \tag{5.3}$$

By Corollary 3.1(d), the aligned function G in (5.2) is unitarily invariant, i.e.,

$$G(XQ) = G(X) \quad \text{for} \quad X \in \mathbb{O}_D^{n\times k} \text{ and } Q \in \mathbb{O}^{k\times k}, \tag{5.4}$$

that is, the aligned NEPv (5.3) is an unitarily invariant NEPv.

We emphasize that both the aligned matrix-valued function G in (5.2) and the aligned NEPv (5.3) have their domain of existence $\mathbb{O}_D^{n\times k}$. Luckily, the domain $\mathbb{O}_D^{n\times k}$ is a relative open set on the Stiefel manifold, because rank-preserving condition (3.2) always hold under small perturbations, as shown in Lemma 5.1 below.

Lemma 5.1. *Let $Y \in \mathbb{O}_D^{n\times k}$. For $X \in \mathbb{O}^{n\times k}$, if $\mathcal{R}(X)$ is sufficiently close to $\mathcal{R}(Y)$, then $X \in \mathbb{O}_D^{n\times k}$.*

Proof. The sufficient closeness of $\mathcal{R}(X)$ to $\mathcal{R}(Y)$ implies that there exists $Q \in \mathbb{O}^{k\times k}$ such that $\|XQ - Y\|$ is sufficiently small. The condition $\operatorname{rank}(Y^{\mathrm{T}}D) = \operatorname{rank}(D)$ implies that $Y^{\mathrm{T}}D$ has exactly $\ell = \operatorname{rank}(D)$ nonzero singular values. Notice

$$Q^{\mathrm{T}}X^{\mathrm{T}}D = Y^{\mathrm{T}}D + (XQ - Y)^{\mathrm{T}}D,$$

where the second term on the right-hand side is in the order of $\|XQ - Y\|$, which can be made as small as needed by letting $\mathcal{R}(X)$ be sufficiently close to $\mathcal{R}(Y)$. Because the singular values of a matrix are continuous with respect to matrix entries [40, 29], $Q^{\mathrm{T}}X^{\mathrm{T}}D$ has at least ℓ nonzero singular values, implying $\operatorname{rank}(Q^{\mathrm{T}}X^{\mathrm{T}}D) \geq \ell$. On the other hand, $\operatorname{rank}(Q^{\mathrm{T}}X^{\mathrm{T}}D) = \operatorname{rank}(X^{\mathrm{T}}D) \leq \operatorname{rank}(D) = \ell$. Hence, we conclude $\operatorname{rank}(X^{\mathrm{T}}D) = \operatorname{rank}(D)$, as was to be shown. □

Let $X \in \mathbb{O}_D^{n\times k}$. The *canonical polar decomposition* of $X^{\mathrm{T}}D$ [22, Chapter 8] refers to

$$X^{\mathrm{T}}D = Q_{\mathbf{o}}M, \tag{5.5a}$$

where, in terms of SVD (3.7),

$$Q_{\mathbf{o}} = U_1V_1^{\mathrm{T}} \qquad \text{and} \qquad M = V_1\Sigma_1V_1^{\mathrm{T}}. \tag{5.5b}$$

By definition, $X^{\mathrm{T}}D$ is factorized as the product of a partial isometry $Q_{\mathbf{o}}$ (i.e., $\|Q_{\mathbf{o}}x\| = \|x\|$ for all $x \in \mathcal{R}(Q_{\mathbf{o}}^{\mathrm{T}})$ where $\|\cdot\|$ is the Euclidian vector norm) and a symmetric positive semi-definite M. Note that $Q_{\mathbf{o}} \in \mathbb{R}^{k\times k}$ is not necessarily orthogonal, as in contrast to an orthogonal factor of polar decomposition (3.10). The canonical polar factors $Q_{\mathbf{o}}$ and M in (5.5b) are uniquely defined (i.e., independent of any freedom in the SVD; see, e.g., [7, p.220], [22, Chapter 8], and [27, 28, 29, 30]).

Over domain $\mathbb{O}_D^{n\times k}$, $G(X)$ also admits an expression in terms of the canonical polar factors of $X^{\mathrm{T}}D$, with which we can conveniently show the differentiability of $G(X)$ and obtain its derivative.

Lemma 5.2. *Let $X \in \mathbb{O}_D^{n\times k}$ and $G(X)$ be defined by* (5.3)*, and let $X^{\mathrm{T}}D$ have the canonical polar decomposition in* (5.5)*. Then we have*

$$G(X) = H_\phi(X) + G_\psi(X), \tag{5.6a}$$

where

$$G_\psi(X) := \operatorname{tr}(M) \cdot H_\psi(X) + \psi(X) \cdot \left(DQ_{\mathbf{o}}^{\mathrm{T}}X^{\mathrm{T}} + XQ_{\mathbf{o}}D^{\mathrm{T}}\right), \tag{5.6b}$$

and $H_\phi(X)$ and $H_\psi(X)$ are from (2.9)*.*

Proof. For any $\widetilde{X} \in [\![X]\!]$, it follows from the expression (3.14) of H that

$$H(\widetilde{X}) = H_\phi(X) + \mathrm{tr}(\widetilde{X}^{\mathrm{T}} D) \cdot H_\psi(X) + \psi(X) \cdot (D\widetilde{X}^{\mathrm{T}} + \widetilde{X} D^{\mathrm{T}}). \tag{5.7}$$

Recall SVD (3.7). By $\mathrm{rank}(X^{\mathrm{T}} D) = \mathrm{rank}(D)$ and Theorem 3.2(b), we have $DV_2 = 0$, which implies

$$D\widetilde{X}^{\mathrm{T}} = D\left(V_1 U_1^{\mathrm{T}} + V_2 \Omega^{\mathrm{T}} U_2^{\mathrm{T}}\right) X^{\mathrm{T}} = D \cdot (V_1 U_1^{\mathrm{T}}) \cdot X^{\mathrm{T}} = D \cdot Q_{\mathbf{o}}^{\mathrm{T}} \cdot X^{\mathrm{T}}, \tag{5.8}$$

where the first equality is by the expression of $[\![X]\!]$ in (3.12), and the last equality is because of the expressions in (5.5b). Also, $\widetilde{X}^{\mathrm{T}} D \equiv M$, independent of which $\widetilde{X} \in [\![X]\!]$. So (5.7) leads directly to $H(\widetilde{X}) = H_\phi(X) + G_\psi(X)$, which does not change as $\widetilde{X}$ varies in $[\![X]\!]$. This proves (5.6). □

The next theorem makes it precise in what sense the original NEPv (2.10) and its aligned NEPv (5.3) are equivalent.

Theorem 5.1. *Let $X \in \mathbb{O}^{n\times k}$ satisfy rank-preserving condition* (3.2)*, i.e., $X \in \mathbb{O}_D^{n\times k}$.*

(a) *If X is a D-regular solution to NEPv* (2.10)*, then X is a solution to the aligned NEPv* (5.3)*.*

(b) *If X is a solution to the aligned NEPv* (5.3)*, then any alignment $\widetilde{X} \in [\![X]\!]$ is a D-regular solution to NEPv* (2.10)*.*

Proof. Consider item (a). That X is D-regular implies $X \in \mathbb{O}_D^{n\times k}$ and $X \in [\![X]\!]$ and hence $G(X) = H([\![X]\!]) = H(X)$. Together with $H(X)X = X\Lambda$, we conclude $G(X)X = X\Lambda$, proving item (a).

Now turn to item (b). By the definition of G in (5.2), $G(X) = H([\![X]\!]) = H(\widetilde{X})$ for any $\widetilde{X} \in [\![X]\!]$. Recalling (3.12) that $\widetilde{X} = XQ$ for some $Q \in \mathbb{O}^{k\times k}$, we have

$$G(X)X = X\Lambda \qquad \Rightarrow \qquad H(\widetilde{X}) \cdot \widetilde{X} = \widetilde{X} \cdot \widetilde{\Lambda},$$

where $\widetilde{\Lambda} = Q^{\mathrm{T}} \Lambda Q$. Namely, $\widetilde{X} \in \mathbb{O}^{n\times k}$ is a solution to NEPv (2.10). Moreover, Corollary 3.1(b) implies $\widetilde{X}^{\mathrm{T}} D \succeq 0$, and so $\widetilde{X}$ is also D-regular. □

The aligned NEPv (5.3) is defined through (5.2) and NEPv (2.10). Alternatively, this new NEPv can be directly derived from the first-order optimality condition of a different maximization problem equivalent to (2.3); see our discussions in Appendix B. This alternative interpretation will become useful when it comes to justify the sufficient condition for the convergence of level-shifted SCF later in Section 6.

5.2 Differentiability for Aligned NEPv

In this subsection, we will show that $G(X)$ for the aligned NEPv (5.3) is differentiable in $X \in \mathbb{R}^{n\times k}$, at X in domain $\mathbb{O}_D^{n\times k}$ defined in (5.1), and establish closed-form expressions to various derivatives for the purpose of analyzing the local convergence of the plain SCF on the aligned NEPv. This will in turn allow us to derive a sharp estimate for the rate of convergence of our SCF-type iteration in Algorithm 4.1 for NEPv (2.10). By the expression of $G(X)$ in (5.6), it is sufficient to consider the differentiability of the canonical polar factors $Q_{\mathbf{o}}$ and M of $X^{\mathrm{T}} D$ with respect to $X \in \mathbb{R}^{n\times k}$.

We begin with an alternative formula for the canonical polar factors of $X^{\mathrm{T}} D$ for $X \in \mathbb{O}_D^{n\times k}$. Let $r_{_D} := \mathrm{rank}(D)$ and factorize D as

$$D = D_1 P^{\mathrm{T}}, \tag{5.9}$$

where $D_1 \in \mathbb{R}^{n\times r_{_D}}$ has full column rank and $P \in \mathbb{O}^{k\times r_{_D}}$. For $X \in \mathbb{O}_D^{n\times k}$, we have

$$\mathrm{rank}(X^{\mathrm{T}} D_1) = \mathrm{rank}(X^{\mathrm{T}} D) = \mathrm{rank}(D) \equiv r_{_D}, \tag{5.10}$$

and, hence, $X^{\mathrm{T}} D_1 \in \mathbb{R}^{k\times r_{_D}}$ has full column rank.

Lemma 5.3. *Let D be factorized as in (5.9), $X \in \mathbb{O}_D^{n\times k}$, and the polar decomposition of $X^{\mathrm{T}}D_1 \in \mathbb{R}^{k\times r_{\mathcal{D}}}$ be*

$$X^{\mathrm{T}}D_1 = Q_1 M_1. \tag{5.11}$$

(a) *The canonical polar decomposition of $X^{\mathrm{T}}D$ is given by $X^{\mathrm{T}}D = Q_{\mathbf{o}}M$, where*

$$Q_{\mathbf{o}} = Q_1 P^{\mathrm{T}} \quad \textit{and} \quad M = P M_1 P^{\mathrm{T}}. \tag{5.12}$$

(b) *Both $Q_{\mathbf{o}}$ and M are Fréchet differentiable with respect to $X \in \mathbb{R}^{n\times k}$, and their Fréchet derivatives along direction $E \in \mathbb{R}^{n\times k}$ are given by*

$$\mathbf{D}M(X)[E] = PLP^{\mathrm{T}}, \quad \mathbf{D}Q_{\mathbf{o}}(X)[E] = \left(E^{\mathrm{T}}DP - Q_{\mathbf{o}}PL\right) M_1^{-1}P^{\mathrm{T}}, \tag{5.13}$$

respectively, where L is the solution to the Lyapunov equation

$$M_1 L + L M_1 = D_1^{\mathrm{T}} \left(XE^{\mathrm{T}} + EX^{\mathrm{T}} \right) D_1. \tag{5.14}$$

(c) *It holds that*

$$\mathrm{tr}\left(\mathbf{D}M(X)[E] \right) = \mathrm{tr}\left(Q_{\mathbf{o}}D^{\mathrm{T}}E\right). \tag{5.15}$$

Proof. For item (a), observe that $Q_{\mathbf{o}}$ in (5.12) is a partial isometry (i.e., $\|Q_{\mathbf{o}}x\| = \|x\|$ for all $x \in \mathcal{R}(Q_{\mathbf{o}}^{\mathrm{T}})$) and $M \succeq 0$. By (5.9), (5.11), and (5.12), we get

$$Q_{\mathbf{o}}M = Q_1 M_1 P^{\mathrm{T}} = X^{\mathrm{T}}D_1 P^{\mathrm{T}} = X^{\mathrm{T}}D,$$

yielding the canonical polar decomposition of $X^{\mathrm{T}}D$, since the canonical polar decomposition is always unique (see, e.g., [7, p.220] and [22, Chapter 8]).

For item (b), observe that $X^{\mathrm{T}}D_1$ has full column rank. According to Lemma A.1 in the appendix, the polar factors M_1 and Q_1 from (5.11) are differentiable in X, and they have derivatives along direction $E \in \mathbb{R}^{n\times k}$ given by [4]

$$\mathbf{D}M_1(X)[E] = L \quad \text{and} \quad \mathbf{D}Q_1(X)[E] = (E^{\mathrm{T}}D_1 - Q_1 L) \cdot M_1^{-1}, \tag{5.16}$$

where L satisfies (5.14). Consequently, by (5.12), the canonical polar factors M and $Q_{\mathbf{o}}$ (depending on M_1 and Q_1) are differentiable in X as well, and their derivatives along direction $E \in \mathbb{R}^{n\times k}$ are

$$\mathbf{D}M(X)[E] = P \cdot \mathbf{D}M_1(X)[E] \cdot P^{\mathrm{T}} \quad \text{and} \quad \mathbf{D}Q_{\mathbf{o}}(X)[E] = \mathbf{D}Q_1(X)[E] \cdot P^{\mathrm{T}}.$$

Then by (5.16) we get (5.13).

For item (c), since $P^{\mathrm{T}}P = I_{r_{\mathcal{D}}}$, (5.13) implies $\mathrm{tr}\left(\mathbf{D}M(X)[E]\right) = \mathrm{tr}\left(L\right)$. Post-multiplying both sides of (5.14) by M_1^{-1} and then taking trace, we obtain

$$2\,\mathrm{tr}(L) = \mathrm{tr}\left(D_1^{\mathrm{T}}XE^{\mathrm{T}}D_1 M_1^{-1}\right) + \mathrm{tr}\left(D_1^{\mathrm{T}}EX^{\mathrm{T}}D_1 M_1^{-1}\right), \tag{5.17}$$

where we have used $\mathrm{tr}(M_1 L M_1^{-1}) = \mathrm{tr}(L)$. The two terms on the right hand side of (5.17) are identical because of $\mathrm{tr}(AM_1^{-1}) = \mathrm{tr}(M_1^{-1}A) = \mathrm{tr}(A^{\mathrm{T}}M_1^{-1})$ for all $A \in \mathbb{R}^{r_{\mathcal{D}}\times r_{\mathcal{D}}}$, where the last equality is by transposing the argument and $M_1^{\mathrm{T}} = M_1$. So (5.17) leads to

$$\mathrm{tr}(L) = \mathrm{tr}\left(D_1^{\mathrm{T}}EX^{\mathrm{T}}D_1 M_1^{-1}\right) = \mathrm{tr}\left(D_1^{\mathrm{T}}EQ_1\right) = \mathrm{tr}\left(Q_1 D_1^{\mathrm{T}}E\right),$$

which yields (5.15) upon noticing $Q_1 D_1^{\mathrm{T}} = (Q_1 P^{\mathrm{T}})(D_1 P^{\mathrm{T}})^{\mathrm{T}} = Q_{\mathbf{o}}D^{\mathrm{T}}$. □

[4] Notice the fact that, for $Z = X^{\mathrm{T}}D_1$ and $Y = E^{\mathrm{T}}D_1$, the Fréchet derivative of a differentiable function $F(Z)$ satisfies $\mathbf{D}F(Z)[Y] = \mathbf{D}\widetilde{F}(X)[E]$ with $\widetilde{F}(X) := F(X^{\mathrm{T}}D_1)$, as can be verified by straightforward expansions of the functions.

Suppose that $H(X)$ in (2.10b) is differentiable at $X \in \mathbb{R}^{n\times k}$ that has orthonormal columns (by Lemma 2.1, this always holds if ϕ and ψ in (2.3) are twice differentiable functions). Then the differentiability of the canonical polar factors, as in Lemma 5.3, implies that $G(X)$ in (5.6) is differentiable at any $X \in \mathbb{O}_D^{n\times k}$, i.e., satisfying rank-preserving condition (3.2). By a straightforward application of the chain rule of differentiation, we derive from (5.6) the following formula for assembling and calculating the derivative $\mathbf{D}G(X)[E]$.

Corollary 5.1. *Let $X \in \mathbb{O}_D^{n\times k}$ and assume ϕ and ψ in* (2.3) *are twice differentiable at X. Then,*

$$\begin{aligned}\mathbf{D}G(X)[E] = \mathbf{D}H_\phi(X)[E] + \operatorname{tr}(\mathbf{D}M(X)[E]) \cdot H_\psi(X) \\ + \operatorname{tr}(M) \cdot \mathbf{D}H_\psi(X)[E] + \mathbf{D}\psi(X)[E] \cdot \big(DQ_{\mathbf{o}}^{\mathrm{T}} X^{\mathrm{T}} + XQ_{\mathbf{o}} D^{\mathrm{T}}\big) \\ + \psi(X) \cdot \big(D \cdot \mathbf{D}Q_{\mathbf{o}}(X)[E]^{\mathrm{T}} \cdot X^{\mathrm{T}} + X \cdot \mathbf{D}Q_{\mathbf{o}}(X)[E] \cdot D^{\mathrm{T}}\big) \\ + \psi(X) \cdot \big(DQ_{\mathbf{o}}^{\mathrm{T}} E^{\mathrm{T}} + EQ_{\mathbf{o}} D^{\mathrm{T}}\big), \quad (5.18)\end{aligned}$$

where $\mathbf{D}M$ and $\mathbf{D}Q_{\mathbf{o}}$ are given by (5.13)*, and H_ϕ and H_ψ are from the expression of $H(X)$ in* (2.10).

5.3 Rate of Convergence

We now perform a local convergence analysis of Algorithm 4.1. Recall that the creation of NEPv (2.10) aims to solve optimization problem (2.3) whose global maximizers are provably D-regular eigenbasis matrices of the NEPv. Because of this, and as we are performing a local convergence analysis, we may assume without loss of generality that the initial guess X_0 is such that $\mathcal{R}(X_0)$ is sufficiently close to $\mathcal{R}(X_*)$, where $X_* \in \mathbb{O}^{n\times k}$ is a D-regular solution of NEPv (2.10), i.e., satisfying both conditions (3.1) and (3.2):

$$X_*^{\mathrm{T}} D \succeq 0 \quad \text{and} \quad \operatorname{rank}(X_*^{\mathrm{T}} D) = \operatorname{rank}(D) =: r_{_D}. \tag{5.19}$$

The key observation of our analysis is that, locally around $\mathcal{R}(X_*)$, we can identify the SCF-type iteration in Algorithm 4.1 as the plain SCF for the aligned NEPv (5.3): each iterative step of the SCF-type iteration in Algorithm 4.1 satisfy

$$G(X_i)X_{i+1} = X_{i+1}\Lambda_{i+1}, \qquad \text{for } i = 0, 1, 2, \ldots, \tag{5.20}$$

where the eigenvalues of $\Lambda_{i+1} \in \mathbb{R}^{k\times k}$ are the k largest eigenvalues of $G(X_i)$, provided $X_i \in \mathbb{O}_D^{n\times k}$, which ensures that $G(X_i)$ in (5.20) is well-defined by (5.2). The conditions $X_i \in \mathbb{O}_D^{n\times k}$ always hold if $\mathcal{R}(X_0)$ is sufficiently close to $\mathcal{R}(X_*)$ in the case of convergence; whereas in the case of divergence, it is still reasonable to assume that, at least for the first few SCF-type iterative steps of Algorithm 4.1 before $\mathcal{R}(X_i)$ deviates too far from $\mathcal{R}(X_*)$.

For the plain SCF (5.20), since $G(X)$ is unitarily invariant according to (5.4), we can apply the existing local convergence results for unitarily invariant NEPv in, e.g., [4, 10]. Our goal in the following is to establish the local convergence rate of Algorithm 4.1 through the plain SCF (5.20).

First, let $G(X_*)$ have the eigenvalue decomposition

$$G(X_*) = [X_*, X_{*\perp}] \begin{bmatrix} \Lambda_* & \\ & \Lambda_{*\perp} \end{bmatrix} [X_*, X_{*\perp}]^{\mathrm{T}}, \tag{5.21}$$

where $\Lambda_* = \operatorname{diag}(\lambda_1, \ldots, \lambda_k)$, $\Lambda_{*\perp} = \operatorname{diag}(\lambda_{k+1}, \ldots, \lambda_n)$, and $\lambda_1 \geq \cdots \geq \lambda_k \geq \lambda_{k+1} \geq \cdots \geq \lambda_n$. As in [4, 10], we assume

$$\lambda_k - \lambda_{k+1} > 0,$$

otherwise the eigenspace for the k largest eigenvalues of $G(X_*)$ is not unique [15, 29, 40]. Define

$$S(X_*) \in \mathbb{R}^{(n-k)\times k} \qquad \text{by} \qquad [S(X_*)]_{ij} = (\lambda_j - \lambda_{k+i})^{-1},$$

and linear operator $\mathscr{L}\colon \mathbb{R}^{(n-k)\times k} \to \mathbb{R}^{(n-k)\times k}$ by

$$\mathscr{L}(Z) := S(X_*) \odot \left(X_{*\perp}^{\mathrm{H}} \cdot \mathbf{D}G(X_*)[X_{*\perp} Z] \cdot X_*\right), \tag{5.22}$$

where $\odot$ is the element-wise multiplication, also known as the matrix Hadamard product, and $\mathbf{D}G$ is given by (5.18).

According to the convergence analysis in [4], the spectral radius $\rho(\mathscr{L})$ of linear operator $\mathscr{L}$ in (5.22) is the local convergence rate of the plain SCF (5.20) and, hence, that of Algorithm 4.1 as well. A restatement of [4, Theorem 4.2] for the plain SCF (5.20) yields the following theorem for the local *convergence-in-subspace* (i.e., the convergence of $\mathcal{R}(X_i)$ as $i \to \infty$) of the SCF-type iteration in Algorithm 4.1.

Theorem 5.2. *Let $X_* \in \mathbb{O}^{n\times k}$ be a D-regular solution to NEPv* (2.10) *such that the corresponding $\Lambda_* \equiv X_*^{\mathrm{T}} H(X_*) X_*$ contains the k largest eigenvalues of $H(X_*)$. Assume that*

$$\lambda_k(H(X_*)) > \lambda_{k+1}(H(X_*)) \tag{5.23}$$

and $H(X)$ is differentiable at X_*, and let $\rho(\mathscr{L})$ be the spectral radius of the linear operator $\mathscr{L}$ defined by (5.22).

(a) *If $\rho(\mathscr{L}) < 1$, then Algorithm 4.1 is locally convergent-in-subspace to X_*, with an asymptotic average convergence rate bounded by $\rho(\mathscr{L})$.*

(b) *If $\rho(\mathscr{L}) > 1$, then Algorithm 4.1 is locally divergent-in-subspace from X_*.*

Convergence-in-subspace, or divergence-in-subspace for that matter, is measured by the canonical angles between subspaces. In the case of Theorem 5.2, it is about

$$\text{whether} \quad \Theta(X_i, X_*) \to 0 \quad \text{or} \quad \Theta(X_i, X_*) \not\to 0 \qquad \text{as } i \to \infty,$$

where $\Theta(X_i, X_*)$ denotes the diagonal matrix of the canonical angles between the subspaces $\mathcal{R}(X_i)$ and $\mathcal{R}(X_*)$ [4, 10, 40]. So, Theorem 5.2(a) can be interpreted as: Given an initial X_0 with sufficiently small $\Theta(X_0, X_*)$, for an arbitrarily small $\epsilon > 0$,

$$\|\Theta(X_{i+m}, X_*)\| \ \le c\,[\rho(\mathscr{L}) + \epsilon]^m \cdot \|\Theta(X_i, X_*)\| \qquad \text{as } i, m \to \infty, \tag{5.24}$$

where c is some constant and $\|\cdot\|$ is any unitarily invariant matrix norm, such as the 2-norm and the Frobenius norm. The reader is referred to [4, Section 4] for more discussions on the local convergence rate of SCF.

We emphasize that the convergence result in Theorem 5.2 holds even if $D \in \mathbb{R}^{n\times k}$ is rank-deficient (i.e., $\mathrm{rank}(D) < k$). In the rank-deficient case, according to (3.12), the alignment function $[\![\widetilde{X}_{i+1}]\!]$ at line 3 Algorithm 4.1 is multi-valued, and so approximation X_{i+1} is not uniquely defined. However, provided X_* is a D-regular eigenbasis matrix satisfying the gap condition (5.23), the sequence of eigenspaces $\{\mathcal{R}(X_i)\}$ by Algorithm 4.1 is unique as each $\mathcal{R}(X_i)$ is sufficiently close to $\mathcal{R}(X_*)$, and the rate of local convergence-in-subspace is given by $\rho(\mathscr{L})$. Later in Section 7, we will numerically demonstrate the results of Theorem 5.2.

Algorithm 6.1 A Level-Shifted SCF-type iteration for NEPv (2.10)

Input: $X_0 \in \mathbb{O}^{n\times k}$ such that $X_0^{\mathrm{T}} D \succeq 0$, and a level-shift σ;
Output: a solution to NEPv (2.10) for the purpose of solving optimization problem (2.3).

1: **for** $i = 0, 1, \ldots$ until convergence **do**
2: solve the symmetric eigenproblem $[H(X_i) + \sigma X_i X_i^{\mathrm{T}}]\widetilde{X}_{i+1} = \widetilde{X}_{i+1}\widetilde{\Lambda}_{i+1}$, where $\widetilde{X}_{i+1} \in \mathbb{O}^{n\times k}$ contains the eigenvectors for the k largest eigenvalues of $H(X_i) + \sigma X_i X_i^{\mathrm{T}}$;
3: align $\widetilde{X}_{i+1}$ to get $X_{i+1} \in [\![\widetilde{X}_{i+1}]\!]$;
4: **end for**
5: **return** the last X_i as a solution to NEPv (2.10).

6 Level-Shifted SCF

For NEPv from the special optimization problems (2.3) studied in [48, 55, 56], Algorithm 4.1 is provably convergent globally from any initial guess. However, for NEPv arising from a general optimization problem (2.3), the algorithm may diverge (even with very accurate initial guesses), as we shall demonstrate by numerical examples in Section 7.

When Algorithm 4.1 diverges, we may apply a level-shifting scheme, which has been commonly adopted to unitarily invariant NEPv for fixing the issue of eigenvalue mispositioning, by which we mean not all eigenvalues of Λ at optimality are among the largest ones and, as a result, Algorithm 4.1 is inevitably divergent; see, e.g., [5, 38, 44, 50]. The level-shifting technique simply modifies the matrix-valued function $H(X)$ to $H(X)+\sigma XX^{\mathrm{T}}$, where $\sigma \in \mathbb{R}$ is a preselected level-shift. Note that the addendum σXX^{T} is unitarily invariant. In an obvious way, this scheme can be conveniently adapted to Algorithm 4.1, by changing $H(X_i)$ at line 2 to $H(X_i) + \sigma X_i X_i^{\mathrm{T}}$. For ease of reference, we outline the level-shifted variant of Algorithm 4.1 in Algorithm 6.1, where again we may stop the iteration if $\mathrm{NRes}(X_i) \le \mathtt{tol}$, as commented earlier for Algorithm 4.1.

Let X_* be a D-regular solution of NEPv (2.10). Similarly to Algorithm 4.1, we can study the local convergence of the level-shifted SCF in Algorithm 6.1 to X_* through the aligned NEPv (5.3). First, let us level-shift the aligned NEPv (5.3) to

$$G_\sigma(X)X = X\Lambda_\sigma \qquad \text{with} \qquad G_\sigma(X) := G(X) + \sigma XX^{\mathrm{T}}. \tag{6.1}$$

The two NEPv are equivalent in that: (X_*, Λ_*) is a solution to NEPv (5.3) $G(X)X = X\Lambda$, if and only if $(X_*, \Lambda_{\sigma*})$ with $\Lambda_{\sigma*} = \Lambda_* + \sigma I$ is a solution to the level-shifted NEPv (6.1). Then, by a straightforward verification, the sequence of $\{X_i\}$ from Algorithm 6.1 satisfy

$$G_\sigma(X_i)X_{i+1} = X_{i+1}\Lambda_{\sigma,i+1} \quad \text{for } i = 0, 1, \ldots, \tag{6.2}$$

where the eigenvalues of $\Lambda_{\sigma,i+1}$ are the k largest eigenvalues of $G_\sigma(X_i)$. We have again assumed $X_i \in \mathbb{O}_D^{n\times k}$, which holds locally if $\mathcal{R}(X_i)$ is close to $\mathcal{R}(X_*)$, as in (5.20). We can see that (6.2) is exactly the plain SCF for the level-shifted NEPv (6.1). By this interpretation, we explain in the following two major benefits for applying the level-shifted SCF in Algorithm 6.1.

Eigenvalue repositioning. Previously, for the purpose of solving optimization problem (2.3), we assumed the desired solutions of the aligned NEPv (5.3) are those with Λ corresponding to the k largest eigenvalues of $G(X)$. This assumption stems more from past research experiences on special cases of optimization problem (2.3) [48, 56, 55] than from rigorous mathematical proofs. For optimization problem (2.3) in its generality, there is indeed no guarantee that Λ has such a property at optimality; see, e.g., [5] for an example of optimization problem (2.3) with $D = 0$, where the target eigenvalue is not an extreme eigenvalue.

Having that said, we also notice that the assumption above is actually not essential for the application of SCF iteration, because otherwise we can level-shift NEPv (5.3) to NEPv (6.1). By the eigendecomposition (5.21), we have

$$G_\sigma(X_*) = [X_*, X_{*\perp}] \begin{bmatrix} \Lambda_* + \sigma I & \\ & \Lambda_{*\perp} \end{bmatrix} [X_*, X_{*\perp}]^{\mathrm{T}}. \tag{6.3}$$

Since Λ_* and $\Lambda_{*\perp}$ are fixed, each of the k eigenvalues in $\Lambda_* + \sigma I$ can be made larger than those in $\Lambda_{*\perp}$ by choosing σ sufficiently large. Particularly, if

$$\sigma > \lambda_{\max}(G(X_*)) - \lambda_{\min}(G(X_*)), \tag{6.4}$$

then the eigenvalues of $\Lambda_{\sigma *} = \Lambda_* + \sigma I$ always consist of the k largest eigenvalues of $G_\sigma(X_*)$. In this case, we can apply the plain SCF (6.2) to the level-shifted NEPv (6.1), using the top k eigenvalues. A similar idea of using level-shifting to reposition the desired eigenvalues of NEPv has also been explored in [5].

Convergence of level-shifted SCF. As another important property, level-shifting can also fix the potential divergence issue of the plain SCF. It is well-known that, under mild assumptions, the level-shifted SCF is always locally convergent when σ is sufficiently large for unitarily invariant NEPv; see, e.g., [4, 11]. For Algorithm 6.1, due to its equivalence to the plain level-shifted SCF in (6.2) for the aligned NEPv, we expect such convergence property to still hold.

We first consider the local convergence rate of Algorithm 6.1. Similarly to $\mathscr{L}$ in (5.22), we define linear operator $\mathscr{L}_\sigma \colon \mathbb{R}^{(n-k)\times k} \to \mathbb{R}^{(n-k)\times k}$ for the level-shifted NEPv (6.1) at a D-regular solution X_* as

$$\mathscr{L}_\sigma(Z) := S_\sigma(X_*) \odot \left(X_{*\perp}^{\mathrm{H}} \ \cdot \mathbf{D}G_\sigma(X_*)[X_{*\perp} Z] \cdot X_* \right), \tag{6.5}$$

where $S_\sigma(X_*) \in \mathbb{R}^{(n-k)\times k}$ is with entries $[S_\sigma(X_*)]_{ij} = (\lambda_j - \lambda_{k+i} + \sigma)^{-1}$, recalling the eigendecomposition (5.21). It follows from (6.1) that $\mathbf{D}G_\sigma(X_*)[Y] \equiv \mathbf{D}G(X_*)[Y] + \sigma(X_* Y^{\mathrm{T}} + Y X_*^{\mathrm{T}})$, and hence

$$\mathscr{L}_\sigma(Z) = S_\sigma(X_*) \odot \mathcal{Q}(Z) + Z, \tag{6.6}$$

where $\mathcal{Q} \colon \mathbb{R}^{(n-k)\times k} \to \mathbb{R}^{(n-k)\times k}$ is a linear operator given by

$$\mathcal{Q}(Z) = \Lambda_{*\perp} Z - Z\Lambda_* + X_{\perp *}^{\mathrm{T}} \mathbf{D}G(X_*)[X_{*\perp} Z] X_*, \tag{6.7}$$

similarly to [4, Theorem 5.1]. By the convergence analysis in [4], as also discussed in Section 5, the spectral radius $\rho(\mathscr{L}_\sigma)$ gives the local convergence rate of the plain level-shifted SCF (6.2) and, thus, that of Algorithm 6.1 as well. To guarantee local convergence of Algorithm 6.1 to X_*, we hence need $\rho(\mathscr{L}_\sigma) < 1$.

Let us establish a sufficient condition on σ for $\rho(\mathscr{L}_\sigma) < 1$ to hold. We observe that the scaling matrix $S_\sigma(X_*)$ in (6.6) has positive entries that go to 0 as $\sigma \to +\infty$. Therefore, assuming $\mathcal{Q}(Z)$ is negative definite in the sense that

$$\mathrm{tr}(\, Z^{\mathrm{T}} \mathcal{Q}(Z)\,) < 0 \qquad \text{for all nonzero } Z \in \mathbb{R}^{(n-k)\times k}, \tag{6.8}$$

we will have $\mathscr{L}_\sigma(Z) < 1$ for sufficiently large σ, noticing that $\mathscr{L}_\sigma(\,\cdot\,)$ approaches the identity operator as $\sigma \to \infty$. Precisely, under condition (6.8), it will hold

$$\mathscr{L}_\sigma(Z) < 1 \quad \text{for all} \quad \sigma > \sigma_L := -\frac{\mu_{\min}}{2} - [\lambda_k(G(X_*)) - \lambda_{k+1}(G(X_*))], \tag{6.9}$$

where $\mu_{\min} < 0$ is the smallest eigenvalue of linear operator $\mathcal{Q}$ in (6.7). The result (6.9) is essentially due to [4, Theorem 5.1], which, although established for NEPv $H(X)X = X\Lambda$ with Λ being the smallest eigenvalues of $H(X)$, can be quickly adapted to (5.3) by working with $-G(X)X = X\Lambda$.

Finally, the following theorem justifies (6.8) is indeed a mild assumption for aligned NEPv (5.3), since the operator $\mathcal{Q}$ in (6.7) is at least negative semi-definite at a global maximizer X_* of (2.3). Its proof is left to Appendix C.

Theorem 6.1. *Assume ϕ and ψ in* (2.3) *are twice differentiable. Let $X_* \in \mathbb{O}^{n\times k}$ be a global maximizer of* (2.3) *and $\mathcal{Q}$ be defined in* (6.7)*. Then it holds that*

$$\mathrm{tr}(\, Z^{\mathrm{T}}\mathcal{Q}(Z)\,) \le 0 \qquad \textit{for all nonzero } Z \in \mathbb{R}^{(n-k)\times k}. \tag{6.10}$$

7 Numerical Experiments

In this section, we verify our theoretical results by numerical experiments. We will compare the observed convergence rate with our theoretical estimate, i.e., the spectral radius for the corresponding $\mathscr{L}$ operator in (5.22). The spectral radius $\rho(\mathscr{L})$ is computed by MATLAB's built-in function `eigs`, and the 'exact' solution X_* is computed by Algorithm 4.1 or, when it fails to converge, Algorithm 6.1 with a suitable level-shift σ. Our stopping criterion is

$$\mathrm{NRes}(X_i) \le \texttt{tol} = 10^{-13},$$

where $\mathrm{NRes}(\,\cdot\,)$ is the normalized residual (4.3) evaluated at the most recent approximation using the matrix 1-norm (i.e., the maximum column sum in absolute value) for computational convenience. We refer to [4] for more details about the computation of $\rho(\mathscr{L})$. All experiments are carried out in MATLAB and run on a Dell desktop with an Intel i9-9900K CPU and 16G memory. The MATLAB scripts implementing the algorithms and the data used to generate the numerical results in this paper can be accessed at `https://github.com/ddinglu/uinepv`.

Our testing NEPv arise from two optimization problems in the form of (2.3), both of which are defined by three matrices A, $B \in \mathbb{R}^{n\times n}$ and $D \in \mathbb{R}^{n\times k}$ along with an additional parameter, where A, B are symmetric and $B \succ 0$, and $k < n$. The first optimization problem is

$$\max_{X\in\mathbb{O}^{n\times k}} f_\alpha(X) \quad \text{with} \quad f_\alpha(X) := (1-\alpha)\cdot\frac{\mathrm{tr}(X^{\mathrm{T}}AX)}{\mathrm{tr}(X^{\mathrm{T}}BX)} + \alpha\cdot\frac{\mathrm{tr}(X^{\mathrm{T}}D)}{\sqrt{\mathrm{tr}(X^{\mathrm{T}}BX)}}, \tag{7.1}$$

where $\alpha \in [0,1]$ is a parameter, and the second one is (1.4) mentioned earlier at the beginning of this paper:

$$\max_{X\in\mathbb{O}^{n\times k}} f_\theta(X) \quad \text{with} \quad f_\theta(X) := \frac{\mathrm{tr}(X^{\mathrm{T}}AX + X^{\mathrm{T}}D)}{[\mathrm{tr}(X^{\mathrm{T}}BX)]^\theta}, \tag{7.2}$$

where θ is a parameter. Problem (7.2) has been studied in [48], where Algorithm 4.1 is proved to be linearly convergent from any initial guess, when $\theta \in \{0,1\}$, or $0 < \theta < 1$ and initially $\mathrm{tr}(X_0^{\mathrm{T}}AX_0 + X_0^{\mathrm{T}}D) \ge 0$, but its theoretical rate of convergence was not investigated. We will fill this gap and also demonstrate what could happen when θ is outside of the interval $[0,1]$.

With $D = 0$ and $\theta = 1$, both problems degenerate to the same one — the trace-ratio optimization problem from the linear discriminant analysis (LDA). For the case, the NEPv approach has been well studied in [53, 54]. The corresponding NEPv (7.4) is unitarily invariant, and Algorithm 4.1, which coincides with the plain SCF (i.e., line 3 replaced by $X_{i+1} = \widetilde{X}_{i+1}$), is globally and locally quadratically convergent to the global optimal solution generically. However, for the general cases with $\alpha D \neq 0$ in (7.1) or $\theta \neq 1$ in (7.2), the two optimization problems are different, and Algorithm 4.1 generally loses quadratic convergence or even possibly diverges (if without the help from level-shifting) for the associated NEPv, as will be shown in a moment.

7.1 NEPv from (7.1)

Optimization problem (7.1) is in the form of (2.3) with unitarily invariant functions

$$\phi(X) = (1-\alpha)\cdot\frac{\mathrm{tr}(X^{\mathrm{T}}AX)}{\mathrm{tr}(X^{\mathrm{T}}BX)} \qquad \text{and} \qquad \psi(X) = \frac{\alpha}{\sqrt{\mathrm{tr}(X^{\mathrm{T}}BX)}}. \tag{7.3}$$

Its KKT condition, by Theorem 2.1, is equivalent to NEPv

$$H_\alpha(X)X = X\Lambda, \tag{7.4a}$$

where the subscript α indicates its dependence on parameter α, and

$$H_\alpha(X) = \frac{2}{\mathrm{tr}(X^{\mathrm{T}}BX)}\left((1-\alpha)A - \phi(X)\cdot B\right) - \frac{\mathrm{tr}(X^{\mathrm{T}}D)\cdot\psi(X)}{\mathrm{tr}(X^{\mathrm{T}}BX)}\cdot B + \psi(X)\cdot(DX^{\mathrm{T}}+XD^{\mathrm{T}}), \tag{7.4b}$$

which is derived, using

$$\begin{cases} \dfrac{\partial\phi(X)}{\partial X} = H_\phi(X)X \qquad \text{with} \qquad H_\phi(X) = \dfrac{2}{\mathrm{tr}(X^{\mathrm{T}}BX)}\Big[(1-\alpha)\cdot A - \phi(X)\cdot B\Big], \\ \dfrac{\partial\psi(X)}{\partial X} = H_\psi(X)X \qquad \text{with} \qquad H_\psi(X) = \dfrac{-\psi(X)}{\mathrm{tr}(X^{\mathrm{T}}BX)}\cdot B. \end{cases} \tag{7.5}$$

By varying $\alpha \in [0,1]$, we can construct a variety of NEPv (7.4) for testing. For $\alpha = 0$, the initial guess X_0 of SCF is set to the eigenvectors of the k largest eigenvalues of the linear problem $Ax = \lambda Bx$. As α increases from 0 to 1, we use the computed solution from the previous NEPv as a starting guess for the next one.

For the purpose of calculating the local rate of convergence, we also obtain by straightforward derivation

$$\begin{cases} \mathbf{D}H_\phi(X)[E] = -2\dfrac{\mathrm{tr}(X^{\mathrm{T}}BE)}{\mathrm{tr}(X^{\mathrm{T}}BX)}\cdot H_\phi(X) - 2\dfrac{\mathrm{tr}(X^{\mathrm{T}}H_\phi(X)E)}{\mathrm{tr}(X^{\mathrm{T}}BX)}\cdot B, \\ \mathbf{D}H_\psi(X)[E] = -3\dfrac{\mathrm{tr}(X^{\mathrm{T}}H_\psi(X)E)}{\mathrm{tr}(X^{\mathrm{T}}BX)}B. \end{cases} \tag{7.6}$$

We construct the aligned NEPv (5.3) through alignment and the derivative operator $\mathbf{D}G$ through (5.18), using the derivatives in (7.6) together with $\mathbf{D}Q_{\mathbf{o}}$ and $\mathbf{D}M$ from (5.13). The corresponding linear operator $\mathscr{L}$ by (5.22) is then obtained.

Example 7.1. We consider problem (7.1) with $D \in \mathbb{R}^n$ being a single vector, i.e., $k = 1$. The alignment operation at line 3 of Algorithm 4.1 is simply $\widetilde{X}_i = \pm X_i$ with $\pm$ corresponding to the sign of $X_i^{\mathrm{T}}D \in \mathbb{R}$. Letting also $n = 3$, we randomly generate

$$A = \begin{bmatrix} -3.242 & -0.450 & 1.807 \\ -0.450 & -1.630 & 0.790 \\ 1.807 & 0.790 & 0.226 \end{bmatrix},\ B = \begin{bmatrix} 0.592 & 1.873 & 0.175 \\ 1.873 & 6.332 & 0.617 \\ 0.175 & 0.617 & 0.488 \end{bmatrix},\ D = \begin{bmatrix} -9.122 \\ 0.421 \\ 3.134 \end{bmatrix}.$$

In what follows, we examine the convergence of Algorithm 4.1 on NEPv (7.4) as α varies in $[0,1]$.

Figure 7.1 compares the observed rates of convergence of Algorithm 4.1 with the theoretical rates $\rho(\mathscr{L})$, as well as the iterative histories of normalized NEPv residuals, at several values of α. The left plot shows linear convergence of Algorithm 4.1 at a few sampled $\alpha \neq 0$ when Algorithm 4.1 actually converges and one α when Algorithm 4.1 does not. The right plot is the curve of spectral radius $\rho(\mathscr{L})$ as α varies in $[0,1]$. We can see that:

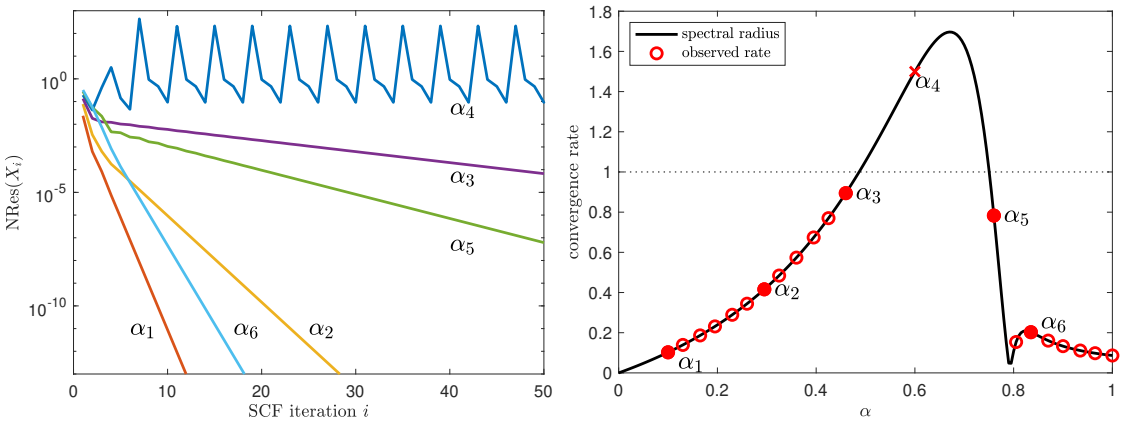

Figure 7.1: Algorithm 4.1 on Example 7.1. *Left:* The iterative history for solving NEPv (7.4) at a few sampled α (correspondingly marked as $\bullet$ and $\times$ on the right plot). *Right:* The curve of spectral radius $\rho(\mathscr{L})$ as a function of parameter $\alpha \in [0, 1]$ (based on 200 equally spaced α), and the observed rates of convergence (marked by $\bullet$ and $\circ$) at a number of values of α, including those sampled α on the left plot, and '$\times$' indicates that SCF is divergent.

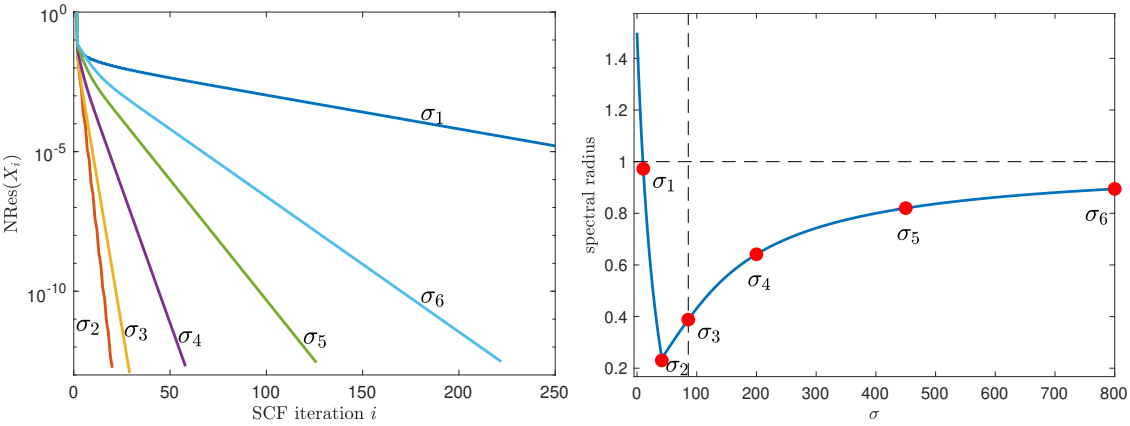

Figure 7.2: Algorithm 6.1 on Example 7.1 with $\alpha = 0.6$ (i.e., α_4 in Figure 7.1, at which Algorithm 4.1 diverges). *Left:* The iterative history of Algorithm 6.1 (level-shifted SCF) with a few sampled level-shifts σ (correspondingly marked as $\bullet$ on the right plot). *Right:* The curve of spectral radius $\rho(\mathscr{L}_\sigma)$ as a function of the level-shift σ and the observed rates of convergence at the sampled level-shifts σ. The vertical dashed line corresponds to the theoretical lower bound $\sigma_L = 85.83$ given by (6.9) (as σ_3 in the plots).

- At those α where Algorithm 4.1 converges, the observed rates of convergence and the theoretical ones by $\rho(\mathscr{L})$ match very well. For example,

$$\text{at } \alpha_3 = 0.46: \qquad \text{observed rate} \approx 0.894493\cdots, \quad \rho(\mathscr{L}) \approx 0.894490\cdots,$$

  matching up to 5 significant decimal digits. We conclude the spectral radius $\rho(\mathscr{L})$ provides sharp estimation for the true convergence rates.

- $\rho(\mathscr{L}) > 1$ approximately for $\alpha \in [0.49, 0.75]$, where Algorithm 4.1 indeed diverges numerically; see, e.g., the convergence behavior for $\alpha_4 = 0.6$ on the left plot. Those $\rho(\mathscr{L})$ are calculated

with solutions X_* to NEPv (7.4) computed by the level-shifted SCF, Algorithm 6.1, with $\sigma = 100$. It is interesting to notice that Algorithm 4.1 converges for α near 0 and 1 but diverges in the middle. This can be explained. In fact, optimization problem (7.1) for $\alpha = 0$ is the same as the one from LDA [53, 54], while for $\alpha = 1$ it is the OCCA subproblem [55]. For both cases, Algorithm 4.1 provably converges.

- $\rho(\mathscr{L}) \to 0$ as $\alpha \to 0$, indicating superlinear convergence of the algorithm at $\alpha = 0$, consistent with the fact that Algorithm 4.1 is quadratically convergent in such a case [54]. At the other end $\alpha = 1$, the spectral radius is a small number $\rho(\mathscr{L}) \approx 0.086$, indicating rapid linear convergence of Algorithm 4.1.

To demonstrate the effectiveness of level-shifting, Figure 7.2 shows the convergence of level-shifted SCF (Algorithm 6.1) on NEPv (7.4) for $\alpha = 0.6$ (i.e., α_4 in Figure 7.1), at which Algorithm 4.1 diverges. The left plot of Figure 7.2 illustrates the linear convergence of Algorithm 6.1 at various level-shifts σ, where a proper choice of σ can lead to a significant acceleration of convergence for SCF. The right plot shows the spectral radius of $\mathscr{L}_\sigma$ by (6.5) as level-shift σ varies, where the optimal level-shift $\sigma_* \approx 41.88$ with the minimal value $\rho(\mathscr{L}_\sigma) \approx 0.239$. It is observed that $\rho(\mathscr{L}_\sigma) < 1$ for all level-shifts $\sigma \gtrsim 9.87$, but the theoretical lower bound in (6.9) can only guarantee that $\rho(\mathscr{L}_\sigma) < 1$ and that Algorithm 6.1 is locally convergent when level-shift $\sigma \gtrsim \sigma_L = 85.83$, greatly overestimating the observed one.

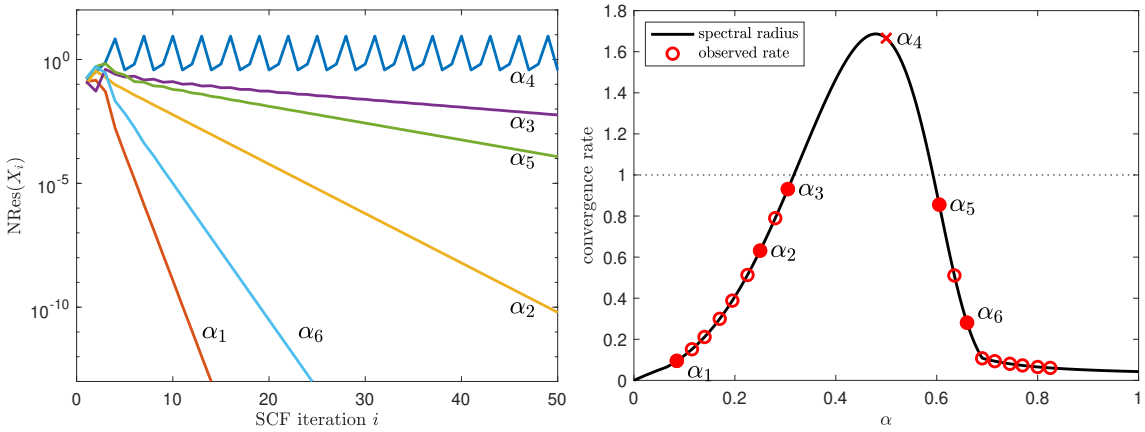

Figure 7.3: Algorithm 4.1 on Example 7.2. *Left:* The iterative history for solving NEPv (7.4) at a few sampled α (correspondingly marked as $\bullet$ and $\times$ on the right plot). *Right:* the curve of spectral radius $\rho(\mathscr{L})$ as a function of parameter $\alpha \in [0, 1]$ (based on 200 equally spaced α), and the observed rates of convergence (marked by $\bullet$ and $\circ$) at a number of values of α, including those sampled α on the left plot, and '$\times$' indicates that SCF is divergent.

Example 7.2. This is an almost repeat of Example 7.1, with the same testing matrices A and B, except $k = 2$ and

$$D = \begin{bmatrix} -1.430 & 2.768 \\ -0.120 & -0.630 \\ 1.098 & 2.229 \end{bmatrix} \in \mathbb{R}^{3\times 2}.$$

Since $k = 2$, the alignment at line 3 of Algorithm 4.1 does require the SVD (or the polar decomposition) of $X_i^{\mathrm{T}} D$. Figure 7.3 shows testing results by Algorithm 4.1 for several values of $\alpha \in [0, 1]$. We

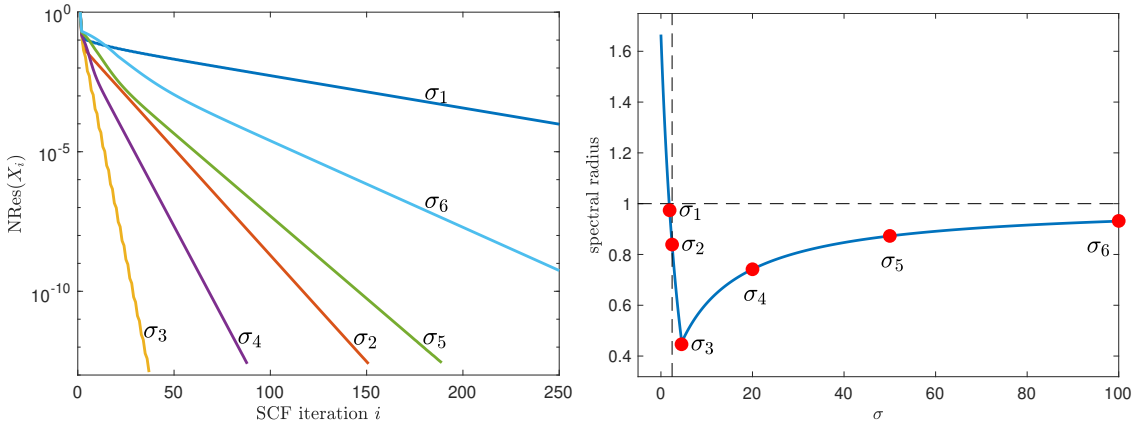

Figure 7.4: Algorithm 6.1 on Example 7.2 with $\alpha = 0.5$ (i.e., α_4 in Figure 7.3, at which Algorithm 4.1 diverges). *Left:* The iterative history of Algorithm 6.1 (level-shifted SCF) with a few sampled level-shifts σ (correspondingly marked as $\bullet$ on the right plot). *Right:* the curve of spectral radius $\rho(\mathscr{L}_\sigma)$ as a function of the level-shift σ and the observed rates of convergence at the sampled level-shifts σ. The vertical dashed line corresponds to the theoretical lower bound $\sigma_L = 2.44$ given by (6.9) (as σ_2 in the plots).

observe a similar performance of the algorithm as in Figure 7.1. We again find that $\rho(\mathscr{L})$ provides a sharp estimation for the convergence rates. For example,

$$\text{at } \alpha_3 = 0.305: \qquad \text{observed rate} \approx 0.930833\cdots, \quad \rho(\mathscr{L}) \approx 0.930798\cdots.$$

From the spectral radius curve as a function of α, it is concluded that Algorithm 4.1 converges rapidly for α at both ends of the interval $[0, 1]$, but fails to converge for α in the middle (approximately from 0.32 to 0.59), where the spectral radius $\rho(\mathscr{L}) > 1$ (those $\rho(\mathscr{L})$ are calculated using the computed solution X_* by Algorithm 6.1 with $\sigma = 50$). Next, we test the level-shifted SCF, Algorithm 6.1, on the NEPv with $\alpha = 0.5$ (i.e., α_4 in Figure 7.3), at which Algorithm 4.1 diverges. As reported in Figure 7.4, the optimal level-shift occurs at $\sigma_* \approx 4.57$ with the spectral radius $\rho(\mathscr{L}_\sigma) \approx 0.455$. For this NEPv, it is guaranteed that $\rho(\mathscr{L}_\sigma) < 1$ with any level-shift $\sigma \gtrsim \sigma_L = 2.44$ according to the theoretical lower bound (6.9) (as σ_2 in both left and right plots).

Example 7.3. We now consider larger n and more general $D \in \mathbb{R}^{n\times k}$ that is possibly rank-deficient unlike the previous examples. Set $A = \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{n\times n}$, a symmetric tridiagonal matrix, and $B = \text{diag}(1, 2, \ldots, n)$, and $n = 200$.

In the first experiment, we consider full-rank $D \in \mathbb{R}^{n\times k}$, generated by the MATLAB function `randn`(n, k). We test $k = 10, 20, 30, 40$, and for each k we repeat the experiment as in the previous examples, i.e., varying parameter α from 0 to 1. The results are summarized in Figure 7.5, where Algorithm 4.1 is locally convergent for all testing cases. In the left plot, the spectral radius moves from 0 to close to 1 as α increases, showing that SCF quickly loses superlinear convergence with a tiny increase of α from 0, and for $\alpha \approx 1$, Algorithm 4.1 converges linearly. Overall, the convergence appears slower than that in Figures 7.1 and 7.2 for the previous examples, and the convergence rate seems not so sensitive to varying α. The right plot again shows that the spectral radius provides sharp estimation for the convergence rate, justifying our theoretical analysis in Section 5.

In the second experiment, we consider rank-deficient $D \in \mathbb{R}^{n\times k}$ with $k = 50$ fixed, constructed as $D = D_1 P^{\mathrm{T}}$ with randomly generated $D_1 \in \mathbb{R}^{n\times r_D}$ and $P \in \mathbb{R}^{k\times r_D}$ that has orthonormal columns.

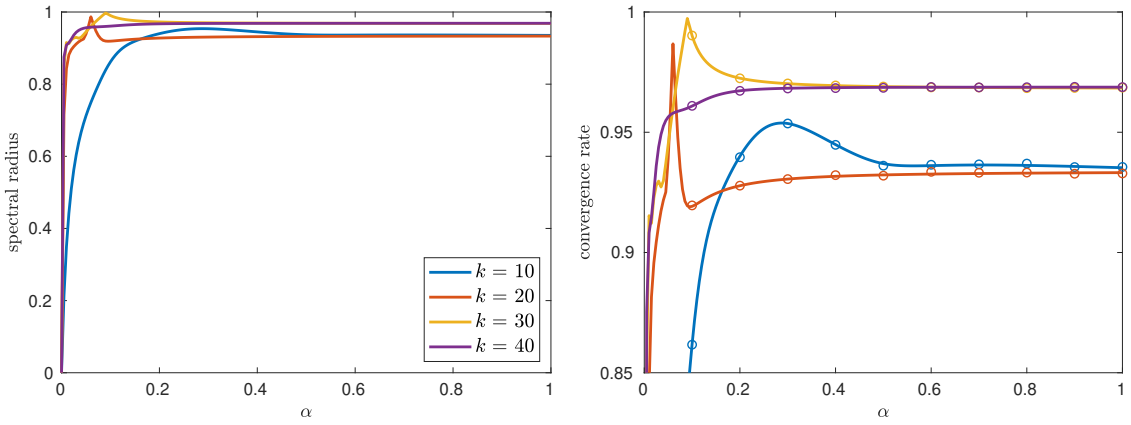

Figure 7.5: Algorithm 4.1 on Example 7.3 with full-rank $D \in \mathbb{R}^{n\times k}$. *Left:* The curves of spectral radius $\rho(\mathscr{L})$ as a function of $\alpha \in [0, 1]$ for different k (based on 200 equally spaced α). *Right:* The observed rates of convergence of Algorithm 4.1 at a few sampled α (marked by $\circ$), zoomed in for rates in $[0.85, 1]$ for readability.

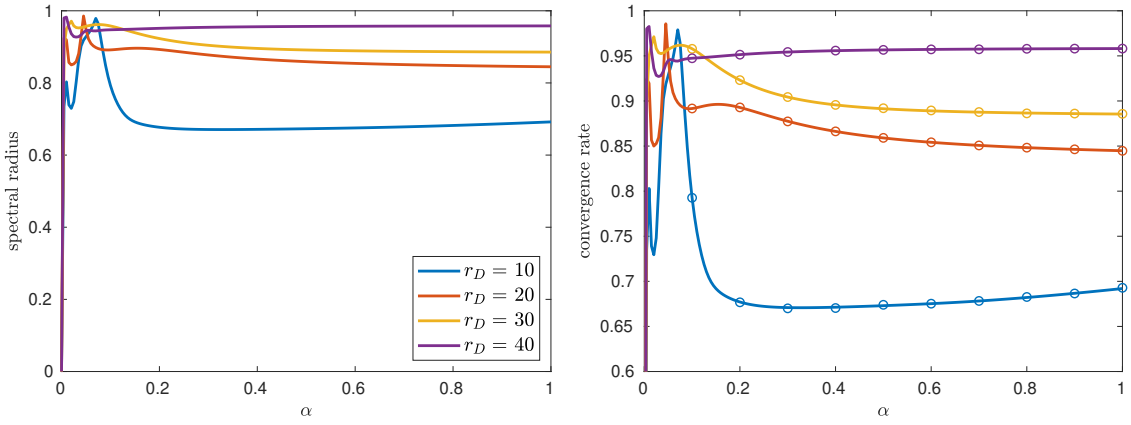

Figure 7.6: Algorithm 4.1 on Example 7.3 with rank-deficient $D \in \mathbb{R}^{n\times k}$ with $k = 50$ and different $r_{_D} \equiv \mathrm{rank}(D)$. *Left:* The curves of spectral radius $\rho(\mathscr{L})$ as a function of $\alpha \in [0, 1]$ (based on 200 equally spaced α). *Right:* The observed rates of convergence of Algorithm 4.1 at sampled α (marked by $\circ$), zoomed in for rates in $[0.6, 1]$ for readability.

We test $r_{_D} = 10, 20, 30, 40$ and repeat the experiment. From the results in Figure 7.6, the spectral radius still provides sharp estimation for the convergence rate, justifying the analysis in Section 5 for rank-deficient D.

7.2 NEPv from (7.2)

Optimization problem (7.2) is another example of (2.3), and it has the unitarily invariant functions

$$\phi(X) = \frac{\mathrm{tr}(X^{\mathrm{T}}AX)}{[\mathrm{tr}(X^{\mathrm{T}}BX)]^{\theta}} \qquad \text{and} \qquad \psi(X) = \frac{1}{[\mathrm{tr}(X^{\mathrm{T}}BX)]^{\theta}}. \tag{7.7}$$

By Theorem 2.1, its KKT condition is equivalent to NEPv

$$H_{\theta}(X)X = X\Lambda, \tag{7.8a}$$

where the subscript θ indicates its dependence on parameter θ, and

$$H_\theta(X) = 2\psi(X)\cdot\Big[A-\theta\cdot\frac{\mathrm{tr}(X^{\mathrm{T}}AX)}{\mathrm{tr}(X^{\mathrm{T}}BX)}\cdot B\Big] - 2\theta\frac{\mathrm{tr}(X^{\mathrm{T}}D)\cdot\psi(X)}{\mathrm{tr}(X^{\mathrm{T}}BX)}\cdot B + \psi(X)\cdot(DX^{\mathrm{T}}+XD^{\mathrm{T}}), \quad (7.8b)$$

which is obtained, using

$$\begin{cases} \dfrac{\partial\phi(X)}{\partial X} = H_\phi(X)X \qquad \text{with} \qquad H_\phi(X) = 2\psi(X)\cdot\Big[A-\theta\cdot\dfrac{\mathrm{tr}(X^{\mathrm{T}}AX)}{\mathrm{tr}(X^{\mathrm{T}}BX)}\cdot B\Big], \\ \dfrac{\partial\psi(X)}{\partial X} = H_\psi(X)X \qquad \text{with} \qquad H_\psi(X) = -2\theta\cdot\dfrac{\psi(X)}{\mathrm{tr}(X^{\mathrm{T}}BX)}\cdot B. \end{cases} \quad (7.9)$$

By varying θ, we construct a variety of NEPv (7.8) for testing.

For the purpose of calculating the local rate of convergence, we also obtain

$$\begin{cases} \mathbf{D}H_\phi(X)[E] = -2\theta\dfrac{\mathrm{tr}(X^{\mathrm{T}}BE)}{\mathrm{tr}(X^{\mathrm{T}}BX)}\cdot H_{\phi,1}(X) - 2\theta\dfrac{\mathrm{tr}(X^{\mathrm{T}}H_\phi(X)E)}{\mathrm{tr}(X^{\mathrm{T}}BX)}\cdot B, \\ \mathbf{D}H_\psi(X)[E] = -2(\theta+1)\dfrac{\mathrm{tr}(X^{\mathrm{T}}H_\psi(X)E)}{\mathrm{tr}(X^{\mathrm{T}}BX)}B, \end{cases} \quad (7.10)$$

where $H_{\phi,1}$ is H_ϕ in (7.9) with $\theta = 1$. We can then obtain the corresponding aligned NEPv (5.3) and linear operator $\mathscr{L}$ by (5.22) in the same way as in Section 7.1.

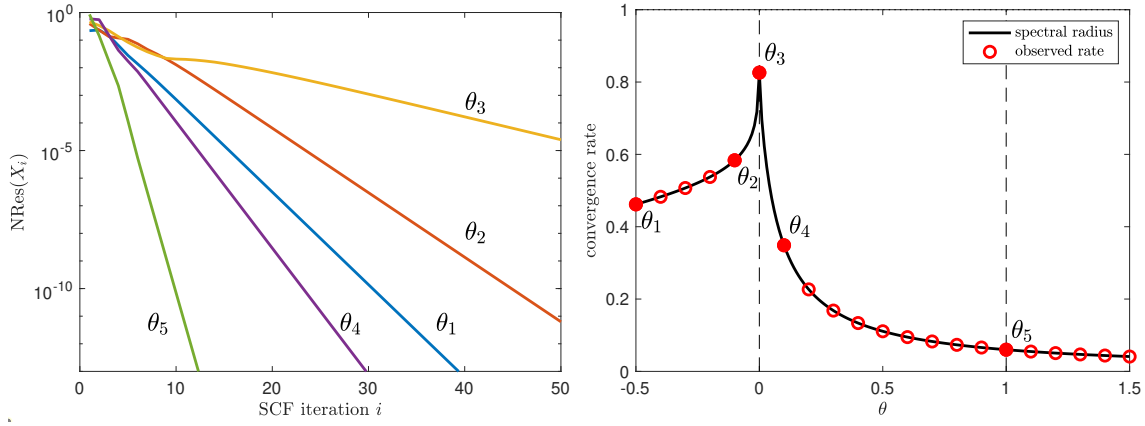

Figure 7.7: Algorithm 4.1 on Example 7.4. *Left:* The iterative history for solving NEPv (7.8) at a few sampled θ (correspondingly marked as $\bullet$ on the right plot). *Right:* The curve of spectral radius $\rho(\mathscr{L})$ as a function of parameter $\theta \in [-0.5, 1.5]$ (based on 200 equally spaced θ), and the observed rates of convergence (marked by $\bullet$ and $\circ$) at a number of values of θ, including those sampled θ on the left plot.

Example 7.4. Consider the random coefficient matrices $A, B \in \mathbb{R}^{n\times n}$ and $D \in \mathbb{R}^{n\times k}$ used in Example 7.1, where $n = 3$ and $k = 1$. We first examine the convergence of Algorithm 4.1 on NEPv (7.8) as θ varies in $[-0.5, 1.5]$. In Figure 7.7, we see that Algorithm 4.1 is convergent for all $\theta \in [0, 1]$, consistent with the global convergence analysis in [48]. We can also see that the spectral radius captures the local convergence rate of the algorithm. For example,

$$\text{at } \theta_4 = 0.1: \qquad \text{observed rate} \approx 0.348738\cdots, \quad \rho(\mathscr{L}) \approx 0.348739\cdots.$$

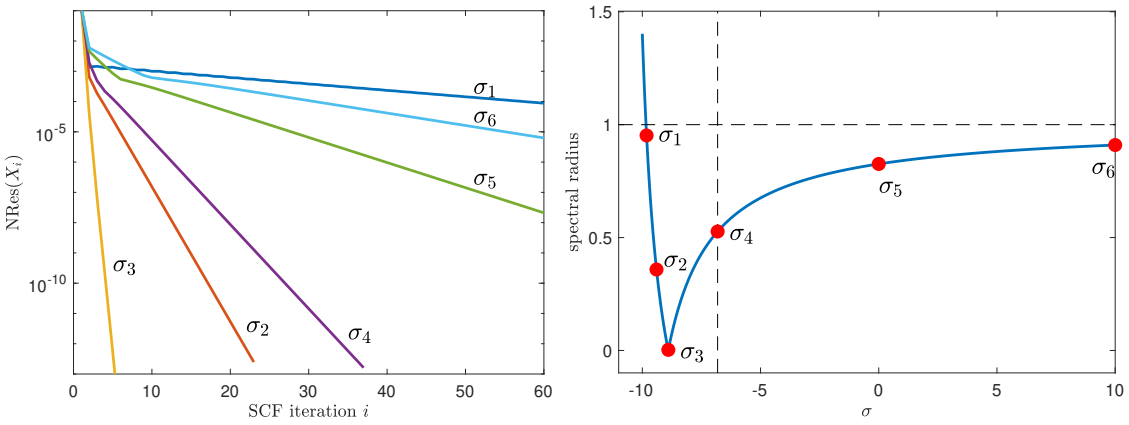

Figure 7.8: Algorithm 6.1 on Example 7.4 with $\theta = 0$ (i.e., α_4 in Figure 7.7, at which Algorithm 4.1 converges slowest). *Left:* The iterative history of Algorithm 6.1 (level-shifted SCF) with a few sampled level-shifts σ (correspondingly marked as $\bullet$ on the right plot). *Right:* The curve of spectral radius $\rho(\mathscr{L}_\sigma)$ as a function of level-shift σ and the observed rates of convergence at the sampled level-shifts σ. The vertical dashed line corresponds to the theoretical lower bound $\sigma_L = -6.81$ by (6.9) (as σ_4 in the plots). Negative shifts can accelerate SCF for the example.

For this example, Algorithm 4.1 runs faster as θ increases from 0. But such a rapid convergence is not guaranteed for general NEPv (7.8), especially as $\theta > 1$; see Example 7.5 below.

Next, we consider the level-shifted SCF (Algorithm 6.1) for NEPv (7.8) at $\theta = 0$, at which the curve of spectral radius in Figure 7.7 peaks. Recall that the NEPv with $\theta = 0$ arises in solving the unbalanced orthogonal Procrustes problems; see, e.g., [56]. Since Algorithm 4.1 is globally convergent in this case, there is no need to use level-shifting to fix the divergence issue of SCF. Nevertheless, the level-shift still helps to speed up the convergence of the algorithm. From the curve of spectral radius in Figure 7.8, the optimal level-shift is achieved at $\sigma_* \approx -8.91$ (as σ_3 in Figure 7.8), at which $\rho(\mathscr{L}_\sigma) \approx 0.7 \times 10^{-3}$, indicating significant acceleration to the algorithm. The theoretical lower bound (6.9) predicts that the level-shifted SCF is locally convergent for any $\sigma \gtrsim \sigma_L = -6.81$ (as σ_4 in Figure 7.8), overestimating the observed one.

An interesting observation here is that the optimal level-shift is a *negative* number. A negative σ would reduce the gap between the kth and $(k+1)$-st eigenvalues (see (6.3)), and it is remarkable that here a negative shift can even greatly accelerate SCF. The negative level-shift for SCF has been briefly mentioned in [4], but its benefits have not been fully understood. Figure 7.8 provides a concrete example to demonstrate this intriguing behavior of negative level-shifting.

Example 7.5. This example demonstrates the potential divergence issue of Algorithm 4.1 for NEPv (7.8) with $\theta > 1$. We consider the following randomly generated matrices

$$A = \begin{bmatrix} 1.145 & -0.095 & 0.514 \\ -0.095 & 0.838 & 1.022 \\ 0.514 & 1.022 & -1.223 \end{bmatrix}, \; B = \begin{bmatrix} 0.582 & -0.037 & 0.025 \\ -0.037 & 0.183 & 0.043 \\ 0.025 & 0.043 & 0.239 \end{bmatrix}, \; D = \begin{bmatrix} 0.760 & 0.258 \\ 0.011 & 0.774 \\ 0.180 & 0.520 \end{bmatrix}.$$

Figure 7.9 reports the convergence of Algorithm 4.1 on NEPv (7.8) as θ varies in $[0, 6]$. From the curve of spectral radius, it can be seen that Algorithm 4.1 converges for all $\theta \in [0, 1]$, and the spectral radius captures very well the observed convergence rates, e.g.,

$$\text{at } \theta_5 = 4.75: \qquad \text{observed rate} \approx 0.977615\cdots, \quad \rho(\mathscr{L}) \approx 0.977613\cdots.$$

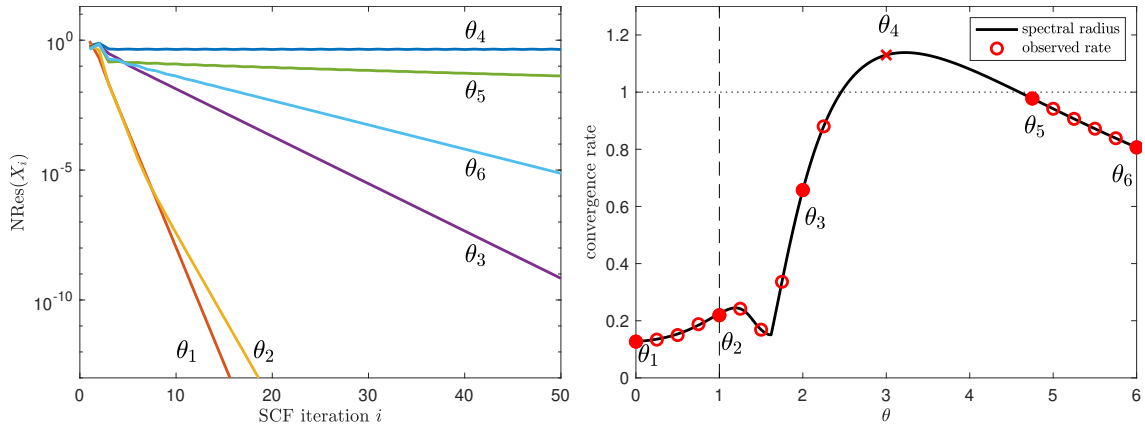

Figure 7.9: Algorithm 4.1 on Example 7.5. *Left:* The iterative history for solving NEPv (7.8) at a few sampled θ (correspondingly marked as as $\bullet$ and $\times$ on the right plot). *Right:* the curve of spectral radius $\rho(\mathscr{L})$ as a function of parameter $\theta \in [0, 6]$ (based on 200 equally spaced θ), and the observed rates of convergence (marked by $\bullet$ and $\circ$) at a number of values of θ, including those sampled θ on the left plot, and '$\times$' indicates that SCF is divergent.

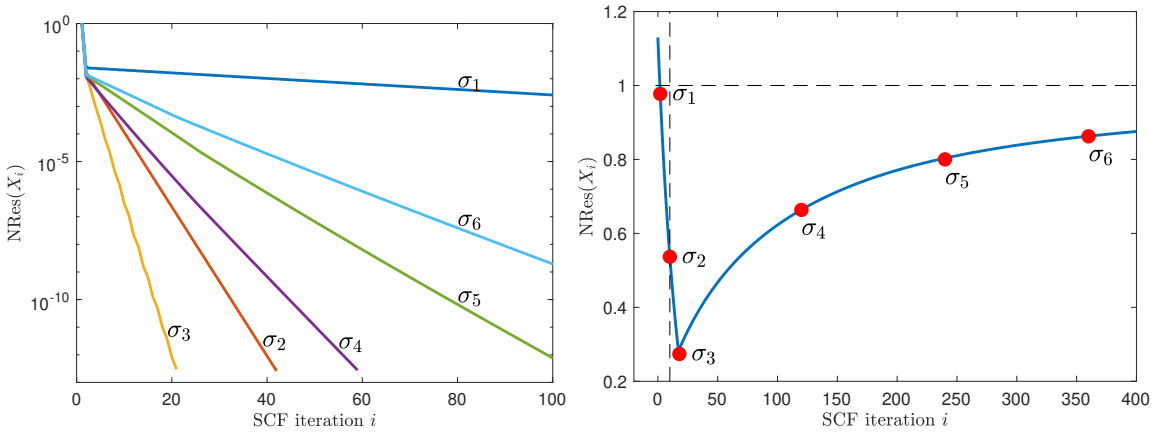

Figure 7.10: Algorithm 6.1 on Example 7.5 with $\theta = 3.0$ (i.e., θ_4 in Figure 7.9, at which Algorithm 4.1 diverges). *Left:* The iterative history of Algorithm 6.1 (level-shifted SCF) with a few sampled level-shifts σ (correspondingly marked as $\bullet$ on the right plot). *Right:* the curve of spectral radius $\rho(\mathscr{L}_\sigma)$ as a function of level-shift σ and the observed rates of convergence at the sampled level-shifts σ. The vertical dashed line corresponds to the theoretical lower bound $\sigma_L = 10.02$ given by (6.9) (as σ_2 in the plots).

The curve of spectral radius indicates that Algorithm 4.1 fails to converge for θ approximately in the interval $[2.46, 4.59]$. For example, at $\theta = 3.0$, the normalized residual $\mathrm{NRes}(X_i)$ by Algorithm 4.1 oscillate between two close numbers $0.453\cdots$ and $0.437\cdots$ after about 20 iterations.

To find the solution of the NEPv when $\rho(\mathscr{L}) > 1$, we apply the level-shifted SCF (Algorithm 6.1) with $\sigma = 40$. The effectiveness of level-shifting is also demonstrated in Figure 7.10 for solving NEPv (7.8) with $\theta = 3.0$ (i.e., θ_4 in Figure 7.9), at which Algorithm 4.1 diverges. The optimal level-shift is $\sigma_* \approx 17.21$ (as σ_3 in the plots) with $\rho(\mathcal{L}_\sigma) \approx 0.282$. The theoretical lower bound of the level-shift σ by (6.9) is $\sigma_L = 10.02$ (as σ_2 in the plots).

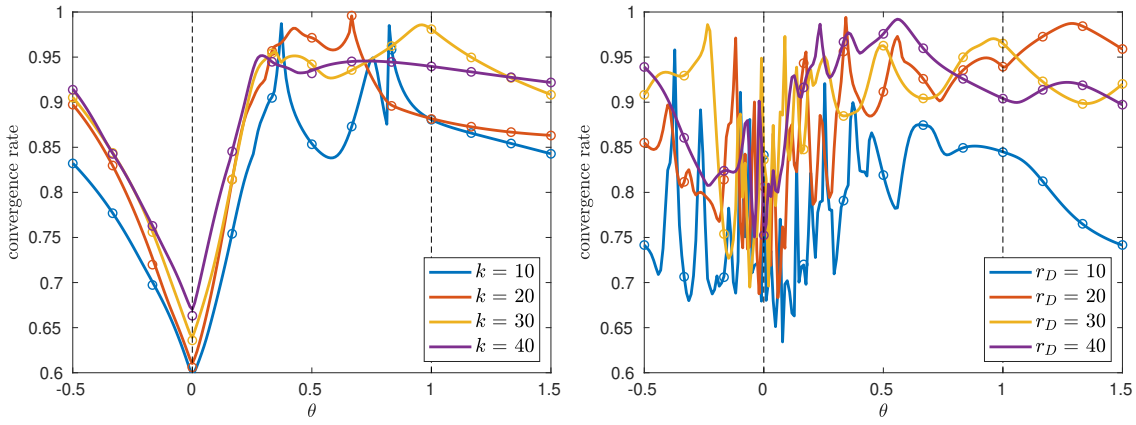

Figure 7.11: Algorithm 4.1 on Example 7.6: The curves of spectral radius $\rho(\mathscr{L})$ as a function of $\theta \in [-0.5, 1.5]$ (based on 200 equally spaced θ) and the observed rates of convergence of Algorithm 4.1 at a few sampled θ (marked by $\circ$). *Left:* Full-rank $D \in \mathbb{R}^{n\times k}$. *Right:* Rank-deficient $D \in \mathbb{R}^{n\times k}$ with $k = 50$ and rank $r_{\scriptscriptstyle D} := \mathrm{rank}(D) < k$.

Example 7.6. This example examines the convergence of Algorithm 4.1 on NEPv (7.8) for more general cases of $D \in \mathbb{R}^{n\times k}$, as in the previous Example 7.3. We use the same testing matrices as in Example 7.3. Recall that a set of four full-rank $D \in \mathbb{R}^{n\times k}$, with $k = 10, 20, 30, 40$, and the other set of four rank-deficient $D \in \mathbb{R}^{n\times 50}$ with $r_{\scriptscriptstyle D} := \mathrm{rank}(D) = 10, 20, 30, 40$, are tested. For each set of testing matrices we apply Algorithm 4.1 on NEPv (7.8) with $\theta \in [-0.5, 1.5]$ and generate the curves of spectral radius $\rho(\mathscr{L})$ as a function in θ as shown in Figure 7.11. It is observed that the spectral radius remains less than 1 in all testing cases, and thus Algorithm 4.1 is locally convergent. The curves of spectral radius demonstrate very different patterns for the full-rank and rank-deficient cases: They appear to be more sensitive to θ for a rank-deficient D. Despite their wild oscillations, the curves of spectral radius still match very well with our theoretical convergence rate at all sampled θ, which further confirms our convergence analysis in Section 5.

8 Concluding Remarks

We investigated a class of NEPv (2.10) as arising from solving optimization problem (2.3) on the Stiefel manifold:

$$\max_{X\in\mathbb{O}^{n\times k}} f(X) \qquad \text{with} \qquad f(X) := \phi(X) + \psi(X)\cdot \mathrm{tr}(X^{\mathrm{T}}D), \tag{2.3}$$

whose objective function $f(\,\cdot\,)$ is not invariant upon substitution $X \leftarrow XQ$ with $Q \in \mathbb{O}^{k\times k}$. Consequently, the resulting NEPv does not have the unitary invariance property, unlike those commonly studied as in [4]. We have shown that any global optimizer X_* of (2.3) is a D-regular eigenbasis matrix of NEPv (2.10), i.e., satisfying

$$X_*^{\mathrm{T}}D \succeq 0 \quad \text{and} \quad \mathrm{rank}(X_*^{\mathrm{T}}D) = \mathrm{rank}(D),$$

and that for any X that has orthonormal columns and is close to X_*, the NEPv can be reformulated to another NEPv, called the aligned NEPv, that is unitarily invariant. This novel reformulation essentially reduces the local convergence analysis of the SCF-type iteration in Algorithm 4.1 for NEPv (2.10) to the case that had been studied in [4] for general unitarily invariant NEPv, once

some technicalities are taken care of. We established closed-form local convergence rates for Algorithms 4.1 and 6.1 and built a theoretical foundation for the application of a level-shifting scheme. Our theoretical analysis has been confirmed by extensive numerical experiments.

Throughout this paper, our presentation is restricted to the real number field $\mathbb{R}$. This restriction is more for simplicity and clarity than the capability of our techniques to deal with a problem similar to (2.3) but in the complex number field $\mathbb{C}$. In fact, our approach can be extended to handle, more generally, the following optimization problem on the complex Stiefel manifold:

$$\max_{X\in\mathbb{C}^{n\times k},\, X^{\mathrm{H}}X=I_k} f(X) \qquad \text{with} \qquad f(X) := \phi(X) + \psi(X)\cdot\eta(\mathrm{Re}(\,\mathrm{tr}(X^{\mathrm{H}}D)\,)), \tag{8.1}$$

where the superscript H takes the complex conjugate transpose of a matrix, $\mathrm{Re}(\,\cdot\,)$ takes the real part of a complex number, ϕ and ψ are real-valued functions in $X\in\mathbb{C}^{n\times k}$, continuously differentiable with respect to the real and imaginary parts of X and satisfying the unitary invariance property:

$$\phi(XQ)=\phi(X), \quad \psi(XQ)=\psi(X) \quad \text{for } X\in\mathbb{C}^{n\times k},\, Q\in\mathbb{C}^{k\times k} \text{ such that } Q^{\mathrm{H}}Q=I_k,$$

$\psi(\,\cdot\,)>0$ is a positive function, and $\eta:\mathbb{R}\to\mathbb{R}$ is a differentiable and monotonically increasing function, i.e., $\eta'(t)>0$ for all $t\in\mathbb{R}$. Due to that X is a complex matrix variable, we will need to use Wirtinger derivatives [9, 26] to establish the KKT condition of (8.1) and its corresponding NEPv. On the other hand, because of the monotonicity in η, we can still apply the same alignment operation as discussed in Section 3.1 and establish analogous necessary conditions for the global maximizers of (8.1) as in Theorem 3.1. A complete treatment is left to future work.

A Differentiability and Polar Decomposition

Recall definition (2.2) for the Fréchet derivative of a (Fréchet) differentiable function $F:\mathbb{R}^{m\times n}\to\mathbb{R}^{p\times q}$ at X along direction Y. In this section, we derive the expressions of the Fréchet derivatives of the polar factors of a matrix having full column rank.

The polar decomposition of a matrix $Z\in\mathbb{R}^{m\times p}$ having full column rank, i.e., $\mathrm{rank}(Z)=p$, is given by

$$Z = Q\,M, \tag{A.1a}$$

where $Q\in\mathbb{O}^{m\times p}$ is the orthonormal polar factor and $M\in\mathbb{R}^{p\times p}$ is the symmetric polar factor that is also positive definite. The polar factors are unique [22, 29] and differentiable with respect to the matrix variable Z. In fact, they can be expressed explicitly as

$$M = \left(Z^{\mathrm{T}}Z\right)^{1/2} \qquad \text{and} \qquad Q = Z\,\left(Z^{\mathrm{T}}Z\right)^{-1/2}, \tag{A.1b}$$

where $(\,\cdot\,)^{1/2}$ is the positive semi-definite square root of a positive semi-definite matrix. The Fréchet derivatives of M and Q with respect to Z are detailed in Lemma A.1 below. These results are not new, but we will provide a quick derivation for self-containedness.

Lemma A.1. *Let $Z\in\mathbb{R}^{m\times p}$ have full column rank. Its polar factors M and Q, as given in* (A.1b)*, are Fréchet differentiable in Z along direction $Y\in\mathbb{R}^{m\times p}$, with their derivatives given by*

$$\mathbf{D}M(Z)[Y] = L \qquad \textit{and} \qquad \mathbf{D}Q(Z)[Y] = (\,Y - QL\,)\cdot M^{-1}, \tag{A.2}$$

where $L\in\mathbb{R}^{m\times p}$ is the solution to the Lyapunov equation

$$M\cdot L + L\cdot M = (Y^{\mathrm{T}}Z + Z^{\mathrm{T}}Y). \tag{A.3}$$

Proof. The differentiability of the factors in (A.1b) follows from the differentiability of the matrix square root function $(\,\cdot\,)^{1/2}$ [22, Chapter 3]. To obtain closed-form formulas for the derivatives, let $M + \delta M(t) = \big[(Z+tY)^{\mathrm{T}}(Z+tY)\big]^{1/2}$ where $t \in \mathbb{R}$ is assumed sufficiently small. Then $(M + \delta M(t))^2 = (Z+tY)^{\mathrm{T}}(Z+tY)$, expanding which to get

$$\delta M(t) \cdot M + M \cdot \delta M(t) = t(Z^{\mathrm{T}}Y + Y^{\mathrm{T}}Z) + \mathcal{O}(t^2).$$

Hence, $\delta M(t) = tL + \mathcal{O}(t^2)$, yielding the first formula in (A.2). Now, it follows from $Q + \delta Q(t) = (Z+tY)(M+\delta M(t))^{-1}$ that

$$\begin{aligned}\delta Q(t) &= tYM^{-1} - ZM^{-1} \cdot \delta M(t) \cdot M^{-1} + \mathcal{O}(t^2)\\ &= t(YM^{-1} - Q \cdot L \cdot M^{-1}) + \mathcal{O}(t^2),\end{aligned}$$

yielding the second formula in (A.2). □

B Aligned NEPv via Bi-level Maximization

We will show that the aligned NEPv (5.3) can also be derived from the KKT condition of another maximization problem that is equivalent to (2.3), as we mentioned at the end of Section 5.1. This approach provides a seemingly more direct derivation of the aligned NEPv (5.3) than what is presented in Section 5.1, but it does not seem to yield rank-preserving condition (3.2) at optimality, a cornerstone of our local convergence analysis of the SCF-type iteration in Algorithm 4.1 for solving NEPv (2.10), whose rationale comes from analyzing the formulation of NEPv (2.10).

We can state optimization (2.3) equivalently as

$$\max_{X \in \mathbb{O}^{n\times k}, Q \in \mathbb{O}^{k\times k}} \phi(XQ) + \psi(XQ) \cdot \mathrm{tr}(Q^{\mathrm{T}}X^{\mathrm{T}}D), \tag{B.1}$$

in the sense that any maximizer of one optimization problem will lead to a maximizer of the other. Due to that ϕ and ψ are unitary invariant, we can drop Q from the argument of both $\phi(XQ)$ and $\psi(XQ)$, and write (B.1) as a *bi-level optimization*:

$$\max_{X \in \mathbb{O}^{n\times k}} \left[\phi(X) + \psi(X) \cdot \left(\max_{Q \in \mathbb{O}^{k\times k}} \mathrm{tr}(Q^{\mathrm{T}}X^{\mathrm{T}}D)\right)\right], \tag{B.2}$$

where we have used $\psi(X) > 0$. Recall that the inner optimization has been considered in Section 3.1, and its solution can be given in terms of the polar decomposition (3.10). Consequently, maximization problem (B.2) is reduced to

$$\max_{X \in \mathbb{O}^{n\times k}} g(X) \qquad \text{with} \qquad g(X) := \phi(X) + \psi(X) \cdot \mathrm{tr}(M), \tag{B.3}$$

where $M \succeq 0$ is the positive semidefinite polar factor of $X^{\mathrm{T}}D = QM$ in (3.10). The two maximization problems (B.3) and (2.3) are clearly equivalent: X_* is a global maximizer of (B.3) if and only if any $\widetilde{X}_* \in [\![X_*]\!]$ is a global maximizer of original (2.3).

To derive the first-order optimality condition of (B.3), we define its Lagrangian function as

$$\mathcal{L}(X, \Gamma) = \phi(X) + \psi(X) \cdot \mathrm{tr}(M) - \frac{1}{2}\mathrm{tr}\big(\Gamma^{\mathrm{T}}[X^{\mathrm{T}}X - I_k]\big), \tag{B.4}$$

where the symmetric $\Gamma \in \mathbb{R}^{k\times k}$ is the matrix of multipliers. Recall that M depends on X. Under the rank-preserving condition $\operatorname{rank}(X^{\mathrm{T}}D) = \operatorname{rank}(D)$, $M \equiv M(X)$ is differentiable with respect to X by Lemma 5.3. Then the function $\gamma(X) := \operatorname{tr}(M)$ has its derivative along direction E given by

$$\mathbf{D}\gamma(X)[E] = \operatorname{tr}\left(\mathbf{D}M(X)[E]\right) = \operatorname{tr}\left(Q_{\mathbf{o}}D^{\mathrm{T}}E\right) \qquad \Rightarrow \qquad \frac{\partial\gamma(X)}{\partial X} = DQ_{\mathbf{o}}^{\mathrm{T}}.$$

Hence,

$$\begin{aligned}\frac{\partial\mathcal{L}(X,\Gamma)}{\partial X} &= H_\phi(X)X + \operatorname{tr}(M)\cdot H_\psi(X)X + \psi(X)\cdot DQ_{\mathbf{o}}^{\mathrm{T}} - X\Gamma \\ &= G(X)X - X\cdot\Lambda(X),\end{aligned} \tag{B.5}$$

where $\Lambda(X) = \psi(X)Q_{\mathbf{o}}D^{\mathrm{T}}X + \Gamma$, and we have used (2.9) for the derivatives of ϕ and ψ, and the definition of $G(X)$ in (5.6). Finally, the first-order optimality condition $\partial\mathcal{L}(X,\Gamma)/\partial X = 0$ leads to the aligned NEPv (5.3).

C Proof of Theorem 6.1

We exploit the fact that the global maximizer X_* of (2.3) is also a global maximizer of (B.3). Let $\mathcal{L}$ be the Lagrangian function for the constrained optimization problem (B.3), as in (B.4). For $X, E \in \mathbb{R}^{n\times k}$, expand the Lagrangian function up to the second order to get

$$\mathcal{L}(X+tE,\Gamma) = \mathcal{L}(X,\Gamma) + t\mathbf{D}_1\mathcal{L}(X,\Gamma)[E] + t^2\frac{1}{2}\mathbf{D}_1^2\mathcal{L}(X,\Gamma)[E,E] + \mathcal{O}(t^3)$$

for $t \in \mathbb{R}$ sufficiently small, where $\mathbf{D}_1\mathcal{L}(\cdot,\cdot)[\cdot]$ stands for partial differentiation with respect to the first matrix argument of $\mathcal{L}$, and $\mathbf{D}_1^2\mathcal{L}(X,\Gamma)(\cdot,\cdot)$, a bilinear form, is the partial Hessian operator of $\mathcal{L}$ with respect to the first matrix argument. By the standard second-order optimality condition [37, Theorem 12.5], any global maximizer X_* of (B.3) must satisfy

$$\mathbf{D}_1^2\mathcal{L}(X_*,\Gamma_*)[E,E] \le 0 \qquad \text{for all } E \in \mathbb{R}^{n\times k} \text{ with } E^{\mathrm{T}}X_* = 0, \tag{C.1}$$

where Γ_* contains the optimal multipliers associated with X_*. On the other hand, we claim that

$$\mathbf{D}_1^2\mathcal{L}(X_*,\Gamma_*)[E,E] = \operatorname{tr}(\,Z^{\mathrm{T}}\mathcal{Q}(Z)\,) \qquad \text{for } E = X_{*\perp}Z, \tag{C.2}$$

where $\mathcal{Q}(Z)$ is given as in (6.7). To justify (C.2), we find, by the gradient of $\mathcal{L}$ in (B.5), that

$$\mathbf{D}_1\mathcal{L}(X,\Gamma)[E] = \operatorname{tr}\Big(E^{\mathrm{T}}G(X)X - E^{\mathrm{T}}X\cdot\Lambda(X)\Big),$$

where $\Lambda(X) = \psi(X)Q_{\mathbf{o}}D^{\mathrm{T}}X + \Gamma$. Differentiate it with respect to X to obtain

$$\begin{aligned}\mathbf{D}_1^2\mathcal{L}(X,\Gamma)[E,E] &= \mathbf{D}_1\Big(\mathbf{D}_1\mathcal{L}(X,\Gamma)[E]\Big)[E] \\ &= \operatorname{tr}\Big(E^{\mathrm{T}}\cdot\mathbf{D}_1G(X)[E]\cdot X + E^{\mathrm{T}}G(X)E - E^{\mathrm{T}}X\cdot\mathbf{D}\Lambda(X)[E] - E^{\mathrm{T}}E\cdot\Lambda(X)\Big).\end{aligned}$$

Noticing for $X = X_*$ and $E = X_{*\perp}Z$, we have

$$E^{\mathrm{T}}X = 0, \qquad E^{\mathrm{T}}E = Z^{\mathrm{T}}Z, \qquad E^{\mathrm{T}}G(X_*)E = Z^{\mathrm{T}}\Lambda_{*\perp}Z,$$

where the last equation follows from (6.3), and by the first-order optimality condition $G(X_*)X_* = X_*\cdot\Lambda(X_*)$ we have $\Lambda(X_*) = X_*^{\mathrm{T}}G(X_*)X_* = \Lambda_*$. Therefore,

$$\mathbf{D}_1^2\mathcal{L}(X_*,\Gamma_*)[E,E] = \operatorname{tr}\Big(Z^{\mathrm{T}}X_{*\perp}^{\mathrm{T}}\cdot\mathbf{D}G(X_*)[X_{*\perp}Z]\cdot X_* + Z^{\mathrm{T}}\Lambda_{*\perp}Z - Z^{\mathrm{T}}Z\Lambda_*\Big) = \operatorname{tr}(Z^{\mathrm{T}}\mathcal{Q}(Z)),$$

which is (C.2). Finally, (6.10) is a consequence of (C.1) and (C.2).

Acknowledgment

The authors wish to thank the anonymous referees for their insightful suggestions and comments that improved the presentation of the paper.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms On Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[2] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen. *LAPACK Users' Guide*. SIAM, Philadelphia, 3rd edition, 1999.

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

[4] 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(1):301–327, 2022.

[5] Z. Bai, D. Lu, and B. Vandereycken. Robust Rayleigh quotient minimization and nonlinear eigenvalue problems. *SIAM J. Sci. Comput.*, 40(5):A3495–A3522, 2018.

[6] W. Bao and Q. Du. Computing the ground state solution of Bose–Einstein condensates by a normalized gradient flow. *SIAM J. Sci. Comput.*, 25(5):1674–1697, 2004.

[7] A. Ben-Israel and T. N. E. Greville. *Generalized Inverses: Theory and Applications*. Springer, New York, 2nd edition, 2003.

[8] I. Borg and J. Lingoes. *Multidimensional Similarity Structure Analysis*. Springer-Verlag, New York, 1987.

[9] D. H. Brandwood. A complex gradient operator and its application in adaptive array theory. In *IEE Proceedings H (Microwaves, Optics and Antennas)*, volume 130, pages 11–16. IET Digital Library, 1983.

[10] Y. Cai, L.-H. Zhang, Z. Bai, and R.-C. Li. On an eigenvector-dependent nonlinear eigenvalue problem. *SIAM J. Matrix Anal. Appl.*, 39(3):1360–1382, 2018.

[11] E. Cancès, R. Chakir, and Y. Maday. Numerical analysis of nonlinear eigenvalue problems. *J. Sci. Comput.*, 45(1):90–117, 2010.

[12] E. Cancès, G. Kemlin, and A. Levitt. Convergence analysis of direct minimization and self-consistent iterations. *SIAM J. Matrix Anal. Appl.*, 42(1):243–274, 2021.

[13] M. T. Chu and N. T. Trendafilov. The orthogonally constrained regression revisited. *J. Comput. Graph. Stat.*, 10(4):746–771, 2001.

[14] J. P. Cunningham and Z. Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *J. Mach. Learning Res.*, 16:2859–2900, 2015.

[15] C. Davis and W. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7:1–46, 1970.

[16] J. P. Van de Geer. Linear relations among k sets of variables. *Psychometrika*, 49:70–94, 1984.

[17] J. Demmel. *Applied Numerical Linear Algebra*. SIAM, Philadelphia, PA, 1997.

[18] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1999.

[19] L. Eldén and H. Park. A Procrustes problem on the Stiefel manifold. *Numer. Math.*, 82:599–619, 1999.

[20] G. H. Golub and C. F. Van Loan. *Matrix Computations*. JHU press, 4th edition, 2013.

[21] J. C. Gower and G. B. Dijksterhuis. *Procrustes Problems*. Oxford University Press, New York, 2004.

[22] N. J. Higham. *Functions of Matrices: Theory and Computation*. SIAM, 2008.

[23] J. R. Hurley and R. B. Cattell. The Procrustes program: producing direct rotation to test a hypothesized factor structure. *Comput. Behav. Sci.*, 7:258–262, 1962.

[24] E. Jarlebring, S. Kvaal, and W. Michiels. An inverse iteration method for eigenvalue problems with eigenvector nonlinearities. *SIAM J. Sci. Comput.*, 36(4):A1978–A2001, 2014.

[25] 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*, pages 263–279. Springer, 2014.

[26] K. Kreutz-Delgado. The complex gradient operator and the CR-calculus. Technical Report Course Lecture Suppl. ECE275A., Dept. Elect. Comput. Eng., University of California San Diego, San Diego, CA, 2005. Available at http://dsp.ucsd.edu/kreutz/.

[27] R.-C. Li. A perturbation bound for the generalized polar decomposition. *BIT*, 33:304–308, 1993.

[28] R.-C. Li. New perturbation bounds for the unitary polar factor. *SIAM J. Matrix Anal. Appl.*, 16:327–332, 1995.

[29] R.-C. Li. Matrix perturbation theory. In L. Hogben, R. Brualdi, and G. W. Stewart, editors, *Handbook of Linear Algebra*, page Chapter 21. CRC Press, Boca Raton, FL, 2nd edition, 2014.

[30] W. Li and W. Sun. Perturbation bounds for unitary and subunitary polar factors. *SIAM J. Matrix Anal. Appl.*, 23:1183–1193, 2002.

[31] 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(2):546–558, 2014.

[32] X.-G. Liu, X.-F. Wang, and W.-G. Wang. Maximization of matrix trace function of product Stiefel manifolds. *SIAM J. Matrix Anal. Appl.*, 36(4):1489–1506, 2015.

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

[34] R. Meyer. Nonlinear eigenvector algorithms for local optimization in multivariate data analysis. *Linear Algebra Appl.*, 264:225–246, 1997.

[35] T. T. Ngo, M. Bellalij, and Y. Saad. The trace ratio optimization problem for dimensionality reduction. *SIAM J. Matrix Anal. Appl.*, 31(5):2950–2971, 2010.

[36] T. T. Ngo, M. Bellalij, and Y. Saad. The trace ratio optimization problem. *SIAM Review*, 54(3):545–569, 2012.

[37] J. Nocedal and S. Wright. *Numerical Optimization*. Springer Science & Business Media, 2006.

[38] V. R. Saunders and I. H. Hillier. A level–shifting method for converging closed shell Hartree–Fock wave functions. *Int. J. Quantum Chem.*, 7(4):699–705, 1973.

[39] R. E. Stanton. Intrinsic convergence in closed-shell SCF calculations. A general criterion. *J. Chem. Phys.*, 75(11):5416–5422, 1981.

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

[41] S. Sun, L. Mao, Z. Dong, and L. Wu. *Multiview Machine Learning*. Springer, 2019.

[42] A. Szabo and N. S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Courier Corporation, 2012.

[43] Z. Teng and R.-C. Li. Variations of orthonormal basis matrices of subspaces. *Numer. Alg., Contr. Optim.*, 2023. to appear.

[44] 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(1):16–27, 2004.

[45] F. Tudisco and D. J. Higham. A nonlinear spectral method for core-periphery detection in networks. *SIAM J. Math. Data Science*, 1(2):269–292, 2019.

[46] P. Upadhyaya, E. Jarlebring, and E. H. Rubensson. A density matrix approach to the convergence of the self-consistent field iteration. *Numer. Alg., Contr. Optim.*, 42:99–115, 2021.

[47] L. Wang, L.-H. Zhang, Z. Bai, and R.-C. Li. Orthogonal canonical correlation analysis and applications. *Opt. Methods Soft.*, 35(4):787–807, 2020.

[48] L. Wang, L.-H. Zhang, and R.-C. Li. Trace ratio optimization with an application to multi-view learning. *Math. Program.*, 201:97–131, 2023. doi.org/10.1007/s10107-022-01900-w.

[49] 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(4):1773–1788, 2009.

[50] 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(5):1854–1875, 2007.

[51] L.-H. Zhang and R.-C. Li. Maximization of the sum of the trace ratio on the Stiefel manifold, I: Theory. *SCIENCE CHINA Math.*, 57(12):2495–2508, 2014.

[52] L.-H. Zhang and R.-C. Li. Maximization of the sum of the trace ratio on the Stiefel manifold, II: Computation. *SCIENCE CHINA Math.*, 58(7):1549–1566, 2015.

[53] L.-H. Zhang, L.-Z. Liao, and M. K. Ng. Fast algorithms for the generalized Foley-Sammon discriminant analysis. *SIAM J. Matrix Anal. Appl.*, 31(4):1584–1605, 2010.

[54] L.-H. Zhang, L.-Z. Liao, and M. K. Ng. Superlinear convergence of a general algorithm for the generalized Foley-Sammon discriminant analysis. *J. Optim. Theory Appl.*, 157(3):853–865, 2013.

[55] 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(2):890–904, 2022.

[56] L.-H. Zhang, W. H. Yang, C. Shen, and J. Ying. An eigenvalue-based method for the unbalanced Procrustes problem. *SIAM J. Matrix Anal. Appl.*, 41(3):957–983, 2020.

[57] Z. Zhang and K. Du. Successive projection method for solving the unbalanced Procrustes problem. *SCIENCE CHINA Math.*, 49(7):971–986, 2006.