# ON OPTIMIZING THE SUM OF RAYLEIGH QUOTIENTS ON THE UNIT SPHERE

by

AOHUD ABDULRAHMAN BINBUHAER

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

May 2019

Acknowledgements

This thesis would not have been possible without my parents support and love. My parents, Abdulrahman Binbuhaer and Aljawharah Alddaham, have been instrumental in making me a strong and confident daughter. My father, a mathematics teacher, was the first inspiring person in my life and the person who believed in me the most and my ambition to be "Dr. Binbuhaer"

I am grateful to my supervisor, Dr. Ren-Cang Li, for his help. I really would like to thank him for his great support and valuable guidance, and for pointing me in the right direction to successfully complete my dissertation. Also, I would like to thank the rest of my dissertation committee members, Dr. David Jorgensen, Dr. Tuncay Aktosun, and Dr. Li Wang, for their advice and time.

I would also like to thank the Mathematics Department at UT Arlington for the help they have provided me over the past years.

Special thanks to The Saudi Arabia Cultural Mission (SACM) and Princess Nourah Bint Abdulrahman University (PNU) for their generous financial support for my Ph.D. study.

Last but not least, my deepest gratitude to my brothers, Turkey, Naif and Meshal, and my sisters, Alanoud and Haya, for their warm love and encouraging words.

April 17, 2019

Abstract

ON OPTIMIZING THE SUM OF RAYLEIGH QUOTIENTS ON THE UNIT SPHERE

Aohud Abdulrahman Binbuhaer, Ph.D.

The University of Texas at Arlington, 2019

Supervising Professor: Ren-Cang Li

Given symmetric matrices $A_i, D \in \mathbb{R}^{n\times n}$ and symmetric positive definite matrices $B_i \in \mathbb{R}^{n\times n}$ for $i = 1,\ldots,k$, we are concerned with the solution of the maximization problem:

$$\max_{\|\mathbf{x}\|_2=1} f(\mathbf{x}) \quad \text{with} \quad f(\mathbf{x}) \coloneqq \sum_{i=1}^{k} \frac{\mathbf{x}^\top A_i \mathbf{x}}{\mathbf{x}^\top B_i \mathbf{x}} + \mathbf{x}^\top D\mathbf{x}$$

on the unit sphere: $\mathcal{M} \coloneqq \{\mathbf{x} \in \mathbb{R}^n \quad | \quad \|\mathbf{x}\|_2 = 1\}$. In this dissertation, we establish necessary optimality conditions for local maximizers. Moreover, a self-consistent-field (SCF) iterative method for solving the above problem is introduced and analyzed. We use the Trust-Region SCF iteration to improve the convergence of the SCF method. Furthermore, we show the first and second order optimality conditions for maximizing the function:

$$\sum_{i=1}^{k} \frac{\mathrm{tr}(V^\top A_i V)}{\mathrm{tr}(V^\top B_i V)} + \mathrm{tr}(V^\top D V)$$

over the Stiefel manifold: $\mathbb{O}^{n\times\ell} \coloneqq \{V \in \mathbb{R}^{n\times\ell} \quad | \quad V^\top V = I_\ell\}$ where $\ell \le n$, and $\mathrm{tr}(\cdot)$ stands for the trace of a square matrix. Also, some necessary conditions for the local maximizers of this problem are investigated.

Table Of Contents

List of Illustrations

# Chapter 1

# Introduction

## 1.1 Introduction

The Rayleigh quotient plays a significant role in finding eigenvalues of symmetric matrices. Moreover, maximizing the sum of the Rayleigh quotient and the generalized Rayleigh quotient over the unit sphere has several applications in the real world. It can arise in the downlink of a multi-user multi-input and multi-output (MIMO) system and in the sparse Fisher discriminant analysis in pattern recognition [10]. The computation of the eigenvalues of matrices appears in a wide variety of problems in engineering and physical sciences. Indeed, eigenspace computation is used in several areas, such as control theory, signal processing, structural dynamics, and data mining [1].

We are concerned with the solution of the optimization problem

$$\max_{\|\mathbf{x}\|_2=1} f(\mathbf{x}) \quad \text{with} \quad f(\mathbf{x}) := \sum_{i=1}^{k} \frac{\mathbf{x}^\top A_i \mathbf{x}}{\mathbf{x}^\top B_i \mathbf{x}} + \mathbf{x}^\top D \mathbf{x} \tag{1.1}$$

on the unit sphere

$$\mathcal{M} := \{\mathbf{x} \in \mathbb{R}^n \quad | \quad \|\mathbf{x}\|_2 = 1\},$$

where $A_i, B_i$, and $D \in \mathbb{R}^{n\times n}$ are symmetric matrices with $B_i$ positive definite for all $i = 1, \ldots, k$.

In this chapter, we recall some basic linear algebra concepts. Most of these definitions and properties and their proofs can be found in [1,4,6,11]. Furthermore, we present special and related cases of our optimization problem (1.1), which have been

studied in [19] and [18], and the optimality conditions including necessary conditions for the local and global maximizers are established.

In chapter 2, we establish necessary optimality conditions for local maximizers.

In chapter 3, a self-consistent-field (SCF) iterative method for solving (1.1) is introduced and analyzed. Also, the local convergence of the SCF iteration is discussed in this chapter.

In chapter 4, we present first and second order optimality conditions of maximizing the objective function

$$f(V) := \sum_{i=1}^{k} \frac{\operatorname{tr}(V^\top A_i V)}{\operatorname{tr}(V^\top B_i V)} + \operatorname{tr}(V^\top D V)$$

over the Stiefel manifold

$$\mathbb{O}^{n\times\ell} := \{V \in \mathbb{R}^{n\times\ell} \quad | \quad V^\top V = I_\ell\},$$

where $\ell \le n$, and $A_i, B_i$ and $D \in \mathbb{R}^{n\times n}$ are symmetric matrices with $B_i$ positive definite for $i = 1, \ldots, k$. Also, necessary conditions for the local maximizers of this problem are investigated.

**Notation.** $\mathbb{R}$ denotes the set of all real numbers, and $\mathbb{R}^n$ consists of all $n$-tuples of $\mathbb{R}$. We use the symbol $\mathbb{R}^{n\times\ell}$ for the set of all $n \times \ell$ real matrices. $I_n$ is the $n \times n$ identity matrix. All vectors are column vectors and are in bold. For a vector $\mathbf{x} \in \mathbb{R}^n, \mathbf{x}^\top$ denotes its transpose, which is a row vector $\mathbf{x}^\top = [x_1, x_2, \ldots, x_n]$,

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

For a matrix $V = [v_{ij}] \in \mathbb{R}^{n\times\ell}$, its transpose is $V^\top = [v_{ji}] \in \mathbb{R}^{\ell\times n}$.

The notations $\mathbb{S}_n$ and $\mathbb{S}_n^{++}$ are for the set of symmetric and symmetric positive definite

matrices of size $n \times n$, respectively. We use $\lambda_1(A) \geq \ldots \geq \lambda_n(A)$ to denote the eigenvalues of symmetric $A$ in the descending order. Moreover, to simplify the presentation, we use the notation $\phi_G(\mathbf{x}) := \mathbf{x}^\top G\mathbf{x}$ for a matrix $G$ to be specified.

## 1.2 Preliminaries

**Definition 1.2.1.** Let $A$ be an $n \times n$ symmetric matrix with entries in $\mathbb{R}$. Any nonvanishing vector $\mathbf{v} \in \mathbb{R}^n$ that satisfies

$$A\mathbf{v} = \lambda\mathbf{v}$$

for some $\lambda \in \mathbb{R}$ is called an *eigenvector* of $A$, $\lambda$ is the associated *eigenvalue*, and the pair $(\lambda, \mathbf{v})$ is called an *eigenpair*.

**Remark.** The set of eigenvalues of $A$ is called the *spectrum* of $A$. The eigenvalues of $A$ are the zeros of the *characteristic polynomial* of $A$.

**Theorem 1.2.2** (Weyl)**.** *Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric, and $\{\lambda_i(A)\}_{i=1}^n, \{\lambda_i(B)\}_{i=1}^n$, and $\{\lambda_i(A+B)\}_{i=1}^n$ denote the sets of eigenvalues of $A, B$,and $A+B$ in increasing order, respectively. Then, for $1 \leq m \leq n$,*

$$\lambda_m(A) + \lambda_1(B) \leq \lambda_m(A+B) \leq \lambda_m(A) + \lambda_n(B).$$

**Remark.** We can decompose a symmetric matrix $A$ into its action in an eigenspace $S$ and its action on the orthogonal complement $S^\perp$:

$$A = E_0 A_0 E_0^\top + E_1 A_1 E_1^\top,$$

where $E_0$ is an orthonormal basis matrix for $S$, and $E_1$ is an orthonormal basis matrix for $S^\perp$. Similarly, for $A + H$:

$$A + H = F_0 \Lambda_0 F_0^\top + F_1 \Lambda_1 F_1^\top,$$

where $F_0$ is an orthonormal basis matrix for $S$, and $F_1$ is an orthonormal basis matrix for $S^\perp$.

**Theorem 1.2.3** (Davis-Kahan $\sin(\Theta)$ theorem)**.** *Let $A \in \mathbb{R}^{n\times n}$, $A = E_0 A_0 E_0^\top + E_1 A_1 E_1^\top$ and $A + H = F_0 \Lambda_0 F_0^\top + F_1 \Lambda_1 F_1^\top$ be symmetric matrices with $[E_0, E_1]$ and $[F_0, F_1]$ orthogonal. If the eigenvalues of $A_0$ in $(a, b)$ and the eigenvalues of $\Lambda_1$ are excluded from the interval $(a-\delta, b+\delta)$ for some $\delta > 0$, then*

$$\|F_1^\top E_0\| \le \frac{\|F_1^\top H E_0\|}{\delta}$$

*for any unitarily invariant norm $\|\cdot\|$.*

**Definition 1.2.4.** The Raleigh quotient of a symmetric matrix $A$ and nonzero vector $\mathbf{x}$ is

$$R(A;\mathbf{x}) \equiv \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \quad \text{for} \quad \mathbf{0} \ne \mathbf{x} \in \mathbb{R}^n.$$

**Definition 1.2.5.** Given a symmetric matrix $A$ and a symmetric positive definite matrix $B$, we define the generalized Rayleigh quotient as

$$R(A, B;\mathbf{x}) = \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top B\mathbf{x}} \quad \text{for} \quad \mathbf{0} \ne \mathbf{x} \in \mathbb{R}^n.$$

**Definition 1.2.6.** There are several important types of square matrices. We say that $A \in \mathbb{R}^{n\times n}$ is

- symmetric if $A^\top = A$;
- positive definite if $\mathbf{x}^\top A\mathbf{x} > 0$ for $\mathbf{0} \ne \mathbf{x} \in \mathbb{R}^n$;
- orthogonal if $A^\top A = I_n$.

**Definition 1.2.7.** A *vector norm* on $\mathbb{R}^n$ is a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ with the following properties:

1. $\|\mathbf{x}\| \ge 0$ for all $\mathbf{x} \in \mathbb{R}^n$ with equality if and only if $\mathbf{x} = 0$.
2. $\|\mathbf{x} + \mathbf{y}\| \le \|\mathbf{x}\| + \|\mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.
3. $\|\alpha\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\|$ for all $\alpha \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$.

A useful class of norms are $\ell_p$*-norms* defined by, for $\mathbf{x} = [x_i] \in \mathbb{R}^n$ and $0 \le p \le \infty$,

$$\|\mathbf{x}\|_p = \left(|x_1|^p + |x_2|^p + \cdots + |x_n|^p\right)^{1/p}.$$

In particular

$$\|\mathbf{x}\|_2 = \left(|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2\right)^{1/2} = \sqrt{\mathbf{x}^\top \mathbf{x}}$$

is the Euclidean norm and most often used.

**Definition 1.2.8.** A *matrix norm* on $\mathbb{R}^{m\times n}$ is a function $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ with the following properties:

1. $\|A\| \geq 0$ and $\|A\| = 0$ if and only if $A = 0$ for $A \in \mathbb{R}^{m\times n}$,
2. $\|\alpha A\| = |\alpha| \cdot \|A\|$ for $\alpha \in \mathbb{R}$ and $A \in \mathbb{R}^{m\times n}$,
3. $\|A + B\| \leq \|A\| + \|B\|$ for $A, B \in \mathbb{R}^{m\times n}$.

In numerical analysis, the $F$-norm (Frobenius norm), and the $p$-norms are the most frequently used matrix norms, where

$$\|A\|_F = \left[\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\right]^{1/2}, \quad A = [a_{ij}]$$

and

$$\|A\|_p = \sup_{\mathbf{x}\neq 0} \frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p},$$

the vector $p$-norms are part of this definition.

**Definition 1.2.9.** Suppose that $\hat{\mathbf{x}} \in \mathbb{R}^n$ is an approximation to $\mathbf{x} \in \mathbb{R}^n$. For a given vector norm $\|\cdot\|$, we say that

$$\epsilon_a = \|\hat{\mathbf{x}} - \mathbf{x}\|$$

is the *absolute error*, and

$$\epsilon_r = \frac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|}, \qquad \mathbf{x} \neq \mathbf{0},$$

the *relative error*.

One of the most important decomposition in matrix computations is the singular value decomposition.

**Theorem** (Singular Value Decomposition (SVD))**.** *If $A \in \mathbb{R}^{m\times n}$ then there exist orthogonal matrices*

$$U = [u_1, \ldots, u_m] \in \mathbb{R}^{m\times m}$$

*and*

$$V = [v_1, \ldots, v_n] \in \mathbb{R}^{n \times n}$$

*such that*

$$U^\top A V = \operatorname{diag}(\sigma_1, \ldots, \sigma_p), \qquad p = \min\{m, n\}$$

*where*

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0.$$

**Definition 1.2.10.** A projector $P$ is a linear transformation from $\mathbb{R}^n$ to itself which is idempotent, i.e.,

$$P^2 = P.$$

**Remark.** We define the kernel and range of a projector $P$, denoted by $\ker(\mathrm{P})$ and $\mathcal{R}(P)$, respectively, as

$$\ker(P) := \{\mathbf{y} \in \mathbb{R}^n \quad | \quad P\mathbf{y} = 0\},$$

and

$$\mathcal{R}(P) := \{\mathbf{x} \in \mathbb{R}^n \quad | \quad \mathbf{x} = P\mathbf{y} \quad \text{for some} \quad \mathbf{y} \in \mathbb{R}^n\}.$$

**Proposition 1.2.1.** *If $P$ is a projector, then so is $(I - P)$, and we have* $\ker(\mathrm{P}) = \mathcal{R}(\mathrm{I} - \mathrm{P})$, *and* $ker(P) \bigcap \mathcal{R}(P) = \{\boldsymbol{0}\}$.

There is a particular case in which the subspace $\ker(P)$ is the orthogonal complement of $\mathcal{R}(P)$.

**Definition 1.2.11.** The projector $P$ is said to be an orthogonal projector onto $\mathcal{R}(P)$ when

$$\ker(P) = \mathcal{R}(P)^\perp.$$

**Remark.**

- We say that a set of vectors $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$ in $\mathbb{R}^m$ is *linearly independent* if

$$\sum_{j=1}^{n} \alpha_j \mathbf{a}_j = 0 \quad \Leftrightarrow \quad \alpha_1 = \cdots = \alpha_n = 0.$$

- The set of all linear combinations of $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^m$ is a subspace referred to as the *span* of $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$:

$$\operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_n\} = \Big\{ \sum_j \beta_j \mathbf{a}_j \quad | \quad \beta_1, \ldots, \beta_n \in \mathbb{R} \Big\}.$$

A smooth mapping $\gamma : \mathbb{R} \to \mathcal{M} : t \mapsto \gamma(t)$ is described as a *curve* in a manifold $\mathcal{M}$. Given a smooth real-valued function $f$ on $\mathcal{M}$, the function $f \circ \gamma : t \mapsto f(\gamma(t))$ is a smooth function from $\mathbb{R}$ to $\mathbb{R}$ with a well-defined classical derivative. The set of these smooth real-valued functions defined on a neighborhood of $\mathbf{x}$ is denoted by $\mathfrak{J}_{\mathbf{x}}(\mathcal{M})$. The following definitions and propositions are well known and can be found in [1]

**Definition 1.2.12.** A *tangent vector* $\xi_{\mathbf{x}}$ to a manifold $\mathcal{M}$ at a point $\mathbf{x}$ is a mapping from $\mathfrak{J}_{\mathbf{x}}(\mathcal{M})$ to $\mathbb{R}$ such that there exists a curve $\gamma$ on $\mathcal{M}$ with $\gamma(0) = \mathbf{x}$, satisfying

$$\xi_{\mathbf{x}} f = \dot{\gamma}(0) f := \left. \frac{d(f(\gamma(t)))}{dt} \right|_{t=0}$$

for all $f \in \mathfrak{J}_{\mathbf{x}}(\mathcal{M})$. Such a curve $\gamma$ is said to realize the tangent vector $\xi_{\mathbf{x}}$.

**Definition 1.2.13.** The *tangent space* to $\mathcal{M}$ at $\mathbf{x}$, denoted by $\mathscr{T}_{\mathbf{x}}\mathcal{M}$, is the set of all tangent vectors to $\mathcal{M}$ at $\mathbf{x}$.

**Definition 1.2.14.** Let $\mathcal{X}(\mathcal{M})$ be the set of smooth vector field on $\mathcal{M}$. An affine connection $\nabla$ on a manifold $\mathcal{M}$ is a mapping

$$\nabla : \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \to \mathcal{X}(\mathcal{M}),$$

which is denoted by $(\eta, \xi) \xrightarrow{\nabla} \nabla_\eta \xi$ and satisfies the following axioms:

i) $\nabla_{f\eta + g\chi} \xi = f \nabla_\eta \xi + g \nabla_\chi \xi$,

ii) $\nabla_\eta (a\xi + b\zeta) = a \nabla_\eta \xi + b \nabla_\eta \zeta$,

iii) $\nabla_\eta (f\xi) = (\eta f)\xi + f \nabla_\eta \xi$,

where $\eta, \xi, \chi, \zeta \in \mathcal{X}(\mathcal{M})$, $f, g \in \mathfrak{J}_{\mathbf{x}}(\mathcal{M})$, and $a, b \in \mathbb{R}$. A connection $\nabla$ on $\mathcal{M}$ is called *Riemannian connection* if in addition to the above axioms, $\nabla$ also satisfies

i) $\nabla_\eta \xi - \nabla_\xi \eta = [\eta, \xi]$,

ii) $\chi\langle \eta, \xi \rangle = \langle \nabla_\chi \eta, \xi \rangle + \langle \eta, \nabla_\chi \xi \rangle$ for any $\eta, \xi, \chi \in \mathcal{X}(\mathcal{M})$.

**Definition 1.2.15.** Given a real-valued function $f$ on a Riemannian manifold $\mathcal{M}$, the *Riemannian Hessian* of $f$ at a point $\mathbf{x}$ in $\mathcal{M}$ is the linear mapping $\operatorname{Hess} f(\mathbf{x})$ from $\mathcal{T}_\mathbf{x}\mathcal{M}$ into itself defined by

$$\operatorname{Hess} f(\mathbf{x})[\xi_\mathbf{x}] = \nabla_{\xi_\mathbf{x}} \operatorname{grad} f$$

for all $\xi_\mathbf{x}$ in $T_\mathbf{x}\mathcal{M}$, where $\nabla$ is the Riemannian connection on $\mathcal{M}$.

**Proposition 1.2.2.** *The Riemannian Hessian satisfies the formula*

$$\langle \operatorname{Hess} f[\xi], \eta \rangle = \xi(\eta f) - (\nabla_\xi \eta) f$$

*for all $\xi, \eta \in \mathcal{X}(\mathcal{M})$, where $\mathcal{X}(\mathcal{M})$ is the set of all (smooth) vector fields on $\mathcal{M}$.*

*Proof.* We have $\langle \operatorname{Hess} f[\xi], \eta \rangle = \langle \nabla_\xi \operatorname{grad} f, \eta \rangle$. Since the Riemannian connection leaves the Riemannian metric invariant, this is equivalent to $\xi\langle \operatorname{grad} f, \eta \rangle - \langle \operatorname{grad} f, \nabla_\xi \eta \rangle$. By the definition of the gradient, this yields $\xi(\eta f) - (\nabla_\xi \eta) f$. ∎

**Proposition 1.2.3.** *The Riemannian Hessian is symmetric (in the sense of the Riemannian metric). That is,*

$$\langle \operatorname{Hess} f[\xi], \eta \rangle = \langle \eta, \operatorname{Hess} f[\xi] \rangle$$

*for all $\xi, \eta \in \mathcal{X}(\mathcal{M})$.*

*Proof.* By the previous proposition, the left-hand side is equal to $\xi(\eta f) - (\nabla_\xi \eta) f$ and the right-hand side is equal to $\langle \operatorname{Hess} f(\mathbf{x})[\eta], \xi \rangle = \eta(\xi f) - (\nabla_\eta \xi) f$. Using the symmetry property of the Riemannian connection on the latter expression, we obtain $\eta(\xi f) - (\nabla_\eta \xi) f = \eta(\xi f) - [\eta, \xi] f - (\nabla_\xi \eta) f = \xi(\eta f) - (\nabla_\xi \eta) f$, and the result is proved. ∎

**Remark.** An affine connection $\nabla$ on a manifold $\mathcal{M}$ is said to be *symmetric* when

$$\nabla_\eta \xi - \nabla_\xi \eta = [\eta, \xi],$$

for $\xi, \eta \in \mathcal{X}(\mathcal{M})$.

## 1.3 Past Work

In this section, we review existing results on some special and related cases of (1.1).

- **Case I:** $k = 1, \quad D = 0$ :

$$\max_{\|\mathbf{x}\|_2=1} f(\mathbf{x}) \quad \text{with} \quad f(\mathbf{x}) := \frac{\mathbf{x}^\top A \mathbf{x}}{\mathbf{x}^\top B \mathbf{x}}. \tag{1.2}$$

  The maximization of the objective function in (1.2) on $\mathcal{M}$ is equivalent to computing the extreme eigenpair of a symmetric-definite matrix pair $(A, B)$ [6].

- **Case II:** $k = 1$ :

$$\max_{\|\mathbf{x}\|_2=1} f(\mathbf{x}) \quad \text{with} \quad f(\mathbf{x}) := \frac{\mathbf{x}^\top A \mathbf{x}}{\mathbf{x}^\top B \mathbf{x}} + \mathbf{x}^\top D \mathbf{x}, \tag{1.3}$$

  where $A, D \in \mathbb{S}_n$ and $B \in \mathbb{S}_n^{++}$. Zhang [18] studies the solution of (1.3). There are several practical applications of (1.3), such as the downlink of a multi-user MIMO system [10], and the sparse Fisher discriminant analysis in pattern recognition.

### 1.3.1 Optimality Conditions for a Local Maximizer

We have the following first order optimality condition:

**Theorem 1.3.1** ( [18]). *Let $A, D \in \mathbb{S}_n$ and $B \in \mathbb{S}_n^{++}$. A point $\mathbf{x} \in \mathcal{M}$ is a critical point of $f_{|\mathcal{M}}(\mathbf{x})$ if and only if it satisfies*

$$E(\mathbf{x})\mathbf{x} = \lambda(\mathbf{x})\mathbf{x}, \tag{1.4}$$

*where*

$$\lambda(\mathbf{x}) \coloneqq \phi_B(\mathbf{x})\phi_D(\mathbf{x}), \quad E(\mathbf{x}) \coloneqq A - \frac{\phi_A(\mathbf{x})}{\phi_B(\mathbf{x})}B + \phi_B(\mathbf{x})D. \tag{1.5}$$

*Proof.* The gradient at $\mathbf{x} \in \mathcal{M}$ is

$$g(\mathbf{x}) \coloneqq \operatorname{grad} f_{|\mathcal{M}}(\mathbf{x}) = P_{\mathbf{x}} \nabla f(\mathbf{x}),$$

where $P_{\mathbf{x}} = I_n - \mathbf{x}\mathbf{x}^\top$ is the orthogonal projection onto $\ker(\mathbf{x}^\top)$. The expression of $\nabla f(\mathbf{x})$ is given as

$$\nabla f(\mathbf{x}) = 2\left(\frac{A\phi_B(\mathbf{x}) - B\phi_A(\mathbf{x})}{\phi_B^2(\mathbf{x})} + D\right)\mathbf{x}.$$

$\mathbf{x} \in \mathcal{M}$ is a critical point if and only if $g(\mathbf{x}) = \mathbf{0}$ where

$$g(\mathbf{x}) \coloneqq \operatorname{grad} f_{|\mathcal{M}}(\mathbf{x}) = P_{\mathbf{x}} \nabla f(\mathbf{x}) = 2\left(\frac{A\phi_B(\mathbf{x}) - B\phi_A(\mathbf{x})}{\phi_B^2(\mathbf{x})} + D - \phi_D(\mathbf{x})I_n\right)\mathbf{x}.$$

The conclusion now follows. ∎

**Remark.** Finding a global maximizer of (1.3) cannot be obtained by a generalized eigenvalue problem or the standard eigenvalue problem, since the matrix $E(\mathbf{x})$ is dependent on the eigenvector $\mathbf{x}$. In general, solving the nonlinear eigenvalue problem (1.4) is a more complicated problem.

Next, the second order optimality conditions can be established using the symmetric Hessian operator at $\mathbf{x} \in \mathcal{M}$:

$$\operatorname{Hess} f_{|\mathcal{M}}(\mathbf{x}) : \mathscr{T}_{\mathbf{x}}\mathcal{M} \to \mathscr{T}_{\mathbf{x}}\mathcal{M} : \mathbf{h} \mapsto \nabla_{\mathbf{h}} \operatorname{grad} f_{|\mathcal{M}}(\mathbf{x}),$$

which is expressed by

$$\operatorname{Hess} f_{|\mathcal{M}}(\mathbf{x})[\mathbf{h}] = P_{\mathbf{x}}(\mathbf{D}g(\mathbf{x})[\mathbf{h}]), \quad \forall \mathbf{h} \in \mathscr{T}_{\mathbf{x}}\mathcal{M},$$

where $\mathbf{D}g(\mathbf{x})[\mathbf{h}]$ is the derivative of $g(\mathbf{x})$ at $\mathbf{x} \in \mathcal{M}$ along $\mathbf{h}$.

We now calculate $\mathbf{D}g(\mathbf{x})[\mathbf{h}]$ :

$$\mathbf{D}g(\mathbf{x})[\mathbf{h}] = \frac{2}{\phi_B(\mathbf{x})}\Bigg[E(\mathbf{x}) - \phi_B(\mathbf{x})\phi_D(\mathbf{x})I_n + \frac{4\phi_A(\mathbf{x})}{\phi_B^2(\mathbf{x})}B\mathbf{x}\mathbf{x}^\top B - \frac{2}{\phi_B(\mathbf{x})}\Big(A\mathbf{x}\mathbf{x}^\top B + B\mathbf{x}\mathbf{x}^\top A\Big)\Bigg]\mathbf{h} - 4(\mathbf{x}^\top D\mathbf{h})\mathbf{x}.$$

The second order optimality conditions are presented in the following theorem.

**Theorem 1.3.2** ( [18]). *Let $A, D \in \mathbb{S}_n$ and $B \in \mathbb{S}_n^{++}$.*

*i) If $\overline{\mathbf{x}}$ is a local maximizer of (1.3), then the matrix*

$$K(\overline{\mathbf{x}}) := E(\overline{\mathbf{x}}) = \phi_B(\overline{\mathbf{x}})\phi_D(\overline{\mathbf{x}})I_n + 2P_{\overline{\mathbf{x}}}(D\overline{\mathbf{x}}\overline{\mathbf{x}}^\top B + B\overline{\mathbf{x}}\overline{\mathbf{x}}^\top D)P_{\overline{\mathbf{x}}}$$

*is negative semidefinite.*

*ii) For any critical point, $\overline{\mathbf{x}}$, of $f_{|\mathcal{M}}(\mathbf{x})$, if $K(\overline{\mathbf{x}}) : \mathscr{T}_{\mathbf{x}}\mathcal{M} \to \mathscr{T}_{\mathbf{x}}\mathcal{M}$ is negative definite, then $\overline{\mathbf{x}}$ is a strictly local maximizer of (1.3).*

**Theorem 1.3.3** ( [18]). *Let $A, D \in \mathbb{S}_n$ and $B \in \mathbb{S}_n^{++}$. If $\overline{\mathbf{x}}$ is a local maximizer of (1.3), then it must be a unit eigenvector corresponding to either the largest or the second largest eigenvalue of $E(\overline{\mathbf{x}})$.*

### 1.3.2 Optimality Condition for a Global Maximizer

The necessary global optimality condition for (1.3) will be presented in this subsection, particularly, in the following theorem.

**Theorem 1.3.4** ( [18]). *Let $A, D \in \mathbb{S}_n$ and $B \in \mathbb{S}_n^{++}$. Then for any global maximizer $\hat{\mathbf{x}}$ of (1.3), $(\lambda(\hat{\mathbf{x}}), \hat{\mathbf{x}})$ must be an eigenpair corresponding to the largest eigenvalue of $E(\hat{\mathbf{x}})$, where $\lambda(\hat{\mathbf{x}}) := \phi_D(\hat{\mathbf{x}})\phi_B(\hat{\mathbf{x}})$.*

- **Case III:** Let $k = 1$, and replace the vector $\mathbf{x}$ with a matrix $V \in \mathbb{R}^{n\times\ell}$ to get

$$\max_{V^\top V = I_\ell}\left\{\frac{\mathrm{tr}(V^\top AV)}{\mathrm{tr}(V^\top BV)} + \mathrm{tr}(V^\top DV)\right\}, \tag{1.6}$$

where $\mathrm{tr}(\cdot)$ stands for the trace of a square matrix, $A, B, D \in \mathbb{R}^{n\times n}$ are real symmetric with $B$ positive definite, and integer $\ell < n$. In [19] many properties for $\ell = 1$ were extended to $\ell > 1$. [19] was focused on the theoretical aspect of the maximization problem (1.6). Introduce

$$\phi_A(V) := \mathrm{tr}(V^\top AV), \quad \phi_B(V) := \mathrm{tr}(V^\top BV), \quad \phi_D(V) := \mathrm{tr}(V^\top DV)$$

for any $V \in \mathbb{R}^{n\times\ell}$. The function in (1.6) can now be written as

$$f(V) := \frac{\phi_A(V)}{\phi_B(V)} + \phi_D(V)$$

over the Stiefel manifold

$$\mathbb{O}^{n\times\ell} := \{V \in \mathbb{R}^{n\times\ell} \quad | \quad V^\top V = I_\ell\}.$$

The tangent space $\mathscr{T}_V\mathbb{O}^{n\times\ell}$ at $V \in \mathbb{O}^{n\times\ell}$ is given by

$$\begin{aligned}\mathscr{T}_V\mathbb{O}^{n\times\ell} :=& \{X \in \mathbb{R}^{n\times\ell} : \quad X^\top V + V^\top X = 0\} \\ =& \{X = VK + (I_n - VV^\top)J : \quad K = -K^\top \in \mathbb{R}^{\ell\times\ell}, J \in \mathbb{R}^{n\times\ell}\}.\end{aligned} \tag{1.7}$$

The standard inner product on $\mathscr{T}_V\mathbb{O}^{n\times\ell}$ is given by

$$\langle X, Y\rangle = \mathrm{tr}(X^\top Y), \quad X, Y \in \mathscr{T}_V\mathbb{O}^{n\times\ell}.$$

The orthogonal projection of $Z \in \mathbb{R}^{n\times\ell}$ onto the tangent space $\mathscr{T}_V\mathbb{O}^{n\times\ell}$ is

$$\begin{aligned}\Pi_{\mathscr{T}}(Z) :=& V\Big(\frac{V^\top Z - Z^\top V}{2}\Big) + (I_n - VV^\top)Z \\ =& Z - V\frac{V^\top Z + Z^\top V}{2} = Z - V\mathrm{sym}(V^\top Z) \in \mathscr{T}_V\mathbb{O}^{n\times\ell},\end{aligned}$$

where $\mathrm{sym}(Z) := \frac{1}{2}(Z^\top + Z)$ is the symmetric part of $Z$. Since

$$\frac{\partial f(V)}{\partial V} = 2\Big[A\frac{1}{\phi_B(V)} - B\frac{\phi_A(V)}{(\phi_B(V))^2} + D\Big]V = 2E(V)V,$$

the gradient of the function $f_{|\mathbb{O}^{n\times\ell}}$ is given by

$$\operatorname{grad} f_{|\mathbb{O}^{n\times\ell}}(V) = \Pi_{\mathscr{T}}\Big(\frac{\partial f(V)}{\partial V}\Big) = 2\Big[E(V)V - V(V^\top E(V)V)\Big].$$

The first order optimality condition which is $\operatorname{grad} f_{|\mathbb{O}^{n\times\ell}} = \mathbf{0}$ leads to the following theorem.

**Theorem 1.3.5.** *[19] If $V \in \mathbb{O}^{n\times\ell}$ is a local maximizer of (1.6), then*

$$E(V)V = VM_V, \tag{1.8}$$

*where*

$$E(V) := \Big[A\frac{1}{\phi_B(V)} - B\frac{\phi_A(V)}{(\phi_B(V))^2} + D\Big],$$

*and*

$$M_V := V^\top E(V)V.$$

*Therefore* $\operatorname{eig}(M_V) \subset \operatorname{eig}(E(V))$*, and $V$ is an orthogonal eigenbasis matrix of $E(V)$ associated with its eigenvalues in* $\operatorname{eig}(M_V)$.

Now, let $g(V) := \operatorname{grad} f_{|\mathbb{O}^{n\times\ell}}(V)$, and assume that $V \in \mathbb{O}^{n\times\ell}$ is a critical point, which means $V$ satisfies (1.8). Using the standard second order optimality conditions, we have the following theorem.

**Theorem 1.3.6** ( [19])**.** *If $V$ is a local maximizer of (1.6), then*

$$\operatorname{tr}(X^\top E(V)X) - \operatorname{tr}(XM_V(V)X^\top) + \operatorname{tr}(X^\top G(V,X)V) \le 0 \quad \textit{for} \quad X \in \mathscr{T}_V\mathbb{O}^{n\times\ell}, \tag{1.9}$$

*where*

$$G(V,X) := 4\frac{\operatorname{tr}(V^\top AV)\operatorname{tr}(X^\top BV)}{[\operatorname{tr}(V^\top BX)]^3} - 2\frac{\operatorname{tr}(X^\top BV)A + \operatorname{tr}(X^\top AV)B}{[\operatorname{tr}(V^\top BV)]^2},$$

*and $M_V = V^\top E(V)V$. On the other hand, if $V \in \mathbb{O}^{n\times\ell}$ satisfies (1.8) and if (1.9) is a strict inequality for $X \neq 0$, then $V$ is a strict local maximizer.*

There are other necessary conditions for both the local and global maximizers. A necessary condition for a local maximizer $V$ of (1.6) is stated in terms of the eigenvalues of $E(V)$. From Theorem 1.3.5, $\mathrm{eig}(M_V) \subset \mathrm{eig}(E(V))$, i.e.,

$$\mathrm{eig}(\mathrm{M_V}) = \{\lambda_{\pi_\mathrm{i}}(\mathrm{E(V)}), \quad \mathrm{i} = 1, 2, \ldots, \ell\}, \tag{1.10}$$

where $1 \le \pi_1 < \pi_2 \ldots < \pi_\ell \le n$.

**Theorem 1.3.7** ( [19]). *Let $V \in \mathbb{O}^{n\times\ell}$ be a local maximizer of* (1.6)*, and denote* $\mathrm{eig}(M_V)$ *by* (1.10)*. Then*

$$\lambda_{\pi_1}(E(V)) \ge \lambda_{2\ell}(E(V)). \tag{1.11}$$

Moreover, regarding a necessary condition for a global maximizer, we have the following result:

**Theorem 1.3.8** ( [19]). *Suppose $D$ is positive definite. If $V$ is a global maximizer of* (1.6)*, then it must be an orthonormal eigenbasis matrix of $E(V)$ corresponding to its $\ell$ largest eigenvalues $\lambda_i(E(V))$ for $1 \le i \le \ell$.*

# Chapter 2

## Optimality Conditions for Local Maximizers

In this chapter, we characterize the solution of the maximization problem (1.1) by providing the optimality conditions. We have the cost function $f(\mathbf{x})$

$$f(\mathbf{x}) := \sum_{i=1}^{k} \frac{\mathbf{x}^\top A_i \mathbf{x}}{\mathbf{x}^\top B_i \mathbf{x}} + \mathbf{x}^\top D \mathbf{x}$$

and its restriction $f_{|\mathcal{M}}(\mathbf{x}) : \mathcal{M} \to \mathbb{R}$.

### 2.1 The First Order Optimality Conditions

We can consider the unit sphere $\mathcal{M}$ as a Riemannian submanifold of the Euclidean space $\mathbb{R}^n$ endowed with the natural inner product. The tangent space $\mathscr{T}_{\mathbf{x}}\mathcal{M}$ at any point $\mathbf{x} \in \mathcal{M}$ can be expressed as [1]:

$$\mathscr{T}_{\mathbf{x}}\mathcal{M} = \{\mathbf{z} \quad | \quad \mathbf{z} = P_{\mathbf{x}}\mathbf{y}, \quad \forall \mathbf{y} \in \mathbb{R}^n\},$$

where $P_{\mathbf{x}} = I_n - \mathbf{x}\mathbf{x}^\top$ is the orthogonal projection onto the kernel $\ker(\mathbf{x}^\top) = \mathcal{R}(\mathbf{x})^\perp$. Using the orthogonal projection $P_{\mathbf{x}}$, we can express the gradient of the smooth real-valued function $f_{|\mathcal{M}}(\mathbf{x})$ by [18]:

$$g(\mathbf{x}) := \operatorname{grad} f_{|\mathcal{M}}(\mathbf{x}) = P_{\mathbf{x}} \nabla f(\mathbf{x}). \tag{2.1}$$

Also, a critical point $\mathbf{x} \in \mathcal{M}$ of a function $f_{|\mathcal{M}}(\mathbf{x})$ is a point that satisfies $g(\mathbf{x}) = \mathbf{0}$. The first order optimality condition is given in the following theorem.

**Theorem 2.1.1.** *Let $A_i, D \in \mathbb{S}_n$ and $B_i \in \mathbb{S}_n^{++}$ for $i = 1, \ldots, k$. A point $\mathbf{x} \in \mathcal{M}$ is a critical point of the function $f_{|\mathcal{M}}(\mathbf{x})$ on $\mathcal{M}$ if and only if it satisfies*

$$E(\mathbf{x})\mathbf{x} = \lambda(\mathbf{x})\mathbf{x}, \tag{2.2}$$

*where*

$$\lambda(\mathbf{x}) := \phi_D(\mathbf{x}) \prod_{i=1}^{k} \phi_{B_i}(\mathbf{x}),$$

$$E(\mathbf{x}) := \sum_{i=1}^{k} \prod_{j=1, j\neq i}^{k} \phi_{B_j}(\mathbf{x}) \Big( A_i - \frac{\phi_{A_i}(\mathbf{x})}{\phi_{B_i}(\mathbf{x})} B_i \Big) + \prod_{i=1}^{k} \phi_{B_i}(\mathbf{x}) D, \tag{2.3}$$

*and* $\phi_G(\mathbf{x}) := \mathbf{x}^\top G \mathbf{x} \quad for \quad G \in \{A_i, B_i, D\}$.

*Proof.* First, we calculate $\nabla f(\mathbf{x})$:

$$\begin{aligned}
\nabla f(\mathbf{x}) &= \nabla \Big( \sum_{i=1}^{k} \frac{\mathbf{x}^\top A_i \mathbf{x}}{\mathbf{x}^\top B_i \mathbf{x}} + \mathbf{x}^\top D x \Big) \\
&= \nabla \Big( \sum_{i=1}^{k} \frac{\phi_{A_i}(\mathbf{x})}{\phi_{B_i}(\mathbf{x})} + \phi_D(\mathbf{x}) \Big) \\
&= 2 \Big( \sum_{i=1}^{k} \frac{\phi_{B_i}(\mathbf{x}) A_i - \phi_{A_i}(\mathbf{x}) B_i}{\phi_{B_i}^2(\mathbf{x})} + D \Big) \mathbf{x}.
\end{aligned}$$

Next, we need to find $g(\mathbf{x}) = P_{\mathbf{x}} \nabla f(\mathbf{x})$:

$$\begin{aligned}
g(\mathbf{x}) &= 2 \Bigg[ \sum_{i=1}^{k} \frac{\phi_{B_i}(\mathbf{x}) A_i - \phi_{A_i}(\mathbf{x}) B_i}{\phi_{B_i}^2(\mathbf{x})} \mathbf{x} + D\mathbf{x} \\
&\quad - \sum_{i=1}^{k} \frac{\phi_{B_i}(\mathbf{x}) \mathbf{x}\mathbf{x}^\top A_i \mathbf{x} - \phi_{A_i}(\mathbf{x}) \mathbf{x}\mathbf{x}^\top B_i \mathbf{x}}{\phi_{B_i}^2(\mathbf{x})} - \mathbf{x}\mathbf{x}^\top D \mathbf{x} \Bigg] \\
&= 2 \Bigg[ \sum_{i=1}^{k} \frac{\phi_{B_i}(\mathbf{x}) A_i - \phi_{A_i}(\mathbf{x}) B_i}{\phi_{B_i}^2(\mathbf{x})} + D - \phi_D(\mathbf{x}) I_n \Bigg] \mathbf{x}.
\end{aligned} \tag{2.4}$$

Since $\mathbf{x} \in \mathcal{M}$ is a critical point if and only if $g(\mathbf{x}) = \mathbf{0}$, we have

$$\begin{aligned}
g(\mathbf{x}) = \mathbf{0} &\Leftrightarrow \Bigg( \sum_{i=1}^{k} \frac{\phi_{B_i}(\mathbf{x}) A_i - \phi_{A_i}(\mathbf{x}) B_i}{\phi_{B_i}^2(\mathbf{x})} + D \Bigg) \mathbf{x} = \phi_D(\mathbf{x}) \mathbf{x} \\
&\Leftrightarrow \Bigg( \sum_{i=1}^{k} \prod_{j=1, j\neq i}^{k} \phi_{B_i}(\mathbf{x}) \Big( A_i - \frac{\phi_{A_i}(\mathbf{x})}{\phi_{B_i}(\mathbf{x})} B_i \Big) + \prod_{i=1}^{k} \phi_{B_i}(\mathbf{x}) D \Bigg) \mathbf{x} = \phi_D(\mathbf{x}) \prod_{i=1}^{k} \phi_{B_i}(\mathbf{x}) \mathbf{x},
\end{aligned}$$

as expected. ∎

From the previous result, we can see that any critical point $\overline{\mathbf{x}}$ is an eigenvector of $E(\overline{\mathbf{x}})$ with the corresponding eigenvalue $\lambda(\overline{\mathbf{x}}) = \phi_D(\overline{\mathbf{x}}) \prod_{i=1}^{k} \phi_{B_i}(\overline{\mathbf{x}})$. Now, we denote the set of all critical points of $f_{|\mathcal{M}}(\mathbf{x})$ by $\mathscr{Y}$ i.e.,

$$\mathscr{Y} = \{ \|\mathbf{x}\|_2 = 1 \quad | \quad E(\mathbf{x})\mathbf{x} = \lambda(\mathbf{x})\mathbf{x} \}. \tag{2.5}$$

Moreover, finding a global maximizer of the problem (1.1) is not easy because (2.2) is an eigenvector-dependent nonlinear problem.

## 2.2 The Second Order Optimality Conditions

In this section, we establish the second order optimality conditions for (1.1). First, we need the Hessian of $f_{|\mathcal{M}}(\mathbf{x})$ in order to prove our results regarding a local maximizer of (1.1). We define the symmetric Hessian operator at $\mathbf{x} \in \mathcal{M}$ as

$$\operatorname{Hess} f_{|\mathcal{M}}(\mathbf{x}) : \mathscr{T}_{\mathbf{x}}\mathcal{M} \to \mathscr{T}_{\mathbf{x}}\mathcal{M} : \mathbf{h} \mapsto \nabla_h \operatorname{grad} f_{|\mathcal{M}}(\mathbf{x}).$$

The Riemannian connection is considered as a natural choice of affine connection since it simplifies the analytical derivations. The action of Riemannian Hessian of $f_{|\mathcal{M}}(\mathbf{x})$ on a tangent vector $\mathbf{h} \in \mathscr{T}_{\mathbf{x}}\mathcal{M}$ is given by [1]:

$$\operatorname{Hess} f_{|\mathcal{M}}(\mathbf{x})[\mathbf{h}] = P_{\mathbf{x}}(\mathbf{D}g(\mathbf{x})[\mathbf{h}]), \quad \mathbf{h} \in \mathscr{T}_{\mathbf{x}}\mathcal{M}. \tag{2.6}$$

Now, we calculate $\mathbf{D}g(\mathbf{x})[\mathbf{h}]$:

$$\begin{aligned}
\mathbf{D}g(\mathbf{x})[\mathbf{h}] &= 2\mathbf{D}\left[\left(\sum_{i=1}^{k} \frac{\phi_{B_i}(\mathbf{x})A_i - \phi_{A_i}(\mathbf{x})B_i}{\phi^2_{B_i}(\mathbf{x})} + D - \phi_D(\mathbf{x})I_n\right)\mathbf{x}\right][\mathbf{h}] \\
&= 2\Bigg[\left(\sum_{i=1}^{k} \frac{\phi_{B_i}(\mathbf{x})A_i - \phi_{A_i}(\mathbf{x})B_i}{\phi^2_{B_i}(\mathbf{x})} + D - \phi_D(\mathbf{x})I_n\right)\mathbf{h} \\
&\quad + \sum_{i=1}^{k} \frac{\Big(A_i(2\mathbf{x}^\top B_i\mathbf{h}) - B_i(2\mathbf{x}^\top A_i\mathbf{h})\Big)\phi^2_{B_i}(\mathbf{x})}{\phi^4_{B_i}(\mathbf{x})} \\
&\quad - \frac{\Big(\phi_{B_i}(\mathbf{x})A_i - \phi_{A_i}(\mathbf{x})B_i\Big)\Big(4\phi_{B_i}(\mathbf{x})\mathbf{x}^\top B_i\mathbf{h}\Big)}{\phi^4_{B_i}(\mathbf{x})} - 2\mathbf{x}^\top D\mathbf{h}\Bigg]\mathbf{x}.
\end{aligned}$$

Then,

$$\begin{aligned}
\mathbf{D}g(\mathbf{x})[\mathbf{h}] = \frac{2}{\prod_{i=1}^{k}\phi_{B_i}(\mathbf{x})}\Bigg[E(\mathbf{x}) - \lambda(\mathbf{x})I_n - 2\frac{\prod_{j=1,j\neq i}^{k}\phi_{B_j}(\mathbf{x})}{\phi_{B_i}(\mathbf{x})}\sum_{i=1}^{k}\Bigg((A_i\mathbf{x}\mathbf{x}^\top B_i + B_i\mathbf{x}\mathbf{x}^\top A_i) \\
+ 2\frac{\phi_{A_i}(\mathbf{x})}{\phi_{B_i}(\mathbf{x})}(B_i\mathbf{x}\mathbf{x}^\top B_i)\Bigg)\Bigg]\mathbf{h} - 4(\mathbf{x}^\top D\mathbf{h})\mathbf{x}.
\end{aligned} \tag{2.7}$$

The second order optimality conditions are established in the following theorem.

**Theorem 2.2.1.** *Let $A_i, D \in \mathbb{S}_n$ and $B_i \in \mathbb{S}_n^{++}, i = 1, \ldots, k$.*

*i) The Hessian operator of $f_{|\mathcal{M}}(\mathbf{x})$ at point $\mathbf{x} \in \mathcal{M}$ acting on $\mathbf{h} \in \mathscr{T}_{\mathbf{x}}\mathcal{M}$ is*

$$\operatorname{Hess} f_{|\mathcal{M}}(\mathbf{x})[\mathbf{h}] = \frac{2}{\prod_{i=1}^{k} \phi_{B_i}(\mathbf{x})} H(\mathbf{x})\mathbf{h},$$

*where $H(\mathbf{x}) \in \mathbb{S}_n$ is given by*

$$\begin{aligned} H(\mathbf{x}) = P_{\mathbf{x}}\Bigg[ E(\mathbf{x}) - \lambda(\mathbf{x}) I_n - 2\frac{\prod_{j=1, j\neq i}^{k} \phi_{B_j}(\mathbf{x})}{\phi_{B_i}(\mathbf{x})} \sum_{i=1}^{k} \Bigg( (A_i \mathbf{x}\mathbf{x}^\top B_i + B_i \mathbf{x}\mathbf{x}^\top A_i) \\ + 2\frac{\phi_{A_i}(\mathbf{x})}{\phi_{B_i}(\mathbf{x})} (B_i \mathbf{x}\mathbf{x}^\top B_i) \Bigg) \Bigg] P_{\mathbf{x}}. \end{aligned} \tag{2.8}$$

*ii) If $\overline{\mathbf{x}}$ is a local maximizer of (1.1), then the matrix*

$$\begin{aligned} K(\overline{\mathbf{x}}) := E(\overline{\mathbf{x}}) - \prod_{i=1}^{k} \phi_{B_i}(\overline{\mathbf{x}}) \phi_D(\overline{\mathbf{x}}) I_n + 2P_{\overline{\mathbf{x}}} \sum_{i=1}^{k} \prod_{j=1, j\neq i}^{k} \phi_{B_j}(\overline{\mathbf{x}}) \Bigg[ (B_i \overline{\mathbf{x}}\overline{\mathbf{x}}^\top D + D\overline{\mathbf{x}}\overline{\mathbf{x}}^\top B_i) \\ + \sum_{\ell=1}^{k} \frac{1}{\phi_{B_\ell}(\overline{\mathbf{x}})} \Bigg( (B_i \overline{\mathbf{x}}\overline{\mathbf{x}}^\top A_\ell + A_\ell \overline{\mathbf{x}}\overline{\mathbf{x}}^\top B_i) - \frac{\phi_{A_\ell}(\overline{\mathbf{x}})}{\phi_{B_\ell}(\overline{\mathbf{x}})} (B_i \overline{\mathbf{x}}\overline{\mathbf{x}}^\top B_\ell + B_\ell \overline{\mathbf{x}}\overline{\mathbf{x}}^\top B_i) \Bigg) \Bigg] P_{\overline{\mathbf{x}}} \end{aligned} \tag{2.9}$$

*is negative semidefinite.*

*iii) If $K(\overline{\mathbf{x}}) : \mathscr{T}_{\mathbf{x}}\mathcal{M} \to \mathscr{T}_{\mathbf{x}}\mathcal{M}$ is negative definite, for any $\overline{\mathbf{x}} \in \mathscr{Y} = \{\|\mathbf{x}\|_2 = 1 | \quad E(\mathbf{x})\mathbf{x} = \lambda(\mathbf{x})\mathbf{x}\}$, then $\overline{\mathbf{x}}$ is a strict local maximizer of the problem (1.1).*

*Proof.* i) Using $P_{\mathbf{x}}\mathbf{x} = \mathbf{0}, P_{\mathbf{x}}\mathbf{h} = \mathbf{h}, \quad \forall \mathbf{h} \in \mathscr{T}_{\mathbf{x}}\mathcal{M}$, by (2.6) and (2.7), we have the desired result.

ii) If $\overline{\mathbf{x}}$ is a local maximizer of (1.1), then we have

$$\begin{aligned} E(\overline{\mathbf{x}})\overline{\mathbf{x}} &= \Big( \sum_{i=1}^{k} \prod_{j=1, j\neq i}^{k} \phi_{B_j}(\overline{\mathbf{x}}) (A_i - \frac{\phi_{A_i}(\overline{\mathbf{x}})}{\phi_{B_i}(\overline{\mathbf{x}})} B_i) + \prod_{i=1}^{k} \phi_{B_i}(\overline{\mathbf{x}}) D \Big) \overline{\mathbf{x}} \\ &= \phi_D(\overline{\mathbf{x}}) \prod_{i=1}^{k} \phi_{B_i}(\overline{\mathbf{x}}) \overline{\mathbf{x}} \\ &= \lambda(\overline{\mathbf{x}}) \overline{\mathbf{x}}. \end{aligned}$$

We claim that

$$P_{\overline{\mathbf{x}}} \Big( E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}}) I_n \Big) P_{\overline{\mathbf{x}}} = E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}}) I_n. \tag{2.10}$$

To prove (2.10), we note

$$
\begin{aligned}
&P_{\overline{\mathbf{x}}}\Big(E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}})I_n\Big)P_{\overline{\mathbf{x}}} \\
&= E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}})I_n - 2E(\overline{\mathbf{x}})\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top} + 2\textstyle\prod_{i=1}^{k}\phi_{B_i}(\overline{\mathbf{x}})\phi_D(\overline{\mathbf{x}})\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top} \\
&= E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}})I_n - 2\Big(E(\overline{\mathbf{x}})\overline{\mathbf{x}} - \lambda(\overline{\mathbf{x}})\overline{\mathbf{x}}\Big)\overline{\mathbf{x}}^{\top} \\
&= E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}})I_n,
\end{aligned}
$$

where we have used:

(a) $E(\overline{\mathbf{x}})\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top} = \overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}E(\overline{\mathbf{x}})$ by Theorem 2.1.1;

(b) $\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}E(\overline{\mathbf{x}})\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top} = \lambda(\overline{\mathbf{x}})\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}$.

Moreover,

$$
\begin{aligned}
&\frac{2\prod_{j=1,j\neq i}^{k}\phi_{B_j}(\overline{\mathbf{x}})}{\phi_{B_i}(\overline{\mathbf{x}})}\sum_{i=1}^{k}\left(2\frac{\phi_{A_i}(\overline{\mathbf{x}})}{\phi_{B_i}(\overline{\mathbf{x}})}(B_i\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}B_i) - (A_i\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}B_i + B_i\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}A_i)\right) \\
&= 2\sum_{i=1}^{k}\frac{\prod_{j=1,j\neq i}^{k}\phi_{B_j}(\overline{\mathbf{x}})}{\phi_{B_i}(\overline{\mathbf{x}})}\left(B_i\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}\Big(\frac{\phi_{A_i}(\overline{\mathbf{x}})}{\phi_{B_i}(\overline{\mathbf{x}})}B_i - A_i\Big) + \Big(\frac{\phi_{A_i}(\overline{\mathbf{x}})}{\phi_{B_i}(\overline{\mathbf{x}})}B_i - A_i\Big)\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}B_i\right) \\
&= 2\sum_{i=1}^{k}\prod_{j=1,j\neq i}^{k}\phi_{B_j}(\overline{\mathbf{x}})\Bigg[(B_i\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}D + D\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}B_i) - \phi_D(\overline{\mathbf{x}})(B_i\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top} + \overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}B_i) \\
&\quad + B_i\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}\sum_{\ell=1,\ell\neq i}^{k}\frac{1}{\phi_{B_\ell}(\overline{\mathbf{x}})}\Big(A_\ell - \frac{\phi_{A_\ell}(\overline{\mathbf{x}})}{\phi_{B_\ell}(\overline{\mathbf{x}})}B_\ell\Big) + \sum_{\ell=1,\ell\neq i}^{k}\frac{1}{\phi_{B_\ell}(\overline{\mathbf{x}})}\Big(A_\ell - \frac{\phi_{A_\ell}(\overline{\mathbf{x}})}{\phi_{B_\ell}(\overline{\mathbf{x}})}B_\ell\Big)\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}B_i\Bigg] \\
&= 2\sum_{i=1}^{k}\prod_{j=1,j\neq i}^{k}\phi_{B_j}(\overline{\mathbf{x}})\Bigg[(B_i\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}D + D\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}B_i) - \phi_D(\overline{\mathbf{x}})(B_i\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top} + \overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}B_i) \\
&\quad + \sum_{\ell=1,\ell\neq i}^{k}\frac{1}{\phi_{B_\ell}(\overline{\mathbf{x}})}\Big((B_i\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}A_\ell + A_\ell\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}B_i) - \frac{\phi_{A_\ell}(\overline{\mathbf{x}})}{\phi_{B_\ell}(\overline{\mathbf{x}})}(B_i\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}B_\ell + B_\ell\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}B_i)\Big)\Bigg].
\end{aligned}
\tag{2.11}
$$

Note that

$$
P_{\overline{\mathbf{x}}}(B_i\overline{\mathbf{x}}\overline{\mathbf{x}}^{\top} + \overline{\mathbf{x}}\overline{\mathbf{x}}^{\top}B_i)P_{\overline{\mathbf{x}}} = \mathbf{0}, \quad i = 1, \ldots, k. \tag{2.12}
$$

Therefore, from (2.10), (2.11) and (2.12), we have $H(\overline{\mathbf{x}}) = K(\overline{\mathbf{x}})$.

iii) The condition in (iii) of Theorem 2.2.1 is a sufficient condition for $\overline{\mathbf{x}}$ to be a strict local maximizer of (1.1).

■

There is a particular result appearing in the following theorem from Theorem 2.2.1 that is related to the eigenvalue of $E(\overline{\mathbf{x}})$ where $\overline{\mathbf{x}}$ is a local maximizer of (1.1), and the following theorem shows this result. We assume that $D$ is a positive definit matrix.

**Theorem 2.2.2.** *Let $A_i \in \mathbb{S}_n$ and $B_i, D \in \mathbb{S}_n^{++}$ for $i = 1, \ldots, k$. If $\overline{\mathbf{x}}$ is a local maximizer of (1.1), then it must be a unit eigenvector corresponding to either the largest or the second largest eigenvalue of $E(\overline{\mathbf{x}})$.*

*Proof.* Let

$$\mathbf{u} = 2P_{\overline{\mathbf{x}}} D\overline{\mathbf{x}},$$

and

$$\mathbf{v}_i = P_{\overline{\mathbf{x}}}\Bigg[B_i\overline{\mathbf{x}} + \frac{1}{2}\sum_{\ell=1,\ell\neq i}^{k} \frac{1}{\phi_{B_\ell}(\overline{\mathbf{x}})D\overline{\mathbf{x}}\overline{\mathbf{x}}^\top}\Bigg((B_i\overline{\mathbf{x}}\overline{\mathbf{x}}^\top A_\ell + A_\ell\overline{\mathbf{x}}\overline{\mathbf{x}}^\top B_i) - \frac{\phi_{A_\ell}(\overline{\mathbf{x}})}{\phi_{B_\ell}(\overline{\mathbf{x}})}(B_i\overline{\mathbf{x}}\overline{\mathbf{x}}^\top B_\ell + B_\ell\overline{\mathbf{x}}\overline{\mathbf{x}}^\top B_i)\Bigg)\overline{\mathbf{x}}\Bigg], \qquad i = 1, \ldots, k.$$

Rewrite $K(\overline{\mathbf{x}})$ in Theorem 2.2.1 to get

$$K(\overline{\mathbf{x}}) = E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}})I_n + \sum_{i=1}^{k}(\mathbf{u}\mathbf{v}_i^\top + \mathbf{v}_i\mathbf{u}^\top).$$

The rank of the matrix $W := \sum_{i=1}^{k}(\mathbf{u}\mathbf{v}_i^\top + \mathbf{v}_i\mathbf{u}^\top) \in \mathbb{S}_n$ is at most 2, and its only possible nonzero eigenvalues are

$$\lambda_{1,2} = \mathbf{u}^\top\sum_{i=1}^{k}\mathbf{v}_i \pm \|\mathbf{u}\|_2\left(\sum_{i=1}^{k}\|\mathbf{v}_i\|_2 + \sqrt{2}\sum_{i=1,i\neq j}^{k}\sqrt{\mathbf{v}_i\mathbf{v}_j}\right).$$

That implies

$$\lambda_1(W) \geq 0 = \lambda_2(W) = \cdots = \lambda_{n-1}(W) \geq \lambda_n(W).$$

As stated in $(ii)$ of Theorem 2.2.1, since $\overline{\mathbf{x}}$ is a local maximizer of (1.1) and $K(\overline{\mathbf{x}})$ is a negative semidefinite matrix, based on Weyl's monotonicity principle Theorem 1.2.2, we have

$$\lambda_1(K(\overline{\mathbf{x}})) + \lambda_m(-W) \geq \lambda_m(K(\overline{\mathbf{x}}) - W) \geq \lambda_n(K(\overline{\mathbf{x}})) + \lambda_m(-W), \quad 1 \leq m \leq n.$$

Then, for $m = 1$,

$\lambda_1(K(\overline{\mathbf{x}})) + \lambda_1(-W) \geq \lambda_1(K(\overline{\mathbf{x}}) - W)$

$\Rightarrow \lambda_1(-W) \geq \lambda_1(K(\overline{\mathbf{x}}) - W)$, because $\lambda_1(K(\overline{\mathbf{x}})) \leq 0$

$\Rightarrow \lambda_1(E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}})I_n) \leq \lambda_1(-W) = -\lambda_n(W)$.

Similarly, for $m = n$, we have

$\lambda_n(E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}})I_n) \leq -\lambda_1(W)$.

Hence, we have

$$\begin{cases} \lambda_1(E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}})I_n) \leq -\lambda_n(W), \\ \lambda_2(E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}})I_n) \leq -\lambda_{n-1}(W) = 0, \\ \qquad\qquad \vdots \\ \lambda_n(E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}})I_n) \leq -\lambda_1(W), \end{cases} \tag{2.13}$$

and note that

$$\lambda_2(E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}})I_n) \leq 0 \leq \lambda_1(E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}})I_n). \tag{2.14}$$

Since $(0, \overline{\mathbf{x}})$ is an eigenpair of $E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}})I_n$, the previous relation (2.14) leads to the following three possible cases:

1. $\lambda_2(E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}})I_n) < 0 = \lambda_1(E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}})I_n)$

   $\Rightarrow \lambda_1(E(\overline{\mathbf{x}})) = \lambda(\overline{\mathbf{x}}) : \lambda(\overline{\mathbf{x}})$ is the largest eigenvalue of $E(\overline{\mathbf{x}})$;

2. $\lambda_2(E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}})I_n) = 0 = \lambda_1(E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}})I_n)$

   $\Rightarrow \lambda_{1,2}(E(\overline{\mathbf{x}})) = \lambda(\overline{\mathbf{x}}) : \lambda(\overline{\mathbf{x}})$ is either the largest eigenvalue or the second largest eigenvalue of $E(\overline{\mathbf{x}})$;

3. $\lambda_2(E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}})I_n) = 0 < \lambda_1(E(\overline{\mathbf{x}}) - \lambda(\overline{\mathbf{x}})I_n)$

   $\Rightarrow \lambda_2(E(\overline{\mathbf{x}})) = \lambda(\overline{\mathbf{x}}) : \lambda(\overline{\mathbf{x}})$ is the second largest eigenvalue of $E(\overline{\mathbf{x}})$.

This completes the proof. ■

# Chapter 3

# Convergence Analysis for the Self-Consistent-Field (SCF) Iteration

## 3.1 SCF Iteration

Currently, the SCF is considered one of the most commonly used algorithms to solve the Kohn-Sham equations in electronic structure calculations (e.g., [8, 9, 12, 15, 16]). In [2, 7, 20], we find the convergence of the SCF iteration for solving the Kohn-Sham equations and numerical algorithms for electronic structure calculations.

In this section, we introduce a self-consistent-field (SCF) iterative method for solving (1.1). In spite of the fact that the solution $\mathbf{x}$ in $E(\mathbf{x})$ of (2.2) is unknown, one can approach the target solution $\mathbf{x}$ using an iterative scheme: a simple self-consistent-field iterative method (Algorithm 1).

There are several remarks regarding this algorithm:

i) Step 1 has the major computational cost of Algorithm 1, where it computes a dominant eigenvector of $E(\mathbf{x}^{(t)})$ in every iteration.

ii) In step 2, the term $r^{(t+1)}$ defines the residual for the approximate $\mathbf{x}^{(t)}$ of the nonlinear eigenvalue problem (2.2).

iii) Once the sequence $\{\mathbf{x}^{(t)}\}$ converges to $\hat{\mathbf{x}}$, $\hat{\mathbf{x}}$ satisfies the necessary conditions for local optimality in Theorem 2.2.2. This can be considered one of the major advantages of the SCF iteration over some optimization-based methods (see [1, 5]).

**Algorithm 1** A self-consistent-field (SCF) iteration

Given $\mathbf{x}^{(0)} \in \mathcal{M}$ and tolerance $\epsilon > 0$; set $t = 0$. This algorithm computes an approximate maximizer for the optimization problem (1.1).

1: Compute the dominant eigenvector $\mathbf{x}^{(t+1)}$ of $E(\mathbf{x}^{(t)})$;

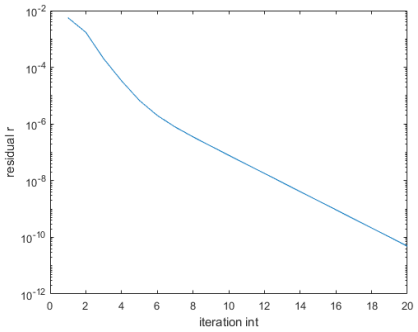
2: Compute the residual $r^{(t+1)} := \dfrac{\left\|E(\mathbf{x}^{(t+1)})\mathbf{x}^{(t+1)} - \lambda(\mathbf{x}^{(t+1)})\mathbf{x}^{(t+1)}\right\|_2}{\left(\left\|E(\mathbf{x}^{(t+1)})\right\|_2 + |\lambda(\mathbf{x}^{(t+1)})|\right)\left\|\mathbf{x}^{(t+1)}\right\|_2}$;

3: **if** $r^{(t+1)} \le \epsilon$ **then**

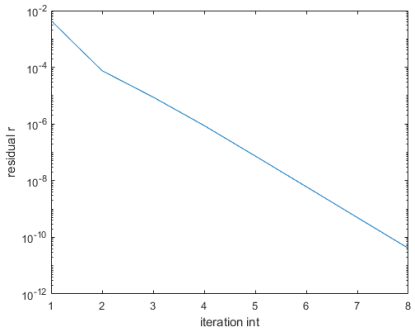
4: stop and return the approximation solution $\mathbf{x}^{(t+1)}$;

otherwise, set $t := t + 1$ and go to step 1.

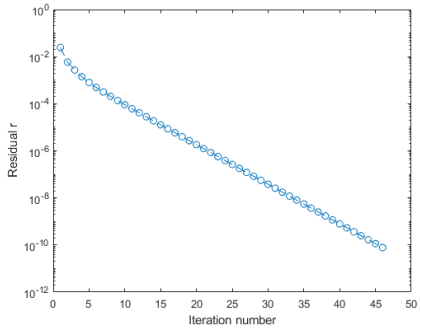
5: **end if**




**(a)** $n = 20, k = 4$

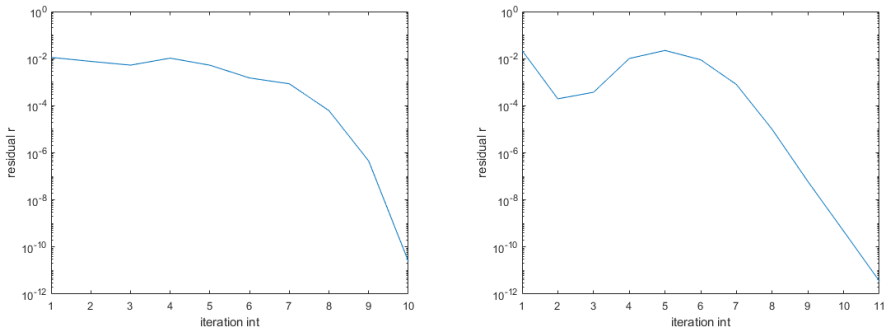



**(b)** $n = 50, k = 10$




**(c)** $n = 30, k = 30$

**Figure 3.1:** The residual $r^{(t)}$ from SCF with $n \ge k$




**(a)** $n = 10, k = 50$ **(b)** $n = 5, k = 10$

**Figure 3.2:** The residual $r^{(t)}$ from SCF with $n < k$

We have presented some results using the SCF iteration, Algorithm 1, for several matrices that are generated by MATLAB, using "randn($n$)" command to return an $n$-by-$n$ matrix of normally distributed random numbers. Then, in order to make this matrix to be symmetric, we use "$(G + G')/2$" for $G \in \{A_i, B_i, D\}$, where "$G'$" is the transpose of $G$ which is $G^\top$. After that, $A_i, B_i$ for $i = 1, \ldots, k$ are sorted using "cell" arry command which is a data type with indexed data containers called cells, where each cell can contain one of the generated matrices $A_i$, and $B_i$ for $i = 1, \ldots, k$. (see Figure 3.1 and Figure 3.2).

## 3.2 Local Convergence of the SCF Iteration

We analyze the local covergence behavior of the SCF iteration. In this section, we provide a condition to ensure the local convergence of $\{\mathbf{x}^{(t)}\}$. It is desirable to utilize the distance between subspaces, in order to measure the convergence rate of the SCF iteration [17]. Yang et al. in [15] have used this technique for analyzing the SCF iteration convergence. Particularly, for (1.1), we analyze the distance between span$(\mathbf{x}^{(t)})$ and span$(\hat{\mathbf{x}})$, where span$(\mathbf{u})$ denotes the one-dimentional subspace

spanned by $\mathbf{u}$, and $\hat{\mathbf{x}}$ is the dominant eigenvector of $E(\hat{\mathbf{x}})$. The distance between $\text{span}(\mathbf{x}^{(t)})$ and $\text{span}(\hat{\mathbf{x}})$ is defined by (see §2.6.3 in [6]):

$$d^{(t)} \coloneqq \text{dist}(\text{span}(\mathbf{x}^{(t)}), \text{span}(\hat{\mathbf{x}})) = \left\| \hat{\mathbf{x}}(\hat{\mathbf{x}})^\top - \mathbf{x}^{(t)}(\mathbf{x}^{(t)})^\top \right\|_2 = \sin \sigma^{(t)}, \tag{3.1}$$

where $\sigma^{(t)} \coloneqq \measuredangle(\text{span}(\mathbf{x}^{(t)}), \text{span}(\hat{\mathbf{x}}))$ is the angle between $\text{span}(\mathbf{x}^{(t)})$ and $\text{span}(\hat{\mathbf{x}})$, given by $\cos \sigma^{(t)} = |(\hat{\mathbf{x}})^\top \mathbf{x}^{(t)}|$.

**Lemma 3.2.1.** *Let* $\mathbf{x}, \hat{\mathbf{x}} \in \mathcal{M}$, $\sigma \coloneqq \measuredangle(\text{span}(\mathbf{x}), \text{span}(\hat{\mathbf{x}}))$, *and* $\eta = 2 \sin \sigma/2$. *The following statements hold:*

*i)* $\sin \sigma \leq \eta \leq \sqrt{2} \sin \sigma$.

*ii) There exists a scalar* $\alpha \in \mathbb{R}$ *with* $|\alpha| = 1$ *and a vector* $\mathbf{y} \perp \hat{\mathbf{x}}$ *with* $\|\mathbf{y}\|_2 = \sin \sigma$ *such that* $\alpha \mathbf{x} = \hat{\mathbf{x}} \cos \sigma + \mathbf{y}$. *As a consequence* $\|\alpha \mathbf{x} - \hat{\mathbf{x}}\|_2 = \eta$.

*Proof.* Since $0 \leq \sigma \leq \pi/2$, we have $\cos(\sigma/2) \geq 1/\sqrt{2}$ and thus

$$\sin \sigma \leq \eta \leq \sqrt{2} \sin \sigma.$$

This proves the first statement. For $(ii)$, let $\mathbf{y} = \alpha \mathbf{x} - \hat{\mathbf{x}} \cos \sigma$. We have

$$\hat{\mathbf{x}}^\top \mathbf{y} = \hat{\mathbf{x}}^\top (\alpha \mathbf{x} - \hat{\mathbf{x}} \cos \sigma) = \hat{\mathbf{x}}^\top (\alpha \mathbf{x}) - \hat{\mathbf{x}}^\top \hat{\mathbf{x}} \cos \sigma = 0,$$

i.e., $\mathbf{y} \perp \hat{\mathbf{x}}$. It follows from $\alpha \mathbf{x} = \hat{\mathbf{x}} \cos \sigma + \mathbf{y}$ and $\mathbf{y} \perp \hat{\mathbf{x}}$ that

$$1^2 = \|\alpha \mathbf{x}\|^2 = \cos^2 \sigma + \|\mathbf{y}\|_2 \quad \Rightarrow \quad \|\mathbf{y}\|_2 = \sin \sigma.$$

Finally,

$$\|\alpha \mathbf{x} - \hat{\mathbf{x}}\|_2^2 = (\cos \sigma - 1)^2 + \|\mathbf{y}\|_2^2 = 2(1 - \cos \sigma) = \left(2 \sin \frac{\sigma}{2}\right)^2,$$

yielding

$$\|\alpha \mathbf{x} - \hat{\mathbf{x}}\|_2 = 2 \sin \frac{\sigma}{2} = \frac{\sin \sigma}{\cos(\sigma/2)} = \eta,$$

as expected. ∎

We present the main result of this section in the following theorem.

**Theorem 3.2.2.** *Let $A_i, D \in \mathbb{S}_n$ and $B_i \in \mathbb{S}_n^{++}, i = 1, \ldots, k$. Suppose $\hat{\mathbf{x}}$ is the dominant eigenvector of $E(\hat{\mathbf{x}})$ and*

$$\delta := \lambda_1(E(\hat{\mathbf{x}})) - \lambda_2(E(\hat{\mathbf{x}})) > 0. \tag{3.2}$$

*If*

$$4\sqrt{2}\chi < \delta, \tag{3.3}$$

*then SCF iteration converges locally to $\hat{\mathbf{x}}$, where*

$$\chi := 2\left\{\sum_{i=1}^{k}\left[2\|A_i\|_2\|B_i^{-1}\|_2^2 + \frac{\|A_i\|_2}{\|B_i\|_2}(1 + \|B_i^{-1}\|_2)\right] + \|D\|_2\right\}\prod_{j=1}^{k}\|B_j\|_2.$$

*Proof.* First, we will drop the superscript of iteration index $t$ in order to simplify the presentation of the proof. We write $\mathbf{x}$ for $\mathbf{x}^{(t)}$, $\phi_{B_i}$ for $\phi_{B_i}(\mathbf{x}^{(t)})$ and $\mathbf{x}_+$ for $\mathbf{x}^{(t+1)}$, and so on.

We will show that $d_+ < \gamma d$ with $\gamma = \frac{4\sqrt{2}\chi}{\delta} < 1$, where $d_+ := \operatorname{dist}(\operatorname{span}(\mathbf{x}_+), \operatorname{span}(\hat{\mathbf{x}}))$ and $d := \operatorname{dist}(\operatorname{span}(\mathbf{x}), \operatorname{span}(\hat{\mathbf{x}}))$. For that, we estimate $\|\Delta E(\mathbf{x})\|_2 = \|E(\mathbf{x}) - E(\hat{\mathbf{x}})\|_2$. Let $\sigma = \angle(\operatorname{span}(\mathbf{x}), \operatorname{span}(\hat{\mathbf{x}}))$ and $\eta = \|\alpha\mathbf{x} - \hat{\mathbf{x}}\|_2$. By Lemma 3.2.1, we have $d \le \eta \le \sqrt{2}d$.

First, we estimate $|\phi_G(\mathbf{x}) - \phi_G(\hat{\mathbf{x}})|$ for $G \in \{A_i, B_i, D\}$. Let $\alpha\mathbf{x}$ be as in Lemma 3.2.1 and set $\tilde{\mathbf{x}} = \alpha\mathbf{x}$. It can be seen that $\phi_G(\mathbf{x}) = \phi_G(\tilde{\mathbf{x}})$. Thus

$$\phi_G(\mathbf{x}) - \phi_G(\hat{\mathbf{x}}) = \phi_G(\tilde{\mathbf{x}}) - \phi_G(\hat{\mathbf{x}}) = \tilde{\mathbf{x}}^\top G\tilde{\mathbf{x}} - \hat{\mathbf{x}}^\top G\hat{\mathbf{x}} = (\tilde{\mathbf{x}} - \hat{\mathbf{x}})^\top G\tilde{\mathbf{x}} + \hat{\mathbf{x}}^\top G(\tilde{\mathbf{x}} - \hat{\mathbf{x}}),$$

yielding

$$|\phi_G(\mathbf{x}) - \phi_G(\hat{\mathbf{x}})| \le 2\|G\|_2\eta. \tag{3.4}$$

Next, estimate $\left|\prod_{i=1}^{k}\phi_{B_i}(\mathbf{x}) - \prod_{i=1}^{k}\phi_{B_i}(\hat{\mathbf{x}})\right|$. Adopt the convention

$$\prod_{i=1}^{0}(\cdots) \equiv 1, \qquad \prod_{i=k+1}^{k}(\cdots) \equiv 1.$$

We have

$$\prod_{i=1}^{k}\phi_{B_i}(\mathbf{x})-\prod_{i=1}^{k}\phi_{B_i}(\hat{\mathbf{x}})=\sum_{j=1}^{k}\left(\prod_{i=1}^{j-1}\phi_{B_i}(\mathbf{x})\right)[\phi_{B_i}(\mathbf{x})-\phi_{B_i}(\hat{\mathbf{x}})]\left(\prod_{i=j+1}^{k}\phi_{B_i}(\hat{\mathbf{x}})\right),$$

yielding with (3.4)

$$\left|\prod_{i=1}^{k}\phi_{B_i}(\mathbf{x})-\prod_{i=1}^{k}\phi_{B_i}(\hat{\mathbf{x}})\right|\leq 2\eta\prod_{i=1}^{k}\|B_i\|_2.$$

For positive definite $B_i, i=1,\ldots,k$, and $\|\mathbf{x}\|_2=1$,

$$\|B_i^{-1}\|_2^{-1}\leq\phi_{B_i}(\mathbf{x})\leq\|B_i\|_2.$$

Also, we estimat $\left|\frac{\phi_{A_i}(\hat{\mathbf{x}})}{\phi_{B_i}(\hat{\mathbf{x}})}-\frac{\phi_{A_i}(\mathbf{x})}{\phi_{B_i}(\mathbf{x})}\right|$ using $\measuredangle(\mathrm{span}(\mathbf{x}),\mathrm{span}(\hat{\mathbf{x}}))$. We have

$$\begin{aligned}\frac{\phi_{A_i}(\hat{\mathbf{x}})}{\phi_{B_i}(\hat{\mathbf{x}})}-\frac{\phi_{A_i}(\mathbf{x})}{\phi_{B_i}(\mathbf{x})}&=\frac{\phi_{A_i}(\hat{\mathbf{x}})\phi_{B_i}(\mathbf{x})-\phi_{A_i}(\mathbf{x})\phi_{B_i}(\hat{\mathbf{x}})}{\phi_{B_i}(\hat{\mathbf{x}})\phi_{B_i}(\mathbf{x})}\\&=\frac{[\phi_{A_i}(\hat{\mathbf{x}})-\phi_{A_i}(\mathbf{x})]\phi_{B_i}(\mathbf{x})+\phi_{A_i}(\mathbf{x})[\phi_{B_i}(\mathbf{x})-\phi_{B_i}(\hat{\mathbf{x}})]}{\phi_{B_i}(\hat{\mathbf{x}})\phi_{B_i}(\mathbf{x})},\end{aligned}$$

yielding

$$\left|\frac{\phi_{A_i}(\hat{\mathbf{x}})}{\phi_{B_i}(\hat{\mathbf{x}})}-\frac{\phi_{A_i}(\mathbf{x})}{\phi_{B_i}(\mathbf{x})}\right|\leq 4\|A_i\|_2\|B_i\|_2\|B_i^{-1}\|_2^2\,\eta.$$

Now estimate $\|\Delta E(\mathbf{x})\|_2=\|E(\mathbf{x})-E(\hat{\mathbf{x}})\|_2$. We have

$$\begin{aligned}E(\mathbf{x})-E(\hat{\mathbf{x}})&=\sum_{i=1}^{k}\left[\left(A_i-\frac{\phi_{A_i}(\mathbf{x})}{\phi_{B_i}(\mathbf{x})}B_i\right)\prod_{\substack{j=1,\\ j\neq i}}^{k}\phi_{B_j}(\mathbf{x})-\left(A_i-\frac{\phi_{A_i}(\hat{\mathbf{x}})}{\phi_{B_i}(\hat{\mathbf{x}})}\right)\prod_{\substack{j=1,\\ j\neq i}}^{k}\phi_{B_j}(\hat{\mathbf{x}})\right]\\&\quad+\left[\prod_{i=1}^{k}\phi_{B_i}(\mathbf{x})-\prod_{i=1}^{k}\phi_{B_i}(\hat{\mathbf{x}})\right]D\\&=\sum_{i=1}^{k}\left[\left(\frac{\phi_{A_i}(\hat{\mathbf{x}})}{\phi_{B_i}(\hat{\mathbf{x}})}-\frac{\phi_{A_i}(\mathbf{x})}{\phi_{B_i}(\mathbf{x})}\right)\prod_{\substack{j=1,\\ j\neq i}}^{k}\phi_{B_j}(\mathbf{x})\right.\\&\qquad\left.+\left(A_i-\frac{\phi_{A_i}(\hat{\mathbf{x}})}{\phi_{B_i}(\hat{\mathbf{x}})}B_i\right)\left(\prod_{\substack{j=1,\\ j\neq i}}^{k}\phi_{B_j}(\mathbf{x})-\prod_{j=1,j\neq i}^{k}\phi_{B_j}(\hat{\mathbf{x}})\right)\right]\\&\quad+\left[\prod_{i=1}^{k}\phi_{B_i}(\mathbf{x})-\prod_{i=1}^{k}\phi_{B_i}(\hat{\mathbf{x}})\right]D,\end{aligned}$$

yielding

$$\begin{aligned}\|E(\mathbf{x}) - E(\hat{\mathbf{x}})\|_2 \le \sum_{i=1}^{k} \Bigg[ & 4\eta \|A_i\|_2 \|B_i^{-1}\|_2^2 \prod_{j=1}^{k} \|B_j\|_2 + 2\eta \|A_i\|_2 (1 + \|B_i^{-1}\|_2) \prod_{\substack{j=1,\\ j\neq i}}^{k} \|B_j\|_2 \Bigg] \\ & + 2\eta \|D\|_2 \prod_{\substack{j=1,\\ j\neq i}}^{k} \|B_j\|_2 \\ = \chi\eta.\end{aligned}$$

Finally, as stated in Theorem 8.1.10 and Corollary 8.1.11 (see Theorem 1.2.3), we have

$$d_+ := \operatorname{dist}(\operatorname{span}(\mathbf{x}_+), \operatorname{span}(\hat{\mathbf{x}})) \le \frac{4\,\|E(\mathbf{x}) - E(\hat{\mathbf{x}})\|_2}{\delta} \le \frac{4\chi\eta}{\delta} \le \frac{4\sqrt{2}\chi}{\delta}\, d.$$

Therefore, the local convergence is ensured.

■

### 3.3 Trust-Region Self-Consistent-Field (TRSCF) Iteration

The SCF iteration may not always converge, and sometimes the iterate $\mathbf{x}^{(t)}$ oscillates and the sequence $\{f^{(t)}\}$ is not monotonically increasing [17]. In electronic structure calculations, there is an observation about this convergence behavior of the SCF iteration. More explanation and analysis of this phenomenon can be found in [15,16]. According to [16], since calculating a dominant eigenvector $\mathbf{x}^{(t+1)}$ of $E(\mathbf{x}^{(t)})$ is equivalent to solving the following problem:

$$\max_{\mathbf{x}\in\mathcal{M}} \quad \mathbf{x}^{\top} E^{(t)} \mathbf{x}, \tag{3.5}$$

the SCF iteration is considered an iterative procedure that maximizes the related objective function of problem (1.1) by maximizing a sequence of the objective function in (3.5). Refer to [16] for more explanation and detailed discussion about this point.

In order to improve the convergence of the SCF iteration, there are several heuristics which are proposed in the material sciences and in quantum chemistry [12, 13, 16]. The trust-region SCF (TRSCF) iteration is one of them. The target of this method is to restrict the approximate solution of the problem (3.5) to a trust region such that the next iterate $\mathbf{x}^{(t+1)}$ is required to not be far away from $\mathbf{x}^{(t)}$. Since $\|\mathbf{x}\mathbf{x}^\top - \mathbf{x}^{(t)}(\mathbf{x}^{(t)})^\top\|_F$ measures the distance between $\text{span}(\mathbf{x}^{(t)})$ and $\text{span}(\mathbf{x})$ [6], with this in mind, the trust-region-based quadratic surrogate problem was designed as [17]

$$\max_{\substack{\|\mathbf{x}\|_2=1,\\ \|\mathbf{x}\mathbf{x}^\top-\mathbf{x}^{(t)}(\mathbf{x}^{(t)})^\top\|_F\le\Delta_t}} \mathbf{x}^\top E^{(t)}\mathbf{x}, \tag{3.6}$$

where $\Delta_t$ is the trust-region radius, and $\|\cdot\|_F$ is the Frobenius norm.

In [17], a technique of transforming the constraint $\|\mathbf{x}(\mathbf{x})^\top - \mathbf{x}^{(t)}(\mathbf{x}^{(t)})^\top\|_F \le \Delta_t$ into a penalty term, i.e.,

$$\max_{\mathbf{x}\in\mathcal{M}}\Big\{\mathbf{x}^\top E^{(t)}\mathbf{x} - \frac{\rho}{2}\left\|\mathbf{x}(\mathbf{x})^\top - \mathbf{x}^{(t)}(\mathbf{x}^{(t)})^\top\right\|_F^2\Big\}, \tag{3.7}$$

is used to solve (3.6), where $\rho > 0$ is a penalty parameter.

Since $\mathbf{x}, \mathbf{x}^{(t)} \in \mathcal{M}$, (3.7) is equivalent to the problem

$$\max_{\mathbf{x}\in\mathcal{M}}\Big\{\mathbf{x}^\top[E^{(t)} + \rho\mathbf{x}^{(t)}(\mathbf{x}^{(t)})^\top]\mathbf{x}\Big\}, \tag{3.8}$$

whose maximizer is the unit dominant eigenvector of $E^{(t)} + \rho\mathbf{x}^{(t)}(\mathbf{x}^{(t)})^\top$. The choice of penalty parameter, $\rho$, was discussed in Theorem 4.1 of [17]. The monotonicity of $\{f^{(t)}\}$ is controlled by the parameter, $\rho$. Also, this parameter plays an important role in adjusting the convergence speed, as well. We present the TRSCF iterative method in Algorithm 2, and compare the results of the sequences $\{f^{(t)}\}$ using the SCF and TRSCF iterative methods, in Figure 3.4. Actually, we can observe that the sequence $\{f^{(t)}\}$ converges using the TRSCF iteration while the sequence is oscillating for the SCF iteration, where we used a tolerance $\epsilon = 10^{-7}$.

**Algorithm 2** The TRSCF iteration.

Choose tolerance $\epsilon$, and $\mathbf{x}^{(0)} \in \mathcal{M}$; set $t = 0, \rho = 0$ and $E_\rho^{(t)} = E(\mathbf{x}^{(t)})$.

1: **while** $r^{(t)} := \dfrac{\left\|E(\mathbf{x}^{(t)})\mathbf{x}^{(t)} - \lambda(\mathbf{x}^{(t)})\mathbf{x}^{(t)}\right\|_2}{\left(\left\|E(\mathbf{x}^{(t)})\right\|_2 + |\lambda(\mathbf{x}^{(t)})|\right)\left\|\mathbf{x}^{(t)}\right\|_2} > \epsilon,$ **do**

2: Compute the dominant eigenvector $\mathbf{x}^{(t+1)}$ of $E_\rho^{(t)}$;

3: **if** $f(\mathbf{x}^{(t)}) > f(\mathbf{x}^{(t+1)})$ **then**

4: $\rho = 2(\lambda_1(E_\rho^{(t)}) - \lambda_2(E_\rho^{(t)}))$; $E_\rho^{(t+1)} := E(\mathbf{x}^{(t+1)}) + \rho\mathbf{x}^{(t+1)}(\mathbf{x}^{(t+1)})^\top$;

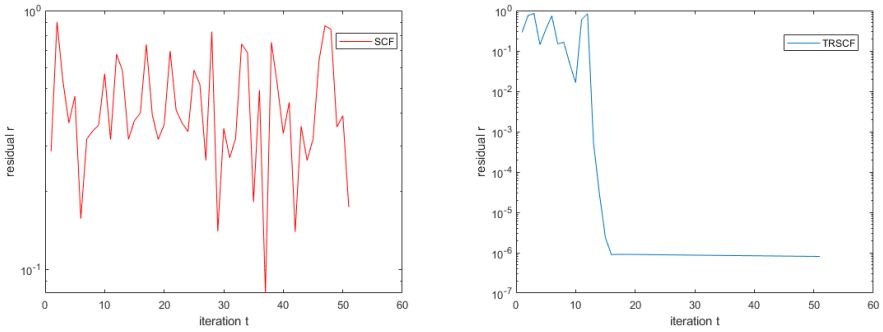
5: **else**

6: $E_\rho^{(t+1)} := E(\mathbf{x}^{(t+1)})$;

7: **end if**

8: $t = t + 1$;

9: **end while**




**(a)** SCF **(b)** TRSCF

**Figure 3.3:** The residual $r^{(t)}$ from SCF and TRSCF.

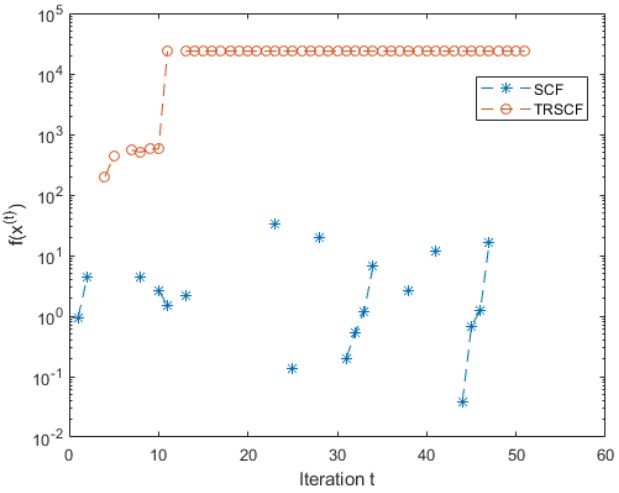



**Figure 3.4:** The sequences $\{f^{(t)}\}$ from SCF and TRSCF.

# Chapter 4

## On Optimizing the Sum of the Trace Ratios on the Stiefel Manifold

In this chapter, we are concerned with the solution of the optimization problem

$$\max_{V^\top V = I_\ell} f(V) \quad \text{with} \quad f(V) := \sum_{i=1}^{k} \frac{\operatorname{tr}(V^\top A_i V)}{\operatorname{tr}(V^\top B_i V)} + \operatorname{tr}(V^\top D V) \tag{4.1}$$

over the Stiefel manifold

$$\mathbb{O}^{n\times\ell} := \{V \in \mathbb{R}^{n\times\ell} \quad | \quad V^\top V = I_\ell\},$$

where $\ell < n$, and $A_i, B_i, D \in \mathbb{R}^{n\times n}$ are real symmetric with $B_i$ positive definite for $i = 1, \ldots, k$.

The problem (4.1) for the case $k = 1$ was investigated in [19], in which many properties for $\ell = 1$ were extended to $\ell > 1$. In the previous sections, we studied the case $\ell = 1$.

We simplify the presentation by using the following notations:

$$\phi_G(V) := \operatorname{tr}(V^\top G V), \quad G \in \{A_i, B_i, D\}, \quad V \in \mathbb{R}^{n\times\ell}.$$

Then, the function in (4.1) can be written as

$$f(V) := \sum_{i=1}^{k} \frac{\phi_{A_i}(V)}{\phi_{B_i}(V)} + \phi_D(V).$$

In fact, maximizing $f(V)$ over $\mathbb{O}^{n\times\ell}$ is more complicated than maximizing each term individually. For the term, $\operatorname{tr}(V^\top D V)$, we have a classical result for

$$\max_{V\in\mathbb{O}^{n\times\ell}} \operatorname{tr}(V^\top D V). \tag{4.2}$$

Any solution $V$ of the problem (4.2) is an orthonormal eigenbasis matrix associated with the $\ell$ largest eigenvalues of $D$. Also, there is no local but non-global maximizer. Generally, it is difficult to establish necessary and sufficient conditions for global maximizers of an optimization problem, and the problem (4.1) is no exception.

Since $\mathbb{O}^{n\times\ell}$ can be viewed as an embeded submanifold of the Euclidean space $\mathbb{R}^{n\times\ell}$ [1, 3, 5], the tangent space $\mathscr{T}_V\mathbb{O}^{n\times\ell}$ at $V \in \mathbb{O}^{n\times\ell}$ is given by (see [1, 3])

$$\begin{aligned}\mathscr{T}_V\mathbb{O}^{n\times\ell} :=& \{X \in \mathbb{R}^{n\times\ell} : \quad X^\top V + V^\top X = 0\} \\ =& \{X = VK + (I_n - VV^\top)J : \quad K = -K^\top \in \mathbb{R}^{\ell\times\ell}, J \in \mathbb{R}^{n\times\ell}\}.\end{aligned} \tag{4.3}$$

On $\mathscr{T}_V\mathbb{O}^{n\times\ell}$ , the standard inner product is

$$\langle X, Y\rangle = \mathrm{tr}(X^\top Y), \quad \text{for} \quad X, Y \in \mathscr{T}_V\mathbb{O}^{n\times\ell}.$$

The orthogonal projection of $Z \in \mathbb{R}^{n\times\ell}$ onto the tangent space $\mathscr{T}_V\mathbb{O}^{n\times\ell}$ is

$$\begin{aligned}\Pi_{\mathscr{T}}(Z) :=& V\Big(\frac{V^\top Z - Z^\top V}{2}\Big) + (I_n - VV^\top)Z \\ =& Z - V\frac{V^\top Z + Z^\top V}{2} = Z - V\mathrm{sym}(V^\top Z) \in \mathscr{T}_V\mathbb{O}^{n\times\ell},\end{aligned} \tag{4.4}$$

where $\mathrm{sym}(Z) := \frac{1}{2}(Z^\top + Z)$ is the symmetric part of $Z$.

## 4.1 First and Second Order Optimality Conditions

The first order optimality condition is given in the following theorem.

**Theorem 4.1.1.** *If $V \in \mathbb{O}^{n\times\ell}$ is a local maximizer of* (4.1)*, then*

$$E(V)V = VW_V, \tag{4.5}$$

*where*

$$E(V) := \sum_{i=1}^{k}\Big[\frac{A_i}{\phi_{B_i}(V)} - \frac{\phi_{A_i}(V)B_i}{(\phi_{B_i}(V))^2}\Big] + D,$$

*and*

$$W_V := V^\top E(V)V.$$

*Thus,* $\operatorname{eig}(W_V) \subset \operatorname{eig}(E(V))$*, and* $V$ *is an orthogonal eigenbasis matrix of* $E(V)$ *corresponding to its eigenvalues in* $\operatorname{eig}(W_V)$.

*Proof.* Using the projection in (4.4), we find the gradient of $f_{|\mathbb{O}^{n\times\ell}}(V)$. We have

$$\frac{\partial f(V)}{\partial V} = 2\Big[\sum_{i=1}^{k} \frac{A_i}{\phi_{B_i}(V)} - \frac{\phi_{A_i}(V)B_i}{(\phi_{B_i}(V))^2} + D\Big]V = 2E(V)V,$$

and thus

$$\begin{aligned}\operatorname{grad} f_{|\mathbb{O}^{n\times\ell}}(V) =& \Pi_{\mathscr{T}}\Big(\frac{\partial f(V)}{\partial V}\Big)\\ =& 2E(V)V - V\frac{V^\top(2E(V)V) + (2E(V)V)^\top V}{2}\\ =& 2\Big[E(V)V - V(V^\top E(V)V)\Big].\end{aligned}$$

Therefore,

$$\operatorname{grad} f_{|\mathbb{O}^{n\times\ell}}(V) = \mathbf{0} \Leftrightarrow E(V)V = VW_V,$$

where $W_V \coloneqq V^\top E(V)V$. ∎

**Theorem 4.1.2.** *If* $V$ *is a local maximizer of* (4.1)*, then*

$$\operatorname{tr}(X^\top E(V)X) - \operatorname{tr}(XW_VX^\top) + \operatorname{tr}(X^\top H(V,X)V) \le 0 \quad \textit{for} \quad X \in \mathscr{T}_V\mathbb{O}^{n\times\ell}, \tag{4.6}$$

*where*

$$H(V,X) \coloneqq 2\sum_{i=1}^{k}\Big\{2\frac{\operatorname{tr}(V^\top A_iV)\operatorname{tr}(X^\top B_iV)B_i}{[\operatorname{tr}(V^\top B_iV)]^3} - \frac{\operatorname{tr}(X^\top B_iV)A_i + \operatorname{tr}(X^\top A_iV)B_i}{[\operatorname{tr}(V^\top B_iV)]^2}\Big\}.$$

*If* $V \in \mathbb{O}^{n\times\ell}$ *satisfies* (4.5)*, and* (4.6) *is a strict inequality for* $X \neq 0$*, then* $V$ *is a strict local maximizer.*

*Proof.* By calculating $\mathbf{D}(\mathbf{D}f(V))[X]$:

$$\begin{aligned}\mathbf{D}(\mathbf{D}f(V))[X] =& 2\mathbf{D}(E(V)V)[X]\\ =& 2E(V)X + 2\mathbf{D}E(V)[X]V,\end{aligned}$$

we get

$$\begin{aligned}\mathbf{D}E(V)[X] &= \sum_{i=1}^{k} -2\frac{A_i \operatorname{tr}(X^\top B_i V)}{[\operatorname{tr}(V^\top B_i V)]^2} - 2\frac{B_i \operatorname{tr}(X^\top A_i V)[\operatorname{tr}(V^\top B_i V)]^2}{[\operatorname{tr}(V^\top B_i V)]^4} \\ &\quad + 2\frac{2B_i \operatorname{tr}(V^\top A_i V) 2\operatorname{tr}(V^\top B_i V)\operatorname{tr}(X^\top B_i V)}{[\operatorname{tr}(V^\top B_i V)]^4} \\ &= 2\sum_{i=1}^{k} \frac{2\operatorname{tr}(V^\top A_i V)\operatorname{tr}(X^\top B_i V)B_i}{[\operatorname{tr}(V^\top B_i V)]^3} - \frac{\operatorname{tr}(X^\top B_i V)A_i + \operatorname{tr}(X^\top A_i V)B_i}{[\operatorname{tr}(V^{top} B_i V)]^2} \\ &=: H(V, X).\end{aligned}$$

Thus,

$$\mathbf{D}(\mathbf{D}f(V))[X] = 2E(V)X + 2H(V, X)V.$$

Furthermore, if $V$ is a critical point, then $V^\top \frac{\partial f(V)}{\partial V} = 2W_V$ and by the necessary second order optimality condition in [14], we have

$$2\operatorname{tr}(X^\top E(V)X) + 2\operatorname{tr}(X^\top H(V, X)V) - 2\operatorname{tr}(W_V X^\top X) \le 0 \quad \text{for} \quad X \in \mathscr{T}_V \mathbb{O}^{n\times\ell},$$

which is (4.6). Therefore, the sufficient condition follows. ∎

Next, in the following theorem, we present the second order optimality condition in terms of $J \in \mathbb{R}^{n\times\ell}$. For more general case, in Theorem 4.1.2, the second order optimality condition has presented in terms of the tangent vector $X \in \mathscr{T}_V \in \mathbb{O}^{n\times\ell}$.

**Theorem 4.1.3.** *If $V$ is a local maximizer of* (1.6)*, then*

$$\begin{aligned}&\operatorname{tr}(J^\top E(V)J) + \operatorname{tr}(V^\top J W_V J^\top V) - \operatorname{tr}(J^\top V W_V V^\top J) - \operatorname{tr}(J W_V J^\top) \\ &\qquad + 4\sum_{i=1}^{k} \frac{\operatorname{tr}(J^\top[I_n - VV^\top]B_i V)}{[\phi_{B_i}(V)]^3}\Big(\phi_{A_i}(V)\operatorname{tr}(J^\top[I_n - VV^\top]B_i V) \\ &\qquad - \phi_{B_i}(V)\operatorname{tr}(J^\top[I_n - VV^\top]A_i V)\Big) \le 0 \quad \textit{for} \quad J \in \mathbb{R}^{n\times\ell}.\end{aligned} \tag{4.7}$$

*If $V \in \mathbb{O}^{n\times\ell}$ satisfies* (4.5)*, and if* (4.7) *is strict for $0 \neq J \in \mathbb{R}^{n\times\ell}$, then $V$ is a strict local maximizer.*

*Proof.* Any element of the tangent space $\mathscr{T}_V\mathbb{O}^{n\times\ell}$ at $V \in \mathbb{O}^{n\times\ell}$, by (4.3), can be expressed as [19]

$$X = VK + (I_n - VV^\top)J \in \mathscr{T}_V\mathbb{O}^{n\times\ell}$$

for any$J \in \mathbb{R}^{n\times\ell}$ and skew-symmetric matrix $K$. Using (4.5), we have

$$E(V)X = VW_VK + E(V)J - VW_VV^\top J. \tag{4.8}$$

Also, we have

$$\begin{aligned}
X^\top E(V)X =& \Big(K^\top V^\top + J^\top(I_nVV^\top)\Big)E(V)X \\
=& K^\top V^\top E(V)X + J^\top(I_nVV^\top)E(V)X \\
=& K^\top V^\top(VW_VK) + K^\top V^\top E(V)J - K^\top V^\top - K^\top V^\top VW_VV^\top J \\
&+ J^\top(I_n - VV^\top)\Big(VW_VK + E(V)J - VW_VV^\top J\Big) \\
=& K^\top W_VK + K^\top V^\top E(V)J - K^\top W_VV^\top J + J^\top(I_nVV^\top)E(V)J \\
=& K^\top W_VK + K^\top W_VV^\top J - K^\top W_VV^\top J + J^\top E(V)J - J^\top VW_VV^\top J \\
=& K^\top W_VK + J^\top E(V)J - J^\top VW_VV^\top J.
\end{aligned}$$

Then

$$\operatorname{tr}(X^\top E(V)X) = \operatorname{tr}(K^\top W_VK) + \operatorname{tr}(J^\top E(V)J) - \operatorname{tr}(J^\top VW_VV^\top J). \tag{4.9}$$

Since $X^\top X = K^\top K + J^\top(I_n - VV^\top J)$, by using $K^\top = -K$ we have

$$\operatorname{tr}(X^\top XM_V) = \operatorname{tr}(K^\top M_VK) + \operatorname{tr}(J^\top JM_V) - \operatorname{tr}(V^\top JW_VJ^\top V). \tag{4.10}$$

Now, in order to write each term of (4.6) in terms of $J \in \mathbb{R}^{n\times\ell}$ instead of $X \in \mathscr{T}_V\mathbb{O}^{n\times\ell}$, we also need to compute $\operatorname{tr}(X^\top H(V,X)V)$ :

$$\begin{aligned}
X^\top H(V,X)V =& 4\sum_{i=1}^{k}\frac{\phi_{A_i}(V)\operatorname{tr}(X^\top B_iV)^2}{[\phi_{B_i}(V)]^3} - \frac{\operatorname{tr}(X^\top A_iV)\operatorname{tr}(X^\top B_iV)}{[\phi_{B_i}(V)]^2} \\
=& 4\sum_{i=1}^{k}\frac{\operatorname{tr}(X^\top B_iV)}{[\phi_{B_i}(V)]^3}\Big(\phi_{A_i}(V)\operatorname{tr}(X^\top B_iV) - \phi_{B_i}(V)\operatorname{tr}(X^\top A_iV)\Big).
\end{aligned}$$

Note that

$$\mathrm{tr}(X^\top A_i V) = \langle X, A_i V\rangle = \langle VK + (I_n - VV^\top)J, A_i V\rangle = \mathrm{tr}(J^\top[I_n - VV^\top]A_i V),$$
$$\mathrm{tr}(X^\top B_i V) = \langle X, B_i V\rangle = \langle VK + (I_n - VV^\top)J, B_i V\rangle = \mathrm{tr}(J^\top[I_n - VV^\top]B_i V).$$

Then,

$$\begin{aligned}\mathrm{tr}(X^\top H(V,X)V) =& 4\sum_{i=1}^{k}\frac{\mathrm{tr}(J^\top[I_n - VV^\top]B_i V)}{[\phi_{B_i}(V)]^3}\\ &\Big(\phi_{A_i}(V)\,\mathrm{tr}(J^\top[I_n - VV^\top]B_i V) - \phi_{B_i}(V)\,\mathrm{tr}(J^\top[I_n - VV^\top]A_i V)\Big).\end{aligned} \tag{4.11}$$

From (4.9), (4.10) and (4.11) with (4.6), we have

$$\begin{aligned}&\mathrm{tr}(X^\top E(V)X) - \mathrm{tr}(XM_V X^\top) + \mathrm{tr}(X^\top H(V,X)V) \le 0,\\ &\Rightarrow \mathrm{tr}(J^\top E(V)J) - \mathrm{tr}(J^\top V W_V V^\top J) - tr(J^\top J M_V) + \mathrm{tr}(V^\top J W_V J^\top V)\\ &\quad + 4\sum_{i=1}^{k}\frac{\mathrm{tr}(J^\top[I_n - VV^\top]B_i V)}{[\phi_{B_i}(V)]^3}\Big(\phi_{A_i}(V)\,\mathrm{tr}(J^\top[I_n - VV^\top]B_i V)\\ &\quad - \phi_{B_i}(V)\,\mathrm{tr}(J^\top[I_n - VV^\top]A_i V)\Big) \le 0,\end{aligned}$$

which leads to (4.7). ■

## 4.2 A Necessary Condition for Local Maximizers

Suppose $V$ is a local maximizer of (4.1). We will establish a necessary condition for a local maximizer of (4.1) in this section. According to Theorem 4.1, we have that

$$\mathrm{eig}(W_V) \subset \mathrm{eig}(E(V)),$$

and thus

$$\mathrm{eig}(W_V) = \{\lambda_{\omega_j}(E(V)), \quad j = 1, 2, \ldots, \ell\}, \tag{4.12}$$

where $1 \le \omega_1 < \cdots < \omega_\ell \le n$.

**Theorem 4.2.1.** *Let $V \in \mathbb{O}^{n\times\ell}$ be a local maximizer of* (4.1)*, and set* $\mathrm{eig}(W_V)$ *as in* (4.12)*. Then*

$$\lambda_{\omega_1}(E(V)) \geq \lambda_{2\ell}(E(V)), \tag{4.13}$$

*provided* $2\ell \leq n$.

*Proof.* We will proof this theorem using the contradiction [19]. So, suppose that

$$\lambda_{\omega_1}(E(V)) < \lambda_{2\ell}(E(V)). \tag{4.14}$$

Let $Y \in \mathbb{R}^{n\times 2\ell}$ be an orthonormal eigenbasis matrix of $E(V)$ associated with its $2\ell$ largest eigenvalues $\lambda_j(E(V))$ for $1 \leq j \leq 2\ell$. Since $V$ is the orthonormal eigenbasis of $E(V)$ associated with its eigenvalues $\lambda_{\omega_j}(E(V))$ for $1 \leq j \leq \ell$, we have $Y^\top V = 0$. Also, let $J_i = YQ_i \in \mathbb{R}^{n\times\ell}$ where $Q_i \in \mathbb{R}^{2\ell\times\ell}$ for $i = 1, \ldots, k$, with orthonormal columns are to be chosen. Then

$$J_i^\top V = Q_i^\top Y^\top V = 0 \quad \text{and} \quad J_i^\top J_i = I_\ell.$$

Therefore, for $i = 1, \ldots, k$,

$$\mathrm{tr}(J_i^\top E(V) J_i) > \mathrm{tr}(V^\top E(V) V) = \mathrm{tr}(W_V) = \mathrm{tr}(J_i^\top J_i W_V) = \mathrm{tr}(J_i W_V J_i^\top). \tag{4.15}$$

Hence, by (4.7), we have

$$\begin{aligned} \mathrm{tr}(J_i^\top E(V) J_i) - \mathrm{tr}(W_V) + 4\sum_{i=1}^{k} \frac{\mathrm{tr}(J_i^\top B_i V)}{[\phi_{B_i}(V)]^3} \Big( \phi_{A_i}(V)\,\mathrm{tr}(J_i^\top B_i V) \\ - \phi_{B_i}(V)\,\mathrm{tr}(J_i^\top A_i V) \Big) \leq 0, \quad i = 1, \ldots, k. \end{aligned} \tag{4.16}$$

We can write that

$$J_i^\top B_i V = Q_i^\top (Y^\top B_i V) = Q_i^\top (U_i \Sigma_i W_i^\top), \quad i = 1, \ldots, k,$$

such that $U_i\Sigma_i W_i^\top$ is the SVD of $Y^\top B_i V$ and $U_i \in \mathbb{R}^{2\ell\times\ell}, \Sigma_i \in \mathbb{R}^{\ell\times\ell}$, and $W_i \in \mathbb{R}^{\ell\times\ell}$. We choose $Q_i$ to be the one makes $[U_i, Q_i] \in \mathbb{R}^{2\ell\times 2\ell}$ orthogonal [19]. Then,

$$J_i^\top B_i V = Q_i^\top (Y^\top B_i V) = Q_i^\top U_i \Sigma_i W_i^\top = 0, \quad i = 1, \ldots, k.$$

Therefore, from (4.16), we have

$$\operatorname{tr}(J_i^\top E(V) J_i) - \operatorname{tr}(W_V) \leq 0 \quad \text{for} \quad i = 1, \ldots, k,$$

which contradicts (4.15). Thus, (4.13) holds. ■

References


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

[2] Y. Cai, L.-H. Zhang, Z. Bai, and R.-C. Li, *On an eigenvector-dependent nonlinear eigenvalue problem*, SIAM Journal on Matrix Analysis and Applications **39** (2018), no. 3, 1360–1382.

[3] M. T. Chu and N. T. Trendafilov, *The orthogonally constrained regression revisited*, Journal of Computational and Graphical Statistics **10** (2001), no. 4, 746–771.

[4] J. W. Demmel, *Applied Numerical Linear Algebra*, Vol. 56, SIAM, 1997.

[5] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM journal on Matrix Analysis and Applications **20** (1998), no. 2, 303–353.

[6] G. Gloub and C. Van Loan, *Matrix Computations*, Johns Hopkins Universtiy Press, 3rd edtion, 1996.

[7] 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 Journal on Matrix Analysis and Applications **35** (2014), no. 2, 546–558.

[8] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods*, Cambridge university press, 2004.

[9] N. S. Ostlund and A. Szabo, *Modern Quantum Chemistry:Introduction to Advanced Electronic Structure Theory*, Dover Publications Inc, 1996.

[10] G. Primolevo, O. Simeone, and U. Spagnolini, *Towards a joint optimization of scheduling and beamforning for MIMO downlink*, 2006 IEEE Ninth International Symposium on Spread Spectrum Techniques and Applications, 2006, pp. 493–497.

[11] Y. Saad, *Numerical Methods for Large Eigenvalue Problem*, SIAM, 1992.

[12] V. Saunders and I. Hillier, *A " level-shifting" method for converging closed shell Hartree-Fock wave functions*, International Journal of Quantum Chemistry **7** (1973), no. 4, 699–705.

[13] 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*, The Journal of Chemical Physics **121** (2004), no. 1, 16–27.

[14] Z. Wen and W. Yin, *A feasible method for optimization with orthogonality constraints*, Mathematical Programming **142** (2013), no. 1-2, 397–434.

[15] C. Yang, W. Gao, and J. Meza, *On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems*, SIAM Journal on Matrix Analysis and Applications **30** (2009), no. 4, 1773–1788.

[16] C. Yang, J. Meza, and L.-W. Wang, *A trust region direct constrained minimization algorithm for the Kohn–Sham equation*, SIAM Journal on Scientific Computing **29** (2007), no. 5, 1854–1875.

[17] L.-H. Zhang, *On a self-consistent-field-like iteration for maximizing the sum of the Rayleigh quotients*, Journal of Computational and Applied Mathematics **257** (2014), 14–28.

[18] L.-H. Zhang, *On optimizing the sum of the Rayleigh quotient and the generalized Rayleigh quotient on the unit sphere*, Computational Optimization and Applications **54** (2013), no. 1, 111–139.

[19] L.-H. Zhang and R.-C. Li, *Maximization of the sum of the trace ratio on the Stiefel manifold, I: Theory*, Science China Mathematics **57** (2014), no. 12, 2495–2508.

[20] X. Zhang, J. Zhu, Z. Wen, and A. Zhou, *Gradient type optimization methods for electronic structure calculations*, SIAM Journal on Scientific Computing **36** (2014), no. 3, C265–C289.

## Biographical Statement

Aohud Abdulrahman Binbuhaer was born in Riyadh, Saudi Arabia on May 1, 1984. Aohud attended a public school and received her diploma in 2002.

Aohud enrolled in Princess Nourah Bint Abdulrahman University, PNU, in 2002 and received a Bachelor's degree in Mathematics in 2006. Then, Aohud received a Master's degree from Texas A&M University-Commerce in August 2013. She began her Ph.D. program at the University of Texas at Arlington in August 2013. In May 2019, she was awarded a Ph.D. in Mathematics under the direction of Dr. Ren-Cang Li.