SCIENCE CHINA
Mathematics

• ARTICLES •

July 2015 Vol. 58 No. 7: 1549–1566
doi: 10.1007/s11425-014-4825-z

Maximization of the sum of the trace ratio on the Stiefel manifold, II: Computation

ZHANG LeiHong[1] & LI RenCang[2,*]

[1]*Department of Applied Mathematics, Shanghai University of Finance and Economics, Shanghai* 200433*, China;*
[2]*Department of Mathematics, University of Texas at Arlington, Arlington, TX* 76019-0408*, USA*

*Email: longzlh@*163*.com, rcli@uta.edu*

Received July 30, 2013; accepted April 5, 2014; published online April 25, 2014

Abstract The necessary condition established in Part I of this paper for the global maximizers of the maximization problem

$$\max_{V}\left\{\frac{\mathrm{tr}(V^\top AV)}{\mathrm{tr}(V^\top BV)}+\mathrm{tr}(V^\top CV)\right\}$$

over the Stiefel manifold $\{V\in\mathbb{R}^{m\times\ell}\mid V^\top V=I_\ell\}$ ($\ell<m$), naturally leads to a self-consistent-field (SCF) iteration for computing a maximizer. In this part, we analyze the global and local convergence of the SCF iteration, and show that the necessary condition for the global maximizers is fulfilled at any convergent point of the sequences of approximations generated by the SCF iteration. This is one of the advantages of the SCF iteration over optimization-based methods. Preliminary numerical tests are reported and show that the SCF iteration is very efficient by comparing with some manifold-based optimization methods.

Keywords trace ratio, Rayleigh quotient, Stiefel manifold, nonlinear eigenvalue problem, optimality condition, self-consistent-field iteration, eigenspace

MSC(2010) 65F15, 65F30, 62H30, 15A18

Citation: Zhang L H, Li R C. Maximization of the sum of the trace ratio on the Stiefel manifold, II: Computation. Sci China Math, 2015, 58: 1549–1566, doi: 10.1007/s11425-014-4825-z

1 Introduction

This is the second paper of ours in the sequel. Building upon the theoretical results in [32], here we will focus on the numerical aspect of the maximization problem:

$$\max_{V^\top V=I_\ell}\left\{\frac{\mathrm{tr}(V^\top AV)}{\mathrm{tr}(V^\top BV)}+\mathrm{tr}(V^\top CV)\right\},\tag{1.1}$$

where $\mathrm{tr}(\,\cdot\,)$ stands for the trace of a square matrix, $A,B,C\in\mathbb{R}^{m\times m}$ are real symmetric with B positive definite, and integer $\ell<m$.

In [32], we showed that any critical point (i.e., KKT point) V of (1.1) is a solution to a nonlinear eigenvalue problem

$$E(V)V=V[V^\top E(V)V],\tag{1.2}$$

*Corresponding author

where $\phi_H(V) := \mathrm{tr}(V^\top HV)$ for $H \in \{A, B\}$, and

$$E(V) := A\frac{1}{\phi_B(V)} - B\frac{\phi_A(V)}{[\phi_B(V)]^2} + C \in \mathbb{R}^{m\times m} \tag{1.3}$$

is symmetric and is dependent on V. Furthermore, if V is a global maximizer, then it must be an orthonormal eigenbasis of $E(V)$ corresponding to its ℓ largest eigenvalues. This together with (1.2) lends themselves to a self-consistent-field (SCF) iteration for computing a global maximizer. The main purpose of this part is to analyze the convergence behavior of the SCF iteration. We will prove a global convergence result for the special case $C = \eta B$ ($\eta > 0$) and establish locally linear and quadratic convergence for the general case. We note that there is no guarantee that the SCF iteration will deliver a global maximizer at convergence but a KKT point that satisfies the necessary condition for the global maximizers. Despite so, this turns out to be an advantage of the SCF iteration over some optimization-based methods [1, 14] which in general only converge to a KKT point that may or may not satisfy the necessary condition for the global maximizers. Numerical tests are presented and they show that the SCF iteration is very efficient, comparing with some manifold-based optimization methods [2, 5], in both accuracy and running time.

The rest of this paper is organized as follows. Section 2 collects some preliminaries that are needed in the convergence analysis of the SCF iteration to be given in Section 3, where, through an example, we argue that the global convergence in general is not necessarily given. For the special case $C = \eta B$ for some $\eta \geqslant 0$, however, we establish a global convergence result in Section 4. For the general case, various local convergence results are given in Section 5. Section 6 reports our numerical experiments. Finally, we present our conclusions in Section 7.

Notation. We will follow the notation as specified at the end of Section 1 in [32]. In particular,

$$f(V) := \frac{\mathrm{tr}(V^\top AV)}{\mathrm{tr}(V^\top BV)} + \mathrm{tr}(V^\top CV). \tag{1.4}$$

For a matrix Z, $\|Z\|_2$, $\|Z\|_{\mathrm{F}}$, and $\|Z\|_{\mathrm{ui}}$ are the spectral norm, the Frobenius norm, and a general unitarily invariant norm, respectively.

2 Preliminaries

2.1 Angles between subspaces

Consider two subspaces $\mathcal{X}$ and $\mathcal{Y}$ of $\mathbb{R}^m$ and suppose

$$k := \dim(\mathcal{X}) \leqslant \dim(\mathcal{Y}) =: \ell. \tag{2.1}$$

Let $X \in \mathbb{R}^{m\times k}$ and $Y \in \mathbb{R}^{m\times \ell}$ be orthonormal basis matrices of $\mathcal{X}$ and $\mathcal{Y}$, respectively, i.e.,

$$X^\top X = I_k,\ \mathcal{X} = \mathcal{R}(X), \quad \text{and} \quad Y^\top Y = I_\ell,\ \mathcal{Y} = \mathcal{R}(Y),$$

and denote by σ_j for $1 \leqslant j \leqslant k$ in descending order, i.e., $\sigma_1 \geqslant \cdots \geqslant \sigma_k$, the singular values of $Y^\top X$. The k *canonical angles* $\theta_j(\mathcal{X}, \mathcal{Y})$ *from* $\mathcal{X}$ *to* $\mathcal{Y}$ [1)] are defined by

$$0 \leqslant \theta_j(\mathcal{X}, \mathcal{Y}) := \arccos\sigma_j \leqslant \frac{\pi}{2} \quad \text{for } 1 \leqslant j \leqslant k. \tag{2.2}$$

They are in ascending order, i.e.,

$$\theta_1(\mathcal{X}, \mathcal{Y}) \leqslant \cdots \leqslant \theta_k(\mathcal{X}, \mathcal{Y}).$$

Set

$$\Theta(\mathcal{X}, \mathcal{Y}) = \mathrm{diag}(\theta_1(\mathcal{X}, \mathcal{Y}), \ldots, \theta_k(\mathcal{X}, \mathcal{Y})). \tag{2.3}$$

[1)] If $k = \ell$, we may say that these angles are *between* $\mathcal{X}$ and $\mathcal{Y}$.

It can be seen that angles so defined are independent of the orthonormal basis matrices X and Y which are not unique. A different way to define these angles is through orthogonal projections onto $\mathcal{X}$ and $\mathcal{Y}$ [24].

When $k = 1$, i.e., X is a vector, there is only one canonical angle from $\mathcal{X}$ to $\mathcal{Y}$ and so we will simply write $\theta(\mathcal{X}, \mathcal{Y})$.

In what follows, we sometimes place a vector or matrix in one or both arguments of $\theta_j(\cdot\,,\cdot)$, $\theta(\cdot\,,\cdot)$, and $\Theta(\cdot\,,\cdot)$ with the understanding that it is about the subspace spanned by the vector or the columns of the matrix argument.

$\|\sin\Theta(\mathcal{X}, \mathcal{Y})\|_2$ defines a distance metric between $\mathcal{X}$ and $\mathcal{Y}$. It can be proved that [20]

$$\|\sin\Theta(\mathcal{X}, \mathcal{Y})\|_2 = \sin\theta_k(\mathcal{X}, \mathcal{Y}) = \|X^\top Y_\perp\|_2,$$

where $Y_\perp$ is an orthonormal basis matrix of the orthogonal complement of $\mathcal{Y}$.

2.2 Unitarily invariant norm

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

1. $\|XBY\|_{\rm ui} = \|B\|_{\rm ui}$ for all unitary matrices X and Y of apt sizes and $B \in \mathbb{C}^{m\times n}$.
2. $\|B\|_{\rm ui} = \|B\|_2$, the spectral norm of B, if $\mathrm{rank}(B) = 1$.

Two commonly used unitarily invariant norms are

$$\text{the spectral norm}: \|B\|_2 = \max_j \sigma_j,$$
$$\text{the Frobenius norm}: \|B\|_{\rm F} = \sqrt{\sum_j \sigma_j^2},$$

where $\sigma_1, \sigma_2, \ldots, \sigma_{\min\{m,n\}}$ are the singular values of B. The trace norm $\|B\|_{\rm tr} = \sum_j \sigma_j$ is a unitarily invariant norm, too.

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

$$\|X\|_2 \leqslant \|X\|_{\rm ui}, \quad \|XYZ\|_{\rm ui} \leqslant \|X\|_2 \cdot \|Y\|_{\rm ui} \cdot \|Z\|_2 \tag{2.4}$$

for any matrices X, Y, and Z of compatible sizes. By [7, p.176],

$$|\,\mathrm{tr}(X)| \leqslant \|X\|_{\rm tr} \leqslant n\|X\|_2 \quad \text{for } X \in \mathbb{C}^{n\times n}. \tag{2.5}$$

Lemma 2.1 (See [3]). *Let H and M be two Hermitian matrices, and let S be a matrix of a compatible size as determined by the Sylvester equation $HY - YM = S$. If either* $\mathrm{eig}(H)$ *is in a closed interval that contains no eigenvalue of M or vice versa, then the equation has a unique solution Y, and moreover* $\|Y\|_{\rm ui} \leqslant \frac{1}{\tau}\|S\|_{\rm ui}$, *where* $\tau = \min|\mu - \omega|$ *over all* $\mu \in \mathrm{eig}(M)$ *and* $\omega \in \mathrm{eig}(H)$.

3 Self-consistent-field iteration

Theorem 4.1 in [32], on one hand, sheds lights on the optimization problem (1.1) by connecting it to a special nonlinear extreme eigenvalue problem

$$E(V)V = VM_V, \quad \mathrm{eig}(M_V) = \{\lambda_i(E(V)),\ i = 1, 2, \ldots, \ell\}, \tag{3.1}$$

where $M_V = V^\top E(V)V$. On the other hand, it naturally lends itself to Algorithm 3.1, a self-consistent-field (SCF) iterative method for solving (1.1). Several remarks are in order for the algorithm.

Algorithm 3.1. A self-consistent-field iteration

Given $V_0 \in \mathbb{O}^{m\times\ell}$ and a tolerance `tol`, this algorithm computes an approximate maximizer for the optimization problem (1.1).

1: **for** $k = 0, 1, \ldots$ **do**
2: compute an orthonormal eigenbasis V_{k+1} of $E(V_k)$ associated with its ℓ largest eigenvalues;
3: **if** $r_k := \|E(V_{k+1})V_{k+1} - V_{k+1}M_{V_{k+1}}\|_2 \leqslant$ `tol` **then**
4: **BREAK**;
5: **end if**
6: **end for**
7: return V_{k+1} as an approximate maximizer.

1. The SCF iteration is currently one of the most widely used algorithms for solving the Kohn-Sham equations in electronic structure calculations (see e.g., [11, 18, 21, 22, 27, 29]). Recent study on numerical algorithms for electronic structure calculations and the convergence of the SCF iteration for solving the Kohn-Sham equations can be found in, e.g., [10, 25, 36]. In dimension reduction and feature extraction, the effective algorithm discussed in [13,23,33,34] is also the SCF iteration for the trace quotient (or trace ratio) optimization problem.

2. The quantity r_k defined at Line 3 is half the norm of the gradient at V_{k+1} of

$$f_{|\mathbb{O}^{m\times\ell}}(V) : \mathbb{O}^{m\times\ell} \to \mathbb{R}$$

and serves as the *residual* for the approximate V_{k+1} of the nonlinear eigenvalue problem (3.1).

3. If the sequence $\{V_k\}$ converges to $\bar{V}$, then not only $\bar{V}$ is a KKT point of (1.1), but also satisfies the necessary condition in [32, Theorem 4.1] for a global maximizer. This is one of the major advantages of the SCF iteration over optimization-based methods in [1, 5, 14] which primarily concern the monotonic (or non-monotonic) change of the objective value and converge to a KKT point that may or may not satisfy the necessary condition in [32, Theorem 4.1]. Because of this, conceivably the SCF iteration is more likely to achieve a global maximizer than some optimization-based methods.

4. The major computational cost of Algorithm 3.1 lies at Line 2, where a dominant orthonormal eigenbasis of $m \times \ell$ matrix $E(V_k)$ has to be computed every iteration. One may employ any state-of-the-art eigensolver, for example, the QR algorithm [4, 6] if m is modest. But the QR algorithm requires $\mathcal{O}(m^3)$ flops and $\mathcal{O}(m^2)$ storage, thus it is impractical for large m. In the latter case, certain iterative method should be used, for example, the Lanczos method [15, 16], the Jacobi-Davidson iteration [19], the conjugate gradient type method [8], to name a few.

Originally, the SCF iteration is referred to the one for solving the Kohn-Sham equation in electronic structure calculations [17, 28]. The Kohn-Sham equation is a nonlinear eigenvalue problem in PDE which after certain discretization becomes a nonlinear matrix eigenvalue problem whose matrix depends on the eigenvectors to be computed, as opposed to the other type of nonlinear matrix eigenvalue problems whose matrices depend on the eigenvalues to be computed. This distinguishing feature of dependency on the eigenvectors is shared by the current nonlinear eigenvalue problem (3.1), making the SCF iteration a natural method to use.

While the SCF iteration is simple to implement, its convergence in general is not well understood, depending closely on the targeted nonlinear eigenvalue problem. For the trace ratio optimization problem (the case $C = 0$) in the linear discriminant analysis (LDA) [33], it is shown that SCF globally converges to a global maximizer *regardless of the initial guess* V_0, and the convergence is locally quadratic under a generic condition. However, in [27] an artificial nonlinear eigenvalue problem is constructed to show that the SCF iteration may generate a sequence of approximate solutions that contain two convergent subsequences alternating each other neither of which converges to the solution for the nonlinear eigenvalue problem. For our problem (3.1), in general SCF can produce sequences with no subsequences converging to a KKT point as the following example illustrates.

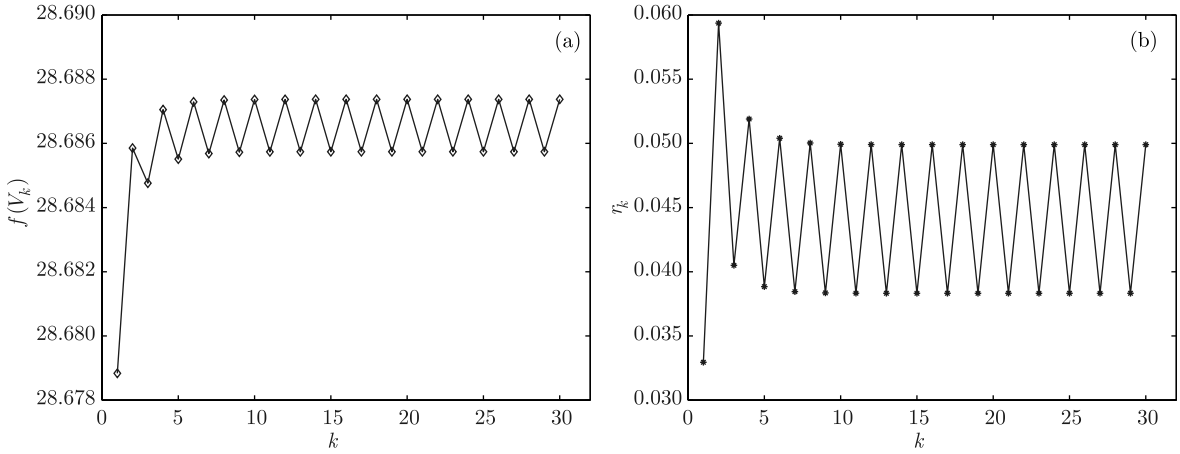

Figure 3.1 (a) $f(V_k)$ and (b) r_k by SCF with $V_0 = [\boldsymbol{e}_1, \boldsymbol{e}_2]$ for Example 3.1

Example 3.1. Let $m = 3,\ \ell = 2$ and

$$A = \begin{bmatrix} 11 & 5 & 8 \\ 5 & 10 & 9 \\ 8 & 9 & 5 \end{bmatrix}, \quad B = \begin{bmatrix} 7 & 7 & 7 \\ 7 & 10 & 8 \\ 7 & 8 & 8 \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} 15 & 10 & 9 \\ 10 & 7 & 6 \\ 9 & 6 & 6 \end{bmatrix}.$$

With $V_0 = [\boldsymbol{e}_1, \boldsymbol{e}_2]$, the sequences $\{f(V_k)\}$ and $\{r_k\}$ generated by the SCF iteration oscillate and diverge (see Figure 3.1). We observe that there are two convergent subsequences (see Figure 3.1) of $\{f(V_k)\}$, but neither approaches a solution to (1.1).

There are some strategies to curb the oscillation behavior of SCF, for example, a trust-region SCF iteration, which has shown some effectiveness in the case of the Kohn-Sham equation [18, 22, 29], but we will not pursue this here but in our future work.

Despite that in general there isn't much we can say about SCF's global convergence behavior, for certain special cases we do know more. One particular example is the case when $C = 0$ investigated in [33]. In the next section, we will add another example to the list of special cases by establishing a global convergence result for the case $C = \eta B$ $(\eta > 0)$.

4 Global convergence for the case $C = \eta B$ $(\eta > 0)$

Lemma 4.1. *Suppose $V, \bar{V} \in \mathbb{O}^{m\times\ell}$. Then there exists an orthogonal matrix $Z \in \mathbb{R}^{\ell\times\ell}$ such that*

$$\|\sin\Theta(V, \bar{V})\|_{\mathrm{ui}} \leqslant \|VZ - \bar{V}\|_2 \leqslant \sqrt{2}\,\|\sin\Theta(V, \bar{V})\|_{\mathrm{ui}}, \tag{4.1}$$

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

Proof. Suppose for the moment that $2\ell \leqslant m$. There exist orthogonal matrices $Q \in \mathbb{R}^{m\times m}$, $Z_i \in \mathbb{R}^{\ell\times\ell}$ such that [20, p. 40]

$$QVZ_1 = \begin{matrix} \\ \ell \\ \ell \\ m-2\ell \end{matrix} \begin{bmatrix} \overset{\ell}{I} \\ 0 \\ 0 \end{bmatrix}, \quad Q\bar{V}Z_2 = \begin{matrix} \\ \ell \\ \ell \\ m-2\ell \end{matrix} \begin{bmatrix} \overset{\ell}{\Gamma} \\ \Sigma \\ 0 \end{bmatrix},$$

where

$$\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_\ell), \quad \gamma_1 \geqslant \cdots \geqslant \gamma_\ell \geqslant 0,$$
$$\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_\ell), \quad 0 \leqslant \sigma_1 \leqslant \cdots \leqslant \sigma_\ell,$$

$$\gamma_i = \cos\theta_i(V,\bar{V}), \quad \sigma_i = \sin\theta_i(V,\bar{V}), \quad \text{for } 1 \leqslant i \leqslant \ell.$$

Take $Z = Z_1 Z_2^\top$. The singular values of $VZ - \bar{V}$ are

$$\sqrt{(1-\cos\theta_i)^2 + \sin^2\theta_i} = 2\sin\frac{\theta_i}{2} \quad \text{for } 1 \leqslant i \leqslant \ell,$$

where $\theta_i = \theta_i(V,\bar{V})$. The inequalities in (4.1) are now simple consequences, upon noticing

$$\sin\theta \leqslant 2\sin\frac{\theta}{2} = \frac{\sin\theta}{\cos\theta/2} \leqslant \sqrt{2}\,\sin\theta \quad \text{for } 0 \leqslant \theta \leqslant \frac{\pi}{2}.$$

The case for $2\ell > m$ can be dealt with in the same way, except that we use [20, p. 41]

$$QVZ_1 = \begin{array}{c} \\ m-\ell \\ 2\ell-m \\ m-\ell \end{array} \begin{array}{c} \begin{array}{cc} m-\ell & 2\ell-m \end{array} \\ \begin{bmatrix} I & 0 \\ 0 & I \\ 0 & 0 \end{bmatrix} \end{array}, \quad Q\bar{V}Z_2 = \begin{array}{c} \\ m-\ell \\ 2\ell-m \\ m-\ell \end{array} \begin{array}{c} \begin{array}{cc} m-\ell & 2\ell-m \end{array} \\ \begin{bmatrix} \varGamma & 0 \\ 0 & I \\ \varSigma & 0 \end{bmatrix} \end{array},$$

where

$$\begin{aligned} &\varGamma = \operatorname{diag}(\gamma_1,\ldots,\gamma_{m-\ell}), \quad \gamma_1 \geqslant \cdots \geqslant \gamma_{m-\ell} \geqslant 0,\\ &\varSigma = \operatorname{diag}(\sigma_1,\ldots,\sigma_{m-\ell}), \quad 0 \leqslant \sigma_1 \leqslant \cdots \leqslant \sigma_{m-\ell},\\ &\gamma_i = \cos\theta_{2\ell-m+i}(V,\bar{V}), \quad \sigma_i = \sin\theta_{2\ell-m+i}(V,\bar{V}), \quad \text{for } 1 \leqslant i \leqslant m-\ell, \end{aligned}$$

and $\theta_i(V,\bar{V}) = 0$ for $1 \leqslant i \leqslant 2\ell - m$. □

The inequality (4.2) below can also be deduced from [9, Theorem 2.1], but since (4.2) is much simpler to prove, a short proof is given for self-containedness.

Lemma 4.2. *Let $H \in \mathbb{R}^{m\times m}$ be a symmetric matrix. For $V, \bar{V} \in \mathbb{O}^{m\times\ell}$, we have*

$$|\operatorname{tr}(\bar{V}^\top H\bar{V}) - \operatorname{tr}(V^\top HV)| \leqslant 2\sqrt{2}\,\|H\|_2\tilde{\epsilon} \leqslant 2\sqrt{2}\,\ell\|H\|_2\epsilon, \tag{4.2}$$

where

$$\tilde{\epsilon} = \|\sin\varTheta(V,\bar{V})\|_{\mathrm{tr}}, \quad \epsilon = \|\sin\varTheta(V,\bar{V})\|_2. \tag{4.3}$$

In particular, for the function f defined by (1.4),

$$\lim_{\|\sin\varTheta(\bar{V},V)\|_2\to 0} f(V) = f(\bar{V}). \tag{4.4}$$

Proof. Let Z be the one in Lemma 4.1, and $\varDelta V = \bar{V} - VZ$. We have

$$\begin{aligned} \operatorname{tr}(\bar{V}^\top H\bar{V}) - \operatorname{tr}(V^\top HV) &= \operatorname{tr}(\bar{V}^\top H\bar{V}) - \operatorname{tr}(Z^\top V^\top HVZ)\\ &= \operatorname{tr}(\bar{V}^\top H\bar{V}) - \operatorname{tr}(\bar{V}^\top HVZ) + \operatorname{tr}(\bar{V}^\top HVZ) - \operatorname{tr}(Z^\top V^\top HVZ)\\ &= \operatorname{tr}(\bar{V}^\top H\varDelta V) + \operatorname{tr}([\varDelta V]^\top HVZ). \end{aligned}$$

Therefore use (2.4) and (2.5) to get

$$|\operatorname{tr}(\bar{V}^\top H\bar{V}) - \operatorname{tr}(V^\top HV)| \leqslant \|\bar{V}^\top H\varDelta V\|_{\mathrm{tr}} + \|[\varDelta V]^\top HVZ\|_{\mathrm{tr}} \leqslant 2\|H\|_2\|\varDelta V\|_{\mathrm{tr}}$$

which is the first inequality in (4.2). The second inequality there holds because $\tilde{\epsilon} \leqslant \ell\epsilon$. □

Theorem 4.3. *Suppose $A \in \mathbb{R}^{m\times m}$ is symmetric and $C = \eta B \in \mathbb{R}^{m\times m}$ $(\eta > 0)$ is symmetric and positive definite. Let $\{V_k\}$ be the sequence generated by* Algorithm 3.1 *with an arbitrarily given $V_0 \in \mathbb{O}^{m\times\ell}$. Let $\mathcal{V}_k = \mathcal{R}(V_k)$, the column space of V_k. Then the following statements hold.*

1. $\{\mathcal{V}_k\}$ *has a convergent subsequence* $\{\mathcal{V}_{k_j}\}$ *in the sense that there is an* ℓ*-dimensional subspace* $\bar{\mathcal{V}} \subset \mathbb{R}^m$ *such that*

$$\lim_{j\to\infty} \|\sin\Theta(\mathcal{V}_{k_j}, \bar{\mathcal{V}})\|_2 = 0. \tag{4.5}$$

2. *The sequence* $\{f(V_k)\}$ *monotonically increases and converges to* $f(\bar{V})$*, where* $\bar{V}$ *is an orthonormal basis matrix of* $\bar{\mathcal{V}}$*.*

3. $\bar{\mathcal{V}}$ *is an invariant subspace of* $E(\bar{V})$ *corresponding to its* ℓ *largest eigenvalues. In the other word, the orthonormal basis matrix of any limit point of* $\{\mathcal{V}_k\}$ *satisfies the necessary condition given in* [32, *Theorem* 4.1].

4. *If* $\lambda_\ell(E(\bar{V})) > \lambda_{\ell+1}(E(\bar{V}))$*, i.e., the dominant* ℓ*-dimensional eigenspace of* $E(\bar{V})$ *is* unique*, then for any positive integer* q*, we have*

$$\lim_{j\to\infty} \|\sin\Theta(V_{k_j\pm q}, \bar{V})\|_2 = 0. \tag{4.6}$$

As an interesting consequence, if $\max|k_{j+1} - k_j|$ *over all* $1 \leqslant j \leqslant \infty$ *is finite, then* $\|\sin\Theta(V_k, \bar{V})\|_2 \to 0$ *as* $k \to \infty$*.*

Proof. All subspaces of dimension ℓ in $\mathbb{R}^m$ form a Grassmann manifold which is compact [12, p. 57] with the metric given by $\|\sin\Theta(\cdot\,,\cdot)\|_2$. $\{\mathcal{V}_k\}$ is a sequence on the manifold and thus has a convergent subsequence $\{\mathcal{V}_{k_j}\}$ that converges to an ℓ-dimensional subspace $\bar{\mathcal{V}} \subset \mathbb{R}^m$ in the sense of (4.5). This is item 1.

Before we prove item 2, we notice that $C = \eta B$ leads to

$$M_V = V^\top E(V)V = V^\top CV = \eta V^\top BV \quad \text{for any } V \in \mathbb{O}^{m\times\ell}. \tag{4.7}$$

Now we have from [32, (4.3)]

$$\Delta f_k := f(V_{k+1}) - f(V_k) = \frac{\phi_B(V_k)[\mathrm{tr}(V_{k+1}^\top E(V_k)V_{k+1}) - \eta\phi_B(V_k)] + \eta[\phi_B(V_{k+1}) - \phi_B(V_k)]^2}{\phi_B(V_{k+1})}. \tag{4.8}$$

Because V_{k+1} is an orthonormal eigenbasis of $E(V_k)$ corresponding to its ℓ largest eigenvalues, implying

$$\mathrm{tr}(V_{k+1}^\top E(V_k)V_{k+1}) \geqslant \mathrm{tr}(V_k^\top E(V_k)V_k) = \eta\phi_B(V_k)$$

by (4.7), it follows that $\Delta f_k \geqslant 0$. Thus $\{f(V_k)\}$ is nondecreasing and convergent since $\{|f(V_k)|\}$ is bounded. Since $\|\sin\Theta(V_{k_j}, \bar{V})\|_2 \to 0$, as $j \to \infty$, by (4.4) we conclude that $z\lim_{k\to\infty} f(V_k) = \lim_{j\to\infty} f(V_{k_j}) = f(\bar{V})$. This proves item 2.

Making use of $\Delta f_k \to 0$, we conclude from (4.8)

$$\lim_{k\to\infty} [\mathrm{tr}(V_{k+1}^\top E(V_k)V_{k+1}) - \eta\phi_B(V_k)] = 0, \tag{4.9}$$

$$\lim_{k\to\infty} [\phi_B(V_{k+1}) - \phi_B(V_k)] = 0. \tag{4.10}$$

Again since $\|\sin\Theta(V_{k_j}, \bar{V})\|_2 \to 0$ as $j \to \infty$, by (4.4) and (4.10) we have

$$\lim_{j\to\infty} \phi_B(V_{k_j+1}) = \lim_{j\to\infty} \phi_B(V_{k_j}) = \phi_B(\bar{V}), \tag{4.11}$$

$$\lim_{j\to\infty} \phi_A(V_{k_j}) = \phi_A(\bar{V}). \tag{4.12}$$

Combine (4.9) with (4.11) to get

$$\lim_{j\to\infty} \mathrm{tr}(V_{k_j+1}^\top E(V_{k_j})V_{k_j+1}) = \eta \lim_{j\to\infty} \phi_B(V_{k_j}) = \eta\phi_B(\bar{V}). \tag{4.13}$$

According to the SCF iteration, we have

$$E(V_k)V_{k+1} = V_{k+1}\widetilde{M}_{k+1} \quad \text{with} \quad \widetilde{M}_{k+1} = V_{k+1}^\top E(V_k)V_{k+1}.$$

Notice that

$$\Delta E_k := E(V_k) - E(\bar{V}) = \left(\frac{1}{\phi_B(V_k)} - \frac{1}{\phi_B(\bar{V})}\right) A + \left(\frac{\phi_A(\bar{V})}{[\phi_B(\bar{V})]^2} - \frac{\phi_A(V_k)}{[\phi_B(V_k)]^2}\right) B.$$

Use (4.11) and (4.12) to see

$$\lim_{j\to\infty} \Delta E_{k_j} = 0. \tag{4.14}$$

Therefore, by the continuity of eigenvalues with respect to matrix entries [20], we know

$$\lim_{j\to\infty} \operatorname{tr}(\widetilde{M}_{k_j+1}) = \lim_{j\to\infty} \sum_{s=1}^{\ell} \lambda_s(E(V_{k_j})) = \sum_{s=1}^{\ell} \lambda_s(E(\bar{V})). \tag{4.15}$$

Together, (4.13) and (4.15) yield

$$\lim_{j\to\infty} \operatorname{tr}(\widetilde{M}_{k_j+1}) = \sum_{s=1}^{\ell} \lambda_s(E(\bar{V})) = \eta\phi_B(\bar{V}). \tag{4.16}$$

This shows that $\eta\phi_B(\bar{V})$ is the sum of the ℓ largest eigenvalues of $E(\bar{V})$. Therefore by (4.7),

$$\operatorname{tr}(\bar{V}^\top E(\bar{V})\bar{V}) = \eta\phi_B(\bar{V}) = \sum_{s=1}^{\ell} \lambda_s(E(\bar{V})).$$

So $\bar{V}$ must be an orthonormal eigenbasis of $E(\bar{V})$ corresponding to its ℓ largest eigenvalues. This proves item 3.

We prove item 4 by induction on q. The assumption that $\lambda_\ell(E(\bar{V})) > \lambda_{\ell+1}(E(\bar{V}))$ implies that the eigenspace of $E(\bar{V})$ associated with its first ℓ largest eigenvalues is unique, i.e., $\bar{\mathcal{V}}$ is unique even though its orthonormal basis matrix $\bar{V}$ is not. $E(V_{k_j}) \to E(\bar{V})$ as $j \to \infty$ according to (4.14) and thus for sufficiently large j, $\lambda_\ell(E(V_{k_j})) > \lambda_{\ell+1}(E(V_{k_j}))$ and the eigenspace $\mathcal{V}_{k_j+1}$ of $E(V_{k_j})$ associated with its first ℓ largest eigenvalues is unique and converges to $\bar{\mathcal{V}}$ by Davis-Kahan $\sin 2\theta$ theorem [3], i.e., (4.6) holds for $q = 1$. Suppose now (4.6) holds for q, and we need to show it is also true for $q+1$. To see this, we simply rename $k_j + q$ to k_j and apply the argument we just did. □

We remark that the convergence behavior of the special case $C = \mu B$ is similar to that for the trace ratio problem (i.e., $C = 0$): they both converge monotonically and globally. However, these two cases are still quite different from each other and one remarkable difference is that the trace ratio problem only admits global maximizers (i.e., any local maximizer is a global maximizer [35, Theorem 1.1], while the case $C = \mu B$ contains both local and global maximizers (see [30, Example 3.1] for an example).

5 Local convergence for the general case

We now investigate the local convergence behavior of the SCF iteration for the general case. Throughout this section,

$$\tilde{\epsilon} = \|\sin\Theta(V, \bar{V})\|_{\mathrm{tr}}, \quad \epsilon = \|\sin\Theta(V, \bar{V})\|_2, \tag{4.3}$$

where $V, \bar{V} \in \mathbb{O}^{m\times\ell}$. Then $\epsilon \leqslant \tilde{\epsilon} \leqslant \ell\epsilon$. R_A and R_B are as defined by (5.2) below with $H \in \{A, B\}$.

Lemma 5.1 is a refinement of Lemma 4.2 when $\mathcal{R}(\bar{V})$ is an approximate invariant subspace of H.

Lemma 5.1. *Let $H \in \mathbb{R}^{m\times m}$ be a symmetric matrix. For $V, \bar{V} \in \mathbb{O}^{m\times\ell}$, we have*

$$|\phi_H(\bar{V}) - \phi_H(V)| \leqslant (2\sqrt{2}\,\|R_H\|_2 + 4\|H\|_2\epsilon)\tilde{\epsilon}, \tag{5.1}$$

where $\phi_H(V) := \operatorname{tr}(V^\top HV)$, and

$$R_H = H\bar{V} - \bar{V}(\underbrace{\bar{V}^\top H\bar{V}}_{=:H_1}). \tag{5.2}$$

Proof. Let $Z \in \mathbb{R}^{\ell\times\ell}$ be the one in Lemma 4.1, and $\Delta V = \bar{V} - VZ$. We have

$$\begin{aligned}|\phi_H(\bar{V}) - \phi_H(V)| &= |\phi_H(\bar{V}) - \phi_H(VZ)| \\ &\leqslant |\operatorname{tr}(\Delta V^\top H\bar{V}) + \operatorname{tr}(\bar{V}^\top H\Delta V)| + |\operatorname{tr}(\Delta V^\top H\Delta V)|. \end{aligned} \tag{5.3}$$

Use $H\bar{V} = R_H + \bar{V}H_1$ to get

$$\begin{aligned}&\operatorname{tr}(\Delta V^\top H\bar{V}) + \operatorname{tr}(\bar{V}^\top H\Delta V) \\ &\quad= \operatorname{tr}(\Delta V^\top R_H) + \operatorname{tr}(\Delta V^\top \bar{V}H_1) + \operatorname{tr}(R_H^\top \Delta V) + \operatorname{tr}(H_1\bar{V}^\top\Delta V) \\ &\quad= 2\operatorname{tr}(\Delta V^\top R_H) + \operatorname{tr}(\Delta V^\top\bar{V}H_1) + \operatorname{tr}(\bar{V}^\top\Delta V H_1) \\ &\quad= 2\operatorname{tr}(\Delta V^\top R_H) + \operatorname{tr}([\Delta V^\top\bar{V} + \bar{V}^\top\Delta V]H_1) \\ &\quad= 2\operatorname{tr}(\Delta V^\top R_H) + \operatorname{tr}(\Delta V^\top\Delta V H_1), \end{aligned} \tag{5.4}$$

where we have used $\Delta V^\top\bar{V} + \bar{V}^\top\Delta V = \Delta V^\top\Delta V$, which can be verified by the substitution $\Delta V = \bar{V} - VZ$. Combine (5.4) with (5.3) to get

$$\begin{aligned}|\phi_H(\bar{V}) - \phi_H(V)| &\leqslant 2|\operatorname{tr}(\Delta V^\top R_H)| + |\operatorname{tr}(\Delta V^\top\Delta V H_1)| + |\operatorname{tr}(\Delta V^\top H\Delta V)| \\ &\leqslant 2\|\Delta V^\top R_H\|_{\mathrm{tr}} + \|\Delta V^\top\Delta V H_1\|_{\mathrm{tr}} + \|\Delta V^\top H\Delta V\|_{\mathrm{tr}} \qquad \text{(by (2.5))} \\ &\leqslant 2\|\Delta V\|_{\mathrm{tr}}\|R_H\|_2 + \|\Delta V\|_2\|\Delta V\|_{\mathrm{tr}}(\|H_1\|_2 + \|H\|_2), \end{aligned}$$

and then use Lemma 4.1 and $\|H_1\|_2 \leqslant \|H\|_2$ to conclude (5.1). □

Note that $R_H = 0$ is equivalent to $\|\sin\Theta(\bar{V}, H\bar{V})\|_2 = 0$, i.e., $\mathcal{R}(\bar{V})$ is an invariant subspace of H.

To establish the main theorem, Theorem 5.9, of this section, we additionally introduce the following notation,

$$\omega_B := \sum_{j=m-\ell+1}^{m} \lambda_j(B), \quad \Omega_B := \sum_{j=1}^{\ell}\lambda_j(B). \tag{5.5}$$

Recall that B is positive definite but A may be indefinite. We have for any $V \in \mathbb{O}^{m\times\ell}$,

$$0 < \omega_B \leqslant \phi_B(V) \leqslant \Omega_B \quad \text{and} \quad |\phi_A(V)| \leqslant \Omega_A := \max\left\{\left|\sum_{j=m-\ell+1}^{m}\lambda_j(A)\right|, \left|\sum_{j=1}^{\ell}\lambda_j(A)\right|\right\}. \tag{5.6}$$

Lemma 5.2. *For $V, \bar{V} \in \mathbb{O}^{m\times\ell}$, we have*

$$\left|\frac{1}{\phi_B(V)} - \frac{1}{\phi_B(\bar{V})}\right| \leqslant \frac{2\|B\|_2}{\omega_B^2}\,\tilde{\epsilon} \leqslant \frac{2\ell\|B\|_2}{\omega_B^2}\,\epsilon, \tag{5.7a}$$

$$\left|\frac{1}{\phi_B(V)} - \frac{1}{\phi_B(\bar{V})}\right| \leqslant \frac{2[\sqrt{2}\,\|R_B\|_2 + 2\|B\|_2\epsilon]}{\omega_B^2}\,\tilde{\epsilon} \leqslant \frac{2\ell[\sqrt{2}\,\|R_B\|_2 + 2\|B\|_2\epsilon]}{\omega_B^2}\,\epsilon, \tag{5.7b}$$

and

$$\left|\frac{\phi_A(\bar{V})}{[\phi_B(\bar{V})]^2} - \frac{\phi_A(V)}{[\phi_B(V)]^2}\right| \leqslant \frac{2}{\omega_B^4}[2\Omega_A\Omega_B\|B\|_2 + \Omega_B^2\|A\|_2]\,\tilde{\epsilon}, \tag{5.8a}$$

$$\begin{aligned}\left|\frac{\phi_A(\bar{V})}{[\phi_B(\bar{V})]^2} - \frac{\phi_A(V)}{[\phi_B(V)]^2}\right| &\leqslant \frac{2}{\omega_B^4}[2\sqrt{2}(\Omega_B^2\|R_A\|_2 + 2\Omega_A\Omega_B\|R_B\|_2) \\ &\quad + 2(\Omega_B^2\|A\|_2 + 2\Omega_A\Omega_B\|B\|_2)\epsilon]\tilde{\epsilon}. \end{aligned} \tag{5.8b}$$

Proof. By Lemmas 4.2 and 5.1, we can write for $H \in \{A, B\}$,

$$\phi_H(V) = \phi_H(\bar{V}) + \epsilon_H, \tag{5.9a}$$

$$|\epsilon_H| \leqslant 2\|H\|_2\tilde{\epsilon} \leqslant 2\ell\|H\|_2\epsilon, \tag{5.9b}$$

$$|\epsilon_H| \leqslant (2\sqrt{2}\,\|R_H\|_2 + 4\|H\|_2\epsilon)\tilde{\epsilon} \leqslant \ell(2\sqrt{2}\,\|R_H\|_2 + 4\|H\|_2\epsilon)\epsilon. \tag{5.9c}$$

Therefore

$$\begin{aligned}
&\left|\frac{1}{\phi_B(V)} - \frac{1}{\phi_B(\bar{V})}\right| = \left|\frac{-\epsilon_B}{\phi_B(V)\phi_B(\bar{V})}\right|,\\
&\left|\frac{\phi_A(\bar{V})}{[\phi_B(\bar{V})]^2} - \frac{\phi_A(V)}{[\phi_B(V)]^2}\right| = \left|\frac{[\phi_A(\bar{V}) - \phi_A(V)][\phi_B(V)]^2 + \phi_A(V)\{[\phi_B(V)]^2 - [\phi_B(\bar{V})]^2\}}{[\phi_B(V)\phi_B(\bar{V})]^2}\right|\\
&\qquad\qquad \leqslant \frac{\Omega_B^2\,|\epsilon_A| + 2\Omega_A\Omega_B\,|\epsilon_B|}{\omega_B^4}.
\end{aligned}$$

Now use (5.9b) and (5.9c) to bound ϵ_B to get (5.7a) and (5.7b), respectively; use (5.9b) to bound ϵ_A and ϵ_B to get (5.8a); use (5.9c) to bound ϵ_A and ϵ_B to get (5.8b). □

Remark 5.3. In this lemma and many expressions in what follows, we find ω_B^2 and ω_B^4 in the denominators of various fractions, as the results of bounding $\phi_B(V)\phi_B(\bar{V})$ and $[\phi_B(V)\phi_B(\bar{V})]^2$ from below, respectively, simply by using (5.6). In cases where ω_B is rather tiny relative to Ω_B, it may make various upper bound estimates unnecessarily too big. A better option may be to use

$$\phi_B(V)\phi_B(\bar{V}) = \phi_B(\bar{V})[\phi_B(\bar{V}) + \epsilon_B]$$

followed by bounding ϵ_B. We omit the details.

Lemma 5.4. *For $V, \bar{V} \in \mathbb{O}^{m\times\ell}$, we have*

$$\|E(\bar{V}) - E(V)\|_2 \leqslant \underbrace{(\chi_1 + \chi_2)\|B\|_2}_{=:\chi}\,\epsilon, \tag{5.10}$$

where

$$\chi_1 = \frac{2\ell\|A\|_2}{\omega_B^2} \quad \text{and} \quad \chi_2 = \frac{2\ell\Omega_B[2\Omega_A\|B\|_2 + \Omega_B\|A\|_2]}{\omega_B^4}. \tag{5.11}$$

Proof. It can be verified that

$$\Delta E := E(V) - E(\bar{V}) = \left[\frac{1}{\phi_B(V)} - \frac{1}{\phi_B(\bar{V})}\right]A + \left[\frac{\phi_A(\bar{V})}{[\phi_B(\bar{V})]^2} - \frac{\phi_A(V)}{[\phi_B(V)]^2}\right]B. \tag{5.12}$$

Therefore

$$\|\Delta E\|_2 \leqslant \left|\frac{1}{\phi_B(V)} - \frac{1}{\phi_B(\bar{V})}\right|\,\|A\|_2 + \left|\frac{\phi_A(\bar{V})}{[\phi_B(\bar{V})]^2} - \frac{\phi_A(V)}{[\phi_B(V)]^2}\right|\,\|B\|_2.$$

Now use (5.7a) and (5.8a) to complete the proof. □

Remark 5.5. Sharper bounds than (5.10) are possible through using (5.7b) and (5.8b) instead. But the use of these possible sharper bounds does not lead to an essentially different conclusion later in our main theorem, Theorem 5.9.

Lemma 5.6. *Let $H \in \mathbb{R}^{m\times m}$ be a symmetric matrix, and $V_+, \bar{V} \in \mathbb{O}^{m\times\ell}$, and R_H be defined by* (5.2). *Let $\bar{V}_\perp \in \mathbb{O}^{m\times(m-\ell)}$ such that $[\bar{V}, \bar{V}_\perp]$ is orthogonal. Then for any unitarily invariant norm $\|\cdot\|_{\rm ui}$,*

$$\|\bar{V}_\perp^\top H V_+\|_{\rm ui} \leqslant \sqrt{2}\,\|H\|_2\|\sin\Theta(V_+, \bar{V})\|_{\rm ui} + \|R_H\|_{\rm ui}\,. \tag{5.13}$$

Proof. Let $Z \in \mathbb{R}^{\ell\times\ell}$ be the one in Lemma 4.1 applied to V_+ and $\bar{V}$ such that

$$\|\bar{V} - V_+ Z\|_{\rm ui} \leqslant \sqrt{2}\|\sin\Theta(V_+, \bar{V})\|_{\rm ui}.$$

Now notice

$$\begin{aligned}
\bar{V}_\perp^\top H V_+ &= \bar{V}_\perp^\top H(V_+Z - \bar{V})Z^\top + \bar{V}_\perp^\top H\bar{V}Z^\top\\
&= \bar{V}_\perp^\top H(V_+Z - \bar{V})Z^\top + \bar{V}_\perp^\top(R_H + \bar{V}H_1)Z^\top\\
&= \bar{V}_\perp^\top H(V_+Z - \bar{V})Z^\top + \bar{V}_\perp^\top R_H Z^\top
\end{aligned}$$

to get $\|\bar{V}_\perp^\top H V_+\|_{\rm ui} \leqslant \|H\|_2\|V_+Z - \bar{V}\|_{\rm ui} + \|R_H\|_{\rm ui} \leqslant \sqrt{2}\,\|H\|_2\|\sin\Theta(V_+, \bar{V})\|_{\rm ui} + \|R_H\|_{\rm ui}$. This completes the proof. □

Lemma 5.7. *Suppose $\bar{V} \in \mathbb{O}^{m\times\ell}$ is an orthonormal eigenbasis of $E(\bar{V})$ associated with its ℓ largest eigenvalues, and let $\delta := \lambda_\ell(E(\bar{V}))-\lambda_{\ell+1}(E(\bar{V}))$. Define R_B and B_1 by (5.2) with $H = B$. For $V \in \mathbb{O}^{m\times\ell}$, let $V_+ \in \mathbb{O}^{m\times\ell}$ be an orthonormal eigenbasis of $E(V)$ associated with its ℓ largest eigenvalues. If*

$$\delta > (\chi + \sqrt{2}\chi_2\|B\|_2)\epsilon, \tag{5.14}$$

then

$$\|\sin\Theta(V_+, \bar{V})\|_2 \leqslant \frac{\tau_1\|R_B\|_2\epsilon + \tau_2\epsilon^2}{\delta - (\chi + \sqrt{2}\chi_2\|B\|_2)\epsilon}, \tag{5.15}$$

where χ and χ_2 are defined in (5.10) *and* (5.11)*, and*

$$\tau_1 = \frac{2\sqrt{2}\ell\|A\|_2}{\omega_B^2} + \chi_2, \quad \tau_2 = \frac{4\ell\|A\|_2\|B\|_2}{\omega_B^2}. \tag{5.16}$$

Proof. By the assumption $E(V)V_+ = V_+\Lambda_1$, where $\Lambda_1 \in \mathbb{R}^{\ell\times\ell}$ and $\mathrm{eig}(\Lambda_1) = \{\lambda_i(E(V))\}_{i=1}^{\ell}$. Let

$$R_1 := E(\bar{V})V_+ - V_+\Lambda_1 = [E(\bar{V}) - E(V)]V_+. \tag{5.17}$$

Let $\bar{V}_\perp \in \mathbb{O}^{m\times(m-\ell)}$ such that $[\bar{V}, \bar{V}_\perp]$ is orthogonal. By the assumption, $\bar{V}_\perp$ is the orthonormal basis matrix associated the smallest $m-\ell$ eigenvalues of $E(\bar{V})$. So we can write $\bar{V}_\perp^\top E(\bar{V}) = \widehat{\Lambda}_2\bar{V}_\perp^\top$, where $\widehat{\Lambda}_2 \in \mathbb{R}^{(m-\ell)\times(m-\ell)}$ and $\mathrm{eig}(\widehat{\Lambda}_2) = \{\lambda_i(E(\bar{V}))\}_{i=\ell+1}^{m}$. We have

$$\bar{V}_\perp^\top R_1 = \bar{V}_\perp^\top[E(\bar{V})V_+ - V_+\Lambda_1] = \widehat{\Lambda}_2\bar{V}_\perp^\top V_+ - \bar{V}_\perp^\top V_+\Lambda_1. \tag{5.18}$$

Since $|\lambda_i(E(V)) - \lambda_i(E(\bar{V}))| \leqslant \|E(\bar{V}) - E(V)\|_2 \leqslant \chi\,\epsilon$ by [20, p. 203] and Lemma 5.4, we have, for $i \leqslant \ell$ and $\ell+1 \leqslant j$,

$$\lambda_i(E(V)) - \lambda_j(E(\bar{V})) \geqslant \lambda_\ell(E(V)) - \lambda_{\ell+1}(E(\bar{V})) \geqslant \lambda_\ell(E(\bar{V})) - \chi\epsilon - \lambda_{\ell+1}(E(\bar{V})) = \delta - \chi\epsilon. \tag{5.19}$$

Therefore by Lemma 2.1, we have

$$\|\sin\Theta(V_+, \bar{V})\|_2 = \|V_\perp^\top V_+\|_2 \leqslant \frac{\|\bar{V}_\perp^\top R_1\|_2}{\delta - \chi\epsilon}. \tag{5.20}$$

To bound $\|\bar{V}_\perp^\top R_1\|_2$, we have, using (5.17) and (5.12),

$$\begin{aligned}\|\bar{V}_\perp^\top R_1\|_2 &= \|\bar{V}_\perp^\top[E(\bar{V}) - E(V)]V_+\|_2\\ &\leqslant \left|\frac{1}{\phi_B(V)} - \frac{1}{\phi_B(\bar{V})}\right|\|\bar{V}_\perp^\top AV_+\|_2 + \left|\frac{\phi_A(\bar{V})}{[\phi_B(\bar{V})]^2} - \frac{\phi_A(V)}{[\phi_B(V)]^2}\right|\|\bar{V}_\perp^\top BV_+\|_2\\ &=: \varepsilon_1 + \varepsilon_2.\end{aligned} \tag{5.21}$$

For ε_1, we use (5.7b) to get

$$\varepsilon_1 \leqslant \frac{2\sqrt{2}\ell\|R_B\|_2\epsilon + 4\ell\|B\|_2\epsilon^2}{\omega_B^2}\|A\|_2. \tag{5.22}$$

For ε_2, we use (5.8a) to get

$$\varepsilon_2 \leqslant \chi_2\epsilon\|\bar{V}_\perp^\top BV_+\|_2 \leqslant \chi_2\epsilon[\|B\|_2\sqrt{2}\|\sin\Theta(V_+, \bar{V})\|_2 + \|R_B\|_2]. \tag{5.23}$$

Finally, combine (5.20)–(5.23) to arrive at, under (5.14),

$$\|\sin\Theta(V_+, \bar{V})\|_2 \leqslant \frac{\varepsilon_1 + \chi_2\|R_B\|_2\epsilon}{\delta - (\chi + \sqrt{2}\chi_2\|B\|_2)\epsilon},$$

which, together with (5.22), leads to (5.15). □

Remark 5.8. The inequality (5.20) remains valid in any unitarily invariant norm:

$$\|\sin\Theta(V_+,\bar{V})\|_{\rm ui} = \|V_\perp^\top V_+\|_{\rm ui} \leqslant \frac{\|\bar{V}_\perp^\top R_1\|_{\rm ui}}{\delta-\chi\epsilon}. \tag{5.20'}$$

The machinery we have built so far in various lemmas allows us to easily bound $\|\bar{V}_\perp^\top R_1\|_{\rm ui}$ in the similar way. The outcome will be theoretically interesting, but may add little to our understanding of SCF than Lemma 5.7 as is. This same remark applies to the next theorem in which the analysis is in terms of $\|\cdot\|_2$ but can be done in terms of any unitarily invariant norm, too.

Theorem 5.9. *Suppose $\bar{V}\in\mathbb{O}^{m\times\ell}$ is an orthonormal eigenbasis of $E(\bar{V})$ associated with its ℓ largest eigenvalues, and let $\delta := \lambda_\ell(E(\bar{V})) - \lambda_{\ell+1}(E(\bar{V}))$. Define R_A, A_1 and R_B, B_1 by* (5.2) *with $H=A$ and $H=B$, respectively, and let χ, χ_2, τ_1, and τ_2 be as defined in* (5.10), (5.11), *and* (5.16), *respectively. Apply the SCF iteration* (*Algorithm* 3.1) *to generate a sequence $\{V_k\}$, given V_0.*

1. *Given any $0<t<1$ and $0<\nu<1$, if*

$$\|R_B\|_2 \leqslant t\nu\,\delta/\tau_1, \tag{5.24}$$

then for any $V_0\in\mathbb{O}^{m\times\ell}$ satisfying

$$\|\sin\Theta(V_0,\bar{V})\|_2 < \frac{(1-t)\nu\delta}{\tau_2+\nu(\chi+\sqrt{2}\chi_2\|B\|_2)}, \tag{5.25}$$

the sequence $\{\mathcal{V}_k := \mathcal{R}(V_k)\}$ converges to $\bar{\mathcal{V}} := \mathcal{R}(\bar{V})$ at least linearly, *and moreover*

$$\|\sin\Theta(V_{k+1},\bar{V})\|_2 \leqslant \nu\|\sin\Theta(V_k,\bar{V})\|_2 \quad \text{for } k=0,1,\ldots \tag{5.26}$$

2. *If $R_B=0$, then for any $V_0\in\mathbb{O}^{m\times\ell}$ such that*

$$\|\sin\Theta(V_0,\bar{V})\|_2 < \frac{\delta}{\tau_2+\chi+\sqrt{2}\chi_2\|B\|_2}, \tag{5.27}$$

the sequence $\{\mathcal{V}_k\}$ converges to $\bar{\mathcal{V}}$ quadratically, *and moreover*

$$\|\sin\Theta(V_{k+1},\bar{V})\|_2 \leqslant \frac{\tau_2}{\delta-(\chi+\sqrt{2}\chi_2\|B\|_2)\|\sin\Theta(V_k,\bar{V})\|_2}\|\sin\Theta(V_k,\bar{V})\|_2^2 \tag{5.28}$$

for $k=0,1,\ldots$

3. *If $R_A=R_B=0$, then for any $V_0\in\mathbb{O}^{m\times\ell}$ satisfying*

$$\|\sin\Theta(V_0,\bar{V})\|_2 < \frac{2\delta}{\chi+\sqrt{\chi^2+4\tau_3\delta}}, \tag{5.29}$$

$\mathcal{V}_1=\bar{\mathcal{V}}$, i.e., convergence is in one step, where

$$\tau_3 := \frac{4\sqrt{2}\|A\|_2\|B\|_2\ell}{\omega_B^2} + \frac{4\sqrt{2}\ell\|B\|_2}{\omega_B^4}[\|A\|_2\Omega_B^2 + 2\Omega_A\Omega_B\|B\|_2]. \tag{5.30}$$

Proof. It suffices to prove (5.26) and (5.28) just for $k=0$ and verify that (5.25) and (5.27) hold with V_0 replaced by V_1.

For clarity, we drop the subscript 0 to V_0 and write V_+ for V_1. Let

$$\epsilon = \|\sin\Theta(V,\bar{V})\|_2, \quad \epsilon_+ = \|\sin\Theta(V_+,\bar{V})\|_2.$$

By Lemma 5.7, we have

$$\epsilon_+ \leqslant \frac{\tau_1\|R_B\|_2+\tau_2\epsilon}{\delta-(\chi+\sqrt{2}\chi_2\|B\|_2)\epsilon}\times\epsilon \tag{5.31}$$

if $\delta > (\chi+\sqrt{2}\chi_2\|B\|_2)\epsilon$. As we will see, this condition is always satisfied under (5.25) or (5.27) in the respective cases.

For item 1, we like to have

$$\begin{aligned}&\frac{\tau_1\|R_B\|_2+\tau_2\epsilon}{\delta-(\chi+\sqrt{2}\chi_2\|B\|_2)\epsilon}\leqslant\nu\\&\Leftrightarrow\quad \tau_1\|R_B\|_2+\tau_2\epsilon\leqslant\nu[\delta-(\chi+\sqrt{2}\chi_2\|B\|_2)\epsilon]\\&\Leftrightarrow\quad \tau_1\|R_B\|_2+[\tau_2+\nu(\chi+\sqrt{2}\chi_2\|B\|_2)]\epsilon\leqslant\nu\delta\\&\Leftarrow\quad \tau_1\|R_B\|_2\leqslant t\nu\delta,\quad [\tau_2+\nu(\chi+\sqrt{2}\chi_2\|B\|_2)]\epsilon\leqslant(1-t)\nu\delta.\end{aligned}\tag{5.32}$$

The two inequalities in (5.32) hold by the assumption $\|R_B\|_2\leqslant t\nu\,\delta/\tau_1$ and (5.25). Also the second inequality in (5.32) implies $\delta>(\chi+\sqrt{2}\chi_2\|B\|_2)\epsilon$.

For item 2, $R_B=0$; so (5.31) becomes

$$\epsilon_+\leqslant\frac{\tau_2\epsilon}{\delta-(\chi+\sqrt{2}\chi_2\|B\|_2)\epsilon}\times\epsilon.\tag{5.33}$$

Now we like to have

$$\begin{aligned}&\frac{\tau_2\epsilon}{\delta-(\chi+\sqrt{2}\chi_2\|B\|_2)\epsilon}<1\\&\Leftrightarrow\quad \tau_2\epsilon<\delta-(\chi+\sqrt{2}\chi_2\|B\|_2)\epsilon\\&\Leftrightarrow\quad [\tau_2+(\chi+\sqrt{2}\chi_2\|B\|_2)]\epsilon<\delta.\end{aligned}\tag{5.34}$$

The inequality in (5.34) is equivalent to the assumption (5.27). Also this inequality implies $\delta>(\chi+\sqrt{2}\chi_2\|B\|_2)\epsilon$.

For item 3, to exploit the additional condition $R_A=0$, we return to (5.20) and (5.21). By (5.7b) and Lemma 5.6 for $H=A$, we have

$$\epsilon_1\leqslant\frac{4\ell\|B\|_2\epsilon}{\omega_B^2}\cdot\sqrt{2}\|A\|_2\epsilon_+\leqslant\frac{4\sqrt{2}\|A\|_2\|B\|_2\ell}{\omega_B^2}\epsilon^2\epsilon_+.\tag{5.35}$$

For ϵ_2, we use (5.8b) and Lemma 5.6 for $H=B$ to get

$$\epsilon_2\leqslant\left|\frac{\phi_A(\bar V)}{[\phi_B(\bar V)]^2}-\frac{\phi_A(V)}{[\phi_B(V)]^2}\right|\sqrt{2}\|B\|_2\epsilon_+\leqslant\frac{4\sqrt{2}\ell\|B\|_2}{\omega_B^4}[\|A\|_2\Omega_B^2+2\Omega_A\Omega_B\|B\|_2]\epsilon^2\epsilon_+.\tag{5.36}$$

Combine (5.20), (5.21), (5.35), and (5.36) to arrive at

$$\epsilon_+\leqslant\frac{\tau_3}{\delta-\chi\epsilon}\epsilon^2\epsilon_+,$$

where τ_3 is defined by (5.30). Equivalently, $\epsilon_+(\delta-\chi\epsilon-\tau_3\epsilon^2)\leqslant 0$, provided $\delta-\chi\epsilon>0$. Thus $\epsilon_+=0$ if $\delta-\chi\epsilon-\tau_3\epsilon^2>0$. This last inequality holds under (5.29) which also ensures $\delta-\chi\epsilon>0$. □

Remark 5.10. The condition (5.24) is rather strong, as the result of a number of upper bound estimations. In actual computations, SCF may still converge even if it fails. However, it is very useful to our understanding of SCF's local convergence behaviors. The condition reveals a remarkable intrinsic connection between the convergence speed of SCF and that $\bar{\mathcal{V}}$ being close to an eigenspace of B. The closer $\bar{\mathcal{V}}$ is to an eigenspace of B, the faster the convergence will be. In fact, when $\bar{\mathcal{V}}$ is an eigenspace of B, the convergence is quadratic. Even more extreme is when $\bar{\mathcal{V}}$ is an eigenspace of both A and B, the convergence is instant for a sufficiently accurate V_0. A very particular case: $AB=BA$ and $CB=BC$ falls into item 3 of Theorem 5.9. Under these assumptions, for any $V\in\mathbb{O}^{m\times\ell}$, $E(V)H=HE(V)$ for $H\in\{A,B\}$ and thus $E(\bar V)\bar V=\bar V M_{\bar V}\Rightarrow E(\bar V)H\bar V=H\bar V M_{\bar V}$ which implies that $H\bar V$ is also an eigenbasis of $E(\bar V)$ corresponding to its ℓ largest eigenvalues. With the aid of $\lambda_\ell(E(\bar V))>\lambda_{\ell+1}(E(\bar V))$, we conclude that for $H\in\{A,B\}$, $\sin\Theta(\bar V,H\bar V)=0\Leftrightarrow R_H=0$.

6 Numerical experiments

We report in this section our numerical experiments on the SCF iteration. We recall Example 3.1, purposely constructed to demonstrate that global convergence cannot be guaranteed in general. But our experience on randomly generated problems has been always fast and globally convergent.

With various pairs (m,ℓ), we have tested the SCF iteration on numerous random symmetric and positive definite matrices A, B, C and random initial V_0 as generated in MATLAB by

$$\texttt{A} = \texttt{randn}(\texttt{m},\texttt{m}); \quad \texttt{A} = \texttt{A}' * \texttt{A}; \quad \texttt{B} = \texttt{randn}(\texttt{m},\texttt{m}); \quad \texttt{B} = \texttt{B}' * \texttt{B};$$
$$\texttt{C} = \texttt{randn}(\texttt{m},\texttt{m}); \quad \texttt{C} = \texttt{C}' * \texttt{C}; \quad \texttt{V}_0 = \texttt{orth}(\texttt{randn}(\texttt{m},\ell),0).$$

We mentioned before that making C positive definite does not loss any generality. The same can be said about A through shifting A to $A+\xi B$ so that it is positive definite and at the same time, the maximizers remain unchanged.

For comparison purpose, we also tested two MATLAB packages for generic Stiefel manifold-based optimization methods for minimizing the objective function [2] $-f(V)$: one is `OptM` [3] [26], and the other is `sg_min` [4] [2, Subsection 9.4] (see also [5]). `OptM` [26] implements a feasible Barzilai-Borwein (BB) method which uses a Crank-Nicolson-like updating scheme to preserve the orthogonality constraints and a curvilinear search with lower per-iteration cost compared to those based on projections and geodesics. The other Stiefel manifold-based optimization package `sg_min` contains four methods, including the Fletcher-Reeves CG iterative search, the Polak-Ribière CG iterative search, the Newton iterative search, and the dog-leg Newton iterative search, all accessible through `sg_min` by

```
[f,V]=sg_min(V0,'frcg','euclidean');
[f,V]=sg_min(V0,'prcg','euclidean');
[f,V]=sg_min(V0,'newton','euclidean');
[f,V]=sg_min(V0,'dog','euclidean');
```

respectively. In order to run `sg_min`, we provide

1. the objective function $-f(V)$ in the MATLAB file `F.m`,
2. the first derivative information in `dF.m`,

$$\texttt{dF}(V) := -\frac{\partial f(V)}{\partial V} = -2E(V)V,$$

3. the second derivative information in `ddF.m`

$$\begin{aligned}\texttt{ddF}(V,X) := \frac{d}{dt}\texttt{dF}(V(t))\bigg|_{t=0} &= -2\{E(V)X + \mathscr{D}E(V)[X]V\}\\ &= -2\{E(V)X + G(V,X)V\},\end{aligned}$$

where $V(t)$ is any smooth curve on $\mathbb{O}^{m\times\ell}$ with $V = V(0)$ and $X = \dot{V}(0) \in \mathfrak{T}_V\mathbb{O}^{m\times\ell}$, and $G(V,X)$ is defined in [32, Theorem 2.2] (see also [31]).

All calculations are carried out in MATLAB 7.13.0 (R2011b) on a MacBook Pro laptop with Intel Core i5@2.50GHz. For the SCF iteration, when $m \leqslant 500$, we use the MATLAB function `eig` to find an orthonormal eigenbasis V_{k+1} of $E(V_k)$ associated with its ℓ largest eigenvalues at Line 2 of Algorithm 3.1, and use `eigs` if $m > 500$. The SCF iteration is terminated whenever the norm of the gradient of $\frac{1}{2}f_{|\mathbb{O}^{m\times\ell}}(V)$ at V_{k+1} satisfies

$$r_k := \|E(V_{k+1})V_{k+1} - V_{k+1}M_{V_{k+1}}\|_2 \leqslant \texttt{tol} = 10^{-8}.$$

[2] Because both `sg_min` and `OptM` *minimize* a function on $\mathbb{O}^{m\times\ell}$, we use $-f(V)$ instead of $f(V)$.

[3] The MATLAB code `OptM` is available at: `http://optman.blogs.rice.edu/`.

[4] The MATLAB code `sg_min` (version 2.4.3) is available at: `http://web.mit.edu/~ripper/www/sgmin.html`.

For `OptM`, we observed numerically that in order to achieve the same small r_k, we need to set `OptM`'s tolerances:

$$\mathrm{tol}_k^x := \frac{\|V_k - V_{k+1}\|_F}{\sqrt{m}} \quad \text{and} \quad \mathrm{tol}_k^f := \frac{|f(V_k) - f(V_{k+1})|}{|f(V_k)| + 1},$$

to about the square of `tol`. This in theory can be explained because the relation (6.1) to be established later suggests

$$f(V_k) - f(V_{k+1}) = \mathcal{O}(r_k^2) + \mathcal{O}(r_{k+1}^2).$$

But we noticed that `OptM` had difficulty with both tol_k^x and tol_k^f set to $\texttt{tol}^2 = 10^{-16}$ which turns out to be too tiny for `OptM` to use. So in our testing, we relaxed the tolerances to $\mathrm{tol}_k^x \leqslant 10^{-8}$, $\mathrm{tol}_k^f \leqslant 10^{-10}$ as the stopping criteria. Finally, for `sg_min`, the default options are used.

For $\ell = 3, 5$ and various m, Tables 1–3 report the numbers of outer iterations, the residuals r_k, and the CPU times (measured by the MATLAB function `cputime`), averaged over 20 random tests, for each method. Similar numerical behaviors are also observed for other ℓ.

Because both `OptM` and `sg_min` are generic black box packages for Riemannian optimization algorithms, it is expected that the customarily designed SCF iteration will outperform them. This is clearly demonstrated in Tables 1–3 in terms of accuracy, the number of iterative steps, and the running time as well. It deserves to point out that SCF shows a remarkable global convergence behavior: we observed that numerically, all the methods converge to the same objective value for each testing problem, but SCF can reach a solution with a residual about 10^{-9} in only about 5 iterations, while `OptM` takes from 14 up to 34 times, and the Fletcher-Reeves CG (`frcg`) takes from 10 up to 28 times, the Polak-Ribière CG (`prcg`) takes from 48 up to 150 times, the Newton iteration (`Newton`) takes from 1.8 up to 2.9 times, and the dog-leg Newton iteration (`dog`) takes from 2.3 up to 4 times as many outer iterations to reduce residuals to only about 10^{-4}.

Our last remark of this section is about the relationship between the accuracy of the objective value $f(V)$ and the residual r_k of the computed solution for each method. Note

$$\begin{aligned}
r_k &= \|E(V_{k+1})V_{k+1} - V_{k+1}M_{V_{k+1}}\|_2 \\
&\leqslant \|E(\bar{V})V_{k+1} - V_{k+1}M_{V_{k+1}}\|_2 + \|[E(V_{k+1}) - E(\bar{V})]V_{k+1}\|_2 \\
&= \|E(\bar{V})V_{k+1} - V_{k+1}M_{V_{k+1}}\|_2 + \mathcal{O}(\epsilon_{k+1}), \\
r_k &\geqslant \|E(\bar{V})V_{k+1} - V_{k+1}M_{V_{k+1}}\|_2 - \|[E(V_{k+1}) - E(\bar{V})]V_{k+1}\|_2 \\
&= \|E(\bar{V})V_{k+1} - V_{k+1}M_{V_{k+1}}\|_2 + \mathcal{O}(\epsilon_{k+1}),
\end{aligned}$$

by Lemma 5.4, where $\epsilon_{k+1} := \sin\Theta(V_{k+1}, \bar{V})$. For sufficiently tiny ϵ_k, if

$$\delta := \lambda_\ell(E(\bar{V})) - \lambda_{\ell+1}(E(\bar{V})) > 0,$$

Table 1 The numbers of outer iterations averaged over 20 random tests

ℓ	m	SCF	sg_min				OptM
			frcg	prcg	Newton	dog	
3	100	5.20	51.80	252.10	9.20	12.10	72.10
	200	5.00	61.10	297.60	9.80	13.20	81.20
	500	4.70	111.12	653.90	11.30	15.70	97.70
	1000	4.60	110.90	875.60	11.30	16.70	123.80
	2000	4.00	110.50	601.50	11.60	16.90	137.70
5	100	5.20	44.40	231.50	9.00	11.70	61.80
	200	4.90	62.60	384.20	9.80	14.00	83.50
	500	4.50	86.60	448.90	10.60	15.10	108.80
	1000	4.00	152.30	1183.50	11.60	17.10	150.20
	2000	4.00	124.80	737.20	11.90	17.40	151.60

Table 2 Residuals r_k averaged over 20 random tests

ℓ	m	SCF	sg_min				OptM
			frcg	prcg	Newton	dog	
3	100	2.0866e−09	9.6073e−05	1.0243e−04	7.2286e−05	1.0136e−04	3.9433e−05
	200	2.8450e−09	2.0088e−04	2.1039e−04	1.5021e−04	2.1498e−04	3.1963e−04
	500	1.2366e−09	5.1486e−04	5.1702e−04	3.9649e−04	5.9318e−04	6.7840e−04
	1000	2.8957e−09	1.0428e−03	1.0215e−03	8.1522e−04	1.2606e−03	1.7737e−04
	2000	1.8831e−09	2.0440e−03	2.0367e−03	1.7409e−03	2.6061e−03	4.7989e−04
5	100	3.5332e−09	1.0265e−04	1.0983e−04	7.5412e−05	1.2384e−04	5.6502e−05
	200	9.9221e−10	2.3758e−04	2.1842e−04	1.7443e−04	2.5598e−04	2.1490e−04
	500	3.6224e−09	5.7844e−04	5.3591e−04	5.0324e−04	7.8546e−04	8.0341e−04
	1000	2.1719e−09	1.2014e−03	1.0986e−03	8.9469e−03	1.6402e−03	1.5912e−04
	2000	4.1370e−10	2.4211e−03	2.0863e−03	1.7516e−03	3.3055e−03	4.2530e−04

Table 3 CPU time averaged over 20 random tests

ℓ	m	SCF	sg_min				OptM
			frcg	prcg	Newton	dog	
3	100	0.0260	0.6730	2.5400	0.5740	1.0710	0.0690
	200	0.0970	1.8519	6.2239	1.2130	3.4610	0.2080
	500	0.8360	10.7350	48.6730	6.5850	38.2220	1.4190
	1000	1.4820	46.8770	219.1930	30.9820	261.3340	6.5720
	2000	4.4930	167.3610	616.7079	132.1800	1861.3160	28.3600
5	100	0.0240	0.7890	2.9840	0.6530	1.1410	0.1100
	200	0.1040	2.0870	8.8760	1.4710	4.4920	0.2770
	500	0.8000	12.6340	43.1360	8.2330	45.9890	1.8090
	1000	1.3070	77.9170	431.3919	38.2249	316.8010	9.1970
	2000	4.3160	220.5770	876.4220	159.2719	2134.8820	34.2160

then by Davis-Kahan $\sin\theta$ theorem [3],

$$\epsilon_{k+1} \leqslant \frac{1}{\delta}\|E(\bar{V})V_{k+1} - V_{k+1}M_{V_{k+1}}\|_2 + \mathcal{O}(\epsilon_{k+1});$$

moreover, by using Lemma 4.1 and the facts that $f(\bar{V}) = f(\bar{V}Z)$ and $E(\bar{V}) = E(\bar{V}Z)$ for any orthogonal matrix $Z \in \mathbb{R}^{\ell\times\ell}$, it is true that

$$\|E(\bar{V})V_{k+1} - V_{k+1}M_{V_{k+1}}\|_2 = \mathcal{O}(\epsilon_{k+1}).$$

Hence roughly speaking, r_k and ϵ_k are about in the same order of magnitude. Since the gradient of $f(V)$ at $\bar{V}$ vanishes, we have

$$f(V_k) = f(\bar{V}) + \mathcal{O}(\epsilon_k^2) = f(\bar{V}) + \mathcal{O}(r_k^2). \tag{6.1}$$

This explains well Table 4 in which the objective values $f(V_k)$ by Stiefel manifold-based optimization methods match the respective ones by SCF to about 9 to 11 decimal digits, even though the residuals by the former methods are about the square roots of the ones by SCF.

In the comparisons so far, although both `OptM` and `sg_min` didn't perform as well as our SCF on the maximization problem (1.1), we point out that the methods in the two packages are not limited to (1.1) and have wider applicability, and can succeed on problems that are difficult for our SCF. One of these problems is Example 3.1 for which our SCF comes close to an optimum and then starts to oscillate while both `OptM` and `sg_min` seem to be able to make progress towards an optimum.

Table 4 The computed objective values at convergence of a typical test problem with $\ell = 5$

m	SCF	sg_min				OptM
		frcg	prcg	Newton	dog	
100	1.732631226491e–3	1.732631226491e–3	1.732631226491e–3	1.732631226491e–3	1.732631226489e–3	1.732631226491e–3
200	3.638994886711e–3	3.638994886709e–3	3.638994886710e–3	3.638994886709e–3	3.638994886705e–3	3.638994886710e–3
500	9.505544535005e–3	9.505544534995e–3	9.505544535000e–3	9.505544535002e–3	9.505544534946e–3	9.505544534980e–3
1000	1.939209465567e–4	1.939209465562e–4	1.939209465565e–4	1.939209465566e–4	1.939209465553e–4	1.939209465555e–4
2000	3.930154578537e–4	3.930154578533e–4	3.930154578536e–4	3.930154578534e–4	3.930154578504e–4	3.930154578514e–4

7 Concluding remarks and future research

Based on a theoretical result in [32], a self-consistent-field (SCF) iteration for solving the maximization problem (1.1) is proposed and analyzed in detail. For the special case $C = \eta B$ ($\eta \geqslant 0$), it is proved that the SCF iteration is globally convergent, but in general global convergence is not guaranteed. However, various local convergence results for the general case are obtained. Our numerical tests (not all are reported here) suggest that the method is very efficient and superior to generic Stiefel manifold-based optimization methods when it works and it usually does.

Still there are certain theoretical and numerical issues that remain unanswered for the SCF iteration. These issues include (1) sufficient conditions for the global maximizers, (2) further convergence analysis for the SCF iteration (like Algorithm 3.1 but for the more general case), and (3) some modifications, if any, of the SCF iteration to ensure its global convergence. They, among others, will be the subjects of our future research.

Acknowledgements The first author was supported by National Natural Science Foundation of China (Grant Nos. 11101257 and 11371102), and the Basic Academic Discipline Program, the 11th Five Year Plan of 211 Project for Shanghai University of Finance and Economics. Part of this work was done while the first author was a visiting scholar at the Department of Mathematics, University of Texas at Arlington from February 2013 to January 2014. The second author was supported by National Science Foundation of USA (Grant Nos. 1115834 and 1317330), and a Research Gift Grant from Intel Corporation. The authors are grateful to two anonymous referees for their careful reading and helpful comments and suggestions.

References

1 Absil P A, Mahony R, Sepulchre R. Optimization Algorithms On Matrix Manifolds. Princeton: Princeton University Press, 2008
2 Bai Z, Demmel J, Dongarra J, et al. eds. Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide. Philadelphia: SIAM, 2000
3 Davis C, Kahan W. The rotation of eigenvectors by a perturbation, III. SIAM J Numer Anal, 1970, 7: 1–46
4 Demmel J. Applied Numerical Linear Algebra. Philadelphia: SIAM, 1997
5 Edelman A, Arias T A, Smith S T. The geometry of algorithms with orthogonality constraints. SIAM J Matrix Anal Appl, 1999, 20: 303–353
6 Golub G H, Van Loan C F. Matrix Computations 3rd ed. Baltimore: Johns Hopkins University Press, 1996
7 Horn R A, Johnson C R. Topics in Matrix Analysis. Cambridge: Cambridge University Press, 1991
8 Knyazev A V. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. SIAM J Sci Comput, 2001, 23: 517–541
9 Knyazev A V, Argentati M E. Rayleigh-Ritz majorization error bounds with applications to FEM. SIAM J Matrix Anal Appl, 2010, 31: 1521–1537
10 Liu X, Wang X, Wen Z, et al. On the convergence of the self-consistent field iteration in Kohn-Sham density functional theory. ArXiv:1302.6022, 2013
11 Martin R M. Electronic Structure: Basic Theory and Practical Methods. Cambridge: Cambridge University Press,

2004
12 Milnor J W, Stasheff J D. Characteristic Classes. Princeton/Tokyo: Princeton University Press & University of Tokyo Press, 1974
13 Ngo T, Bellalij M, Saad Y. The trace ratio optimization problem for dimensionality reduction. SIAM J Matrix Anal Appl, 2010, 31: 2950–2971
14 Nocedal J, Wright S. Numerical Optimization, 2nd ed. New York: Springer, 2006
15 Parlett B N. The Symmetric Eigenvalue Problem. Philadelphia: SIAM, 1998
16 Saad Y. Numerical Methods for Large Eigenvalue Problems. Manchester: Manchester University Press, 1992
17 Saad Y, Chelikowsky J R, Shontz S M. Numerical methods for electronic structure calculations of materials. SIAM Rev, 2010, 52: 3–54
18 Saunders V R, Hillier I H. A "level-shifting" method for converging closed shell Hartree-Fock wave functions. Internat J Quantum Chem, 1973, 7: 699–705
19 Sleijpen G L G, van der Vorst H A. A Jacobi-Davidson iteration method for linear eigenvalue problems. SIAM J Matrix Anal Appl, 1996, 17: 401–425
20 Stewart G W, Sun J G. Matrix Perturbation Theory. Boston: Academic Press, 1990
21 Szabo A, Ostlund N S. Modern Quantum Chemistry: An Introduction to Advanced Electronic Structure Theory. New York: Dover Publications, 1996
22 Thøgersen L, Olsen J, Yeager D, et al. The trust-region self-consistent field method: Towards a black-box optimization in Hartree-Fock and Kohn-Sham theories. J Chem Phys, 2004, 121: 16–27
23 Wang H, Yan S, Xu D, et al. Trace ratio vs. ratio trace for dimensionality reduction. In: IEEE Conference on Computer Vision and Pattern Recognition, 2007, CVPR'07, 1–8
24 Wedin P Å. On angles between subspaces. In: Kågström B, Ruhe A, eds. Matrix Pencils. New York: Springer, 1983, 263–285
25 Wen Z, Milzarek A, Ulbrich M, et al. Adaptive regularized self-consistent field iteration with exact Hessian for electronic structure calculation. SIAM J Sci Comput, 2013, 35: A1299–A1324
26 Wen Z, Yin W. A feasible method for optimization with orthogonality constraints. Math Programming, 2013, 142: 397–434
27 Yang C, Gao W, Meza J C. On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems. SIAM J Matrix Anal Appl, 2009, 30: 1773–1788
28 Yang C, Meza J C, Lee B, et al. KSSOLV – a MATLAB toolbox for solving the Kohn-Sham equations. ACM Trans Math Softw, 2009, 36: 1–35
29 Yang C, Meza J C, Wang L W. A trust region direct constrained minimization algorithm for the Kohn-Sham equation. SIAM J Sci Comput, 2007, 29: 1854–1875
30 Zhang L H. On optimizing the sum of the Rayleigh quotient and the generalized Rayleigh quotient on the unit sphere. Comput Opt Appl, 2013, 54: 111–139
31 Zhang L H, Li R C. Maximization of the sum of the trace ratio on the Stiefel manifold. Technical Report 2013-04. Department of Mathematics, University of Texas at Arlington, May 2013, http://www.uta.edu/math/preprint/
32 Zhang L H, Li R C. Maximization of the sum of the trace ratio on the Stiefel manifold, I: Theory. Sci China Math, 2014, 57: 2495–2508
33 Zhang L H, Liao L Z, Ng M K. Fast algorithms for the generalized Foley-Sammon discriminant analysis. SIAM J Matrix Anal Appl, 2010, 31: 1584–1605
34 Zhang L H, Liao L Z, Ng M K. Superlinear convergence of a general algorithm for the generalized Foley-Sammon discriminant analysis. J Optim Theory Appl, 2013, 157: 853–865
35 Zhang L H, Yang W, Liao L Z. A note on the trace quotient problem. Optim Lett, 2014, 8: 1637–1645
36 Zhang X, Zhu J, Wen Z, Zhou A. Gradient type optimization methods for electronic structure calculations. ArXiv:1308.2864, 2013