

# APPROXIMATE MATRIX AND TENSOR DIAGONALIZATION BY UNITARY TRANSFORMATIONS: CONVERGENCE OF JACOBI-TYPE ALGORITHMS\*

KONSTANTIN USEVICH<sup>†</sup>, JIANZE LI<sup>‡</sup>, AND PIERRE COMON<sup>§</sup>

**Abstract.** We propose a gradient-based Jacobi algorithm for a class of maximization problems on the unitary group, with a focus on approximate diagonalization of complex matrices and tensors by unitary transformations. We provide weak convergence results, and prove local linear convergence of this algorithm. The convergence results also apply to the case of real-valued tensors.

**Key words.** optimization on manifolds, unitary group, Givens rotations, approximate tensor diagonalization, Lojasiewicz gradient inequality, local convergence

**AMS subject classifications.** 90C30,53B21,53B20,15A69,65K10,65Y20

**1. Introduction.** In this paper, we consider the following optimization problem

$$(1.1) \quad U_* = \arg \max_{U \in \mathcal{U}_n} f(U),$$

where  $\mathcal{U}_n$  is the unitary group and  $f : \mathcal{U}_n \rightarrow \mathbb{R}$  is real differentiable. An important class of such problems stems from approximate matrix and tensor diagonalization in numerical linear algebra [13], signal processing [18] and machine learning [5].

Jacobi-type algorithms are widely used for maximization of these cost functions. Inspired by the classic Jacobi algorithm [24] for the symmetric eigenvalue problem, they proceed by successive Givens rotations that update only a pair of columns of  $U$ . The popularity of these approaches is explained by low computational complexity of the updates. Despite their popularity, their convergence has not yet been studied thoroughly, except the case of matrices [24] and a pair of commuting matrices [13].

For tensor problems in the real-valued case (orthogonal group), a gradient-based Jacobi-type algorithm was proposed in [32], and its weak convergence<sup>1</sup> was proved<sup>2</sup>. In [38], its global (single-point) convergence<sup>3</sup> for joint real 3rd order tensor or matrix diagonalization was proved. The proof in [38] based on the Lojasiewicz gradient inequality, a popular tool for studying convergence properties of nonlinear optimization algorithms [2, 37, 47, 6], including various tensor approximation problems [48, 31].

In this paper, we address the complex-valued case (1.1), and focus on tensor and matrix approximate diagonalization problems. Unlike the real case, where the Givens rotations are univariate (“line-search” type), in the complex case the updates correspond to maximization on a sphere (similar in spirit to subspace methods). The main

---

\*Submitted to the editors DATE.

**Funding:** This work was funded in part by the ERC project “DECODA” no.320594, in the frame of the European program FP7/2007-2013, by the National Natural Science Foundation of China (No.11601371), and by the Agence Nationale de Recherche (ANR grant LeaFleT, ANR-19-CE23-0021).

<sup>†</sup>Université de Lorraine, CNRS, CRAN, Nancy, France ([konstantin.usevich@univ-lorraine.fr](mailto:konstantin.usevich@univ-lorraine.fr)).

<sup>‡</sup>Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China ([lijianze@gmail.com](mailto:lijianze@gmail.com)).

<sup>§</sup>Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-Lab, France ([pierre.comon@gipsa-lab.fr](mailto:pierre.comon@gipsa-lab.fr)).

<sup>1</sup>every accumulation point is a stationary point.

<sup>2</sup>The algorithm [32] was proposed for a particular problem of Tucker approximation, but the convergence result of [32] are valid for arbitrary smooth functions, see discussion in [38].

<sup>3</sup>i.e., for any starting point, the iterations converge to a single limit point. Note that global convergence does not imply convergence to a global minimum; also, the notion of “global convergence” often has a different meaning in the numerical linear algebra community [27].

contributions of the paper are: (i) we generalize the algorithm of [32] to the complex case, prove its weak convergence, and find global rates of convergence based on the results of [9]; (ii) we show that the local convergence can be studied by combining the tools of Lojasiewicz gradient inequality, geodesic convexity and recent results on Lojasiewicz exponent for Morse-Bott functions. In particular, local linear convergence holds for local maxima satisfying second order regularity conditions. One of the motivations for this work was that the case of the unitary group is not common in the optimization literature, unlike the orthogonal group or other matrix manifolds [3].

The structure of the paper is as follows. In Section 2, we recall the cost functions of interest, the principle of Jacobi-type algorithms, present the gradient-based algorithm and a summary of main results. Section 3, contains all necessary facts for differentiation on the unitary group. Section 4 contains expressions for the first- and second-order derivatives, as well as expressions for Jacobi rotations for cost functions of interest. In Section 5, we present the results on weak convergence and global convergence rates. The results of [9] are summarized in the same section. In Section 6, we recall results based on Lojasiewicz gradient inequality, and facts on Morse-Bott functions. Section 7 contains main results and lemmas.

## 2. Background, problem statement, and summary of results.

**2.1. Main notation.** For an  $\mathbf{X} \in \mathbb{C}^{m \times n}$ , we denote by  $\mathbf{X}^*$  its elementwise conjugate, and by  $\mathbf{X}^\top$ ,  $\mathbf{X}^\mathbb{H}$  its transpose and Hermitian transpose. We use the following notation for the real and imaginary parts  $\mathbf{X} = \mathbf{X}^\Re + i\mathbf{X}^\Im$  of matrices, and  $\Re(z)$ ,  $\Im(z)$  for  $z \in \mathbb{C}$ . Let  $\mathbb{T} \subset \mathbb{C}$  be the unit circle, and  $\mathcal{U}_n \subset \mathbb{C}^{n \times n}$  be the unitary group.

In this paper, we make no distinction between tensors and multi-way arrays; for simplicity, we consider only fully contravariant tensors [42]. For a tensor or a matrix  $\mathcal{A} \in \mathbb{C}^{n \times \dots \times n}$ , we denote by  $\text{diag}\{\mathcal{A}\} \in \mathbb{C}^n$  the vector of all the diagonal elements  $\mathcal{A}_{i_i \dots i_i}$  and by  $\text{tr}\{\mathcal{A}\}$  the sum of the diagonal elements. We denote by  $\|\cdot\|$  the Frobenius norm of a tensor/matrix, or the Euclidean norm of a vector. For a  $d$ -th order tensor  $\mathcal{A} \in \mathbb{C}^{n \times \dots \times n}$  its contraction on the  $k$ th index with  $\mathbf{v} \in \mathbb{C}^n$  (resp.  $\mathbf{M} \in \mathbb{C}^{m \times n}$ ) is

$$(\mathcal{A} \bullet_k \mathbf{v})_{i_1 \dots \cancel{j_k} \dots i_d} \stackrel{\text{def}}{=} \sum_{j=1}^n \mathcal{A}_{i_1 \dots i_{k-1} j i_{k+1} \dots i_d} v_j, \quad (\mathcal{A} \bullet_k \mathbf{M})_{i_1 \dots i_d} \stackrel{\text{def}}{=} \sum_{j=1}^n \mathcal{A}_{i_1 \dots i_{k-1} j i_{k+1} \dots i_d} M_{i_k, j}.$$

By writing multiple contractions  $\mathcal{A} \bullet_{k_1} \mathbf{v}_1 \dots \bullet_{k_\ell} \mathbf{v}_\ell$  we assume that they are performed simultaneously, i.e., the indexing of the tensor does not change before contractions are complete. For a matrix  $\mathbf{S} \in \mathbb{C}^{n \times n}$ , we will also denote the double contraction as

$$(\mathcal{A} \bullet_{k, \ell} \mathbf{S})_{i_1 \dots \cancel{j_k} \dots \cancel{j_\ell} \dots i_d} \stackrel{\text{def}}{=} \sum_{j, s=1}^{n, n} \mathcal{A}_{i_1 \dots i_{k-1} j i_{k+1} \dots i_{\ell-1} s i_{\ell+1} \dots i_d} S_{j, s}$$

For a matrix  $\mathbf{U} \in \mathbb{C}^{n \times n}$ , we will denote its columns as  $\mathbf{U} = [\mathbf{u}_1 \ \dots \ \mathbf{u}_n]$ .

**2.2. Motivation.** This paper is motivated by following maximization problems:

(i) joint *approximate Hermitian diagonalization of matrices*  $\mathbf{A}^{(\ell)} \in \mathbb{C}^{n \times n}$ ,  $1 \leq \ell \leq L$ :

$$(2.1) \quad f(\mathbf{U}) = \sum_{\ell=1}^L \|\text{diag}\{\mathbf{U}^\mathbb{H} \mathbf{A}^{(\ell)} \mathbf{U}\}\|^2 = \sum_{\ell=1}^L \sum_{p=1}^n |\mathbf{u}_p^\mathbb{H} \mathbf{A}^{(\ell)} \mathbf{u}_p|^2;$$

(ii) *approximate diagonalization of a 3rd order tensor*  $\mathcal{A} \in \mathbb{C}^{n \times n \times n}$ :

$$(2.2) \quad f(\mathbf{U}) = \|\text{diag}\{\mathcal{A} \bullet_1 \mathbf{U}^\mathbb{H} \bullet_2 \mathbf{U}^\top \bullet_3 \mathbf{U}^\top\}\|^2 = \sum_{p=1}^n |\mathcal{A} \bullet_1 \mathbf{u}_p^* \bullet_2 \mathbf{u}_p \bullet_3 \mathbf{u}_p|^2;$$

- (iii) *approximate diagonalization of a 4th order tensor*  $\mathcal{B} \in \mathbb{C}^{n \times n \times n \times n}$  satisfying a Hermitian symmetry condition  $\mathcal{B}_{ijkl} = \mathcal{B}_{klij}^*$  for any  $1 \leq i, j, k, l \leq n$ :

$$(2.3) \quad f(\mathbf{U}) = \text{tr}\{\mathcal{B} \bullet_1 \mathbf{U}^H \bullet_2 \mathbf{U}^H \bullet_3 \mathbf{U}^T \bullet_4 \mathbf{U}^T\} = \sum_{p=1}^n \mathcal{B} \bullet_1 \mathbf{u}_p^* \bullet_2 \mathbf{u}_p^* \bullet_3 \mathbf{u}_p \bullet_4 \mathbf{u}_p.$$

Such maximization problems appear in blind source separation [18] in the context of:

- (i) joint diagonalization of covariance matrices [14, 15];
- (ii) diagonalization of the cumulant tensor [19] with  $\mathcal{A}_{ijk} = \text{Cum}(v_i, v_j, v_k^*)$ ;
- (iii) diagonalization of the cumulant tensor [16]  $\mathcal{A}_{ijkl} = \text{Cum}(v_i, v_j, v_k^*, v_l^*)$  of a complex random vector  $\mathbf{v}$ , which may itself stem from a Fourier transform [22].

*Remark 2.1.* Due to invariance of  $\|\cdot\|$  to unitary transformations, maximizing (2.1) or (2.2) is equivalent to minimizing sums of squares of the off-diagonal elements of the rotated tensors/matrices, hence the name “approximate diagonalization”. For example, in the single matrix case (i.e., (2.1) and  $L = 1$ ), we can equivalently minimize the squared norm of the off-diagonal elements (so called off-norm)

$$(2.4) \quad \|\text{off}(\mathbf{U}^H \mathbf{A} \mathbf{U})\|^2 = \|\mathbf{A}\|^2 - \|\text{diag}\{\mathbf{U}^H \mathbf{A} \mathbf{U}\}\|^2,$$

which is typically done in the numerical linear algebra community [24].

In this paper, we consider a class of functions that generalizes<sup>4</sup> (2.1)–(2.3). For a set of tensors  $\mathcal{A}^{(1)}, \dots, \mathcal{A}^{(L)}$  of orders  $d_1, \dots, d_L$  (potentially different), integers  $t_\ell$ ,  $0 \leq t_\ell \leq d_\ell$ , and  $\alpha_\ell \in \mathbb{R}$  (possibly negative), we define the cost function as

$$(2.5) \quad f(\mathbf{U}) = \sum_{\ell=1}^L \alpha_\ell \|\text{diag}\{\mathcal{A}^{(\ell)} \bullet_1 \mathbf{U}^H \dots \bullet_{t_\ell} \mathbf{U}^H \bullet_{t_\ell+1} \mathbf{U}^T \dots \bullet_{d_\ell} \mathbf{U}^T\}\|^2,$$

i.e., a conjugate transformation is applied  $t_\ell$  times and a non-conjugate  $d_\ell - t_\ell$  times. If all  $\alpha_k > 0$ , maximization of (2.5) can be viewed as joint diagonalization of several tensors (as in Remark 2.1); the general case of negative  $\alpha_k$  allows for more flexibility. Also, (2.5) includes symmetric diagonalization problems (without conjugations), e.g.,

$$f(\mathbf{U}) = \sum_{\ell=1}^L \|\text{diag}\{\mathbf{U}^T \mathbf{A}^{(\ell)} \mathbf{U}\}\|^2, \quad \text{or } f(\mathbf{U}) = \|\text{diag}\{\mathcal{A} \bullet_1 \mathbf{U}^T \bullet_2 \mathbf{U}^T \bullet_3 \mathbf{U}^T\}\|^2;$$

It can be shown that  $f$  admits representation (2.5) (with  $d = \max(d_1, \dots, d_L)$ ) if and only if there exists a  $2d$ -th order tensor  $\mathcal{B}$  that is Hermitian [44], i.e.,

$$(2.6) \quad \mathcal{B}_{i_1 \dots i_d j_1 \dots j_d} = \mathcal{B}_{j_1 \dots j_d i_1 \dots i_d}^*$$

such that  $f$  has a representation which generalizes (2.3):

$$(2.7) \quad f(\mathbf{U}) = \text{tr}\{\mathcal{B} \bullet_1 \mathbf{U}^H \dots \bullet_d \mathbf{U}^H \bullet_{d+1} \mathbf{U}^T \dots \bullet_{2d} \mathbf{U}^T\}.$$

The equivalence between (2.5) and (2.7) is analogous to the spectral theorem for Hermitian matrices; a proof can be found in Section 4 (see also [33, Prop. 3.5]).

<sup>4</sup>It is easy to see that (2.5) generalizes (2.1) (for  $d_1 = \dots = d_L = 2, t_1 = \dots = t_L = 1$ ) and (2.2) (for  $L = 1, d_1 = 3, t_1 = 1$ ).



The convergence of the iterations for cyclic Jacobi algorithms is unknown, except in the single matrix case<sup>6</sup> [24]. Most of the results for the matrix case are on the convergence of  $f(\mathbf{U}_k)$  to  $\|\mathbf{A}\|^2$  (or the off-norm (2.4) to zero). The rate is linear and asymptotically quadratic, for the cyclic strategies of choice of pairs and a class of other strategies, see [24, §8.4.3] and [26, 27] for an overview. Moreover, the result [40] guarantees that in this case  $\mathbf{U}_k^H \mathbf{A} \mathbf{U}_k$  converges to a diagonal matrix. However, this implies convergence of  $\mathbf{U}_k$  to a limit point only if the eigenvalues of  $\mathbf{A}$  are distinct (for multiple eigenvalues, convergence of subspaces is proved [20]). All these results are specific to matrices, and cannot be directly applied to our case. Finally, note that an extension of the Jacobi algorithm to compact Lie groups was proposed in [34], but their setup is different: it is the notion of diagonality of a matrix that is generalized to Lie groups in [34], while we consider higher-order cost functions.

**2.4. Jacobi-G algorithm and an overview of results.** Recently, a gradient-based Jacobi algorithm (Jacobi-G) was proposed [32] in a context of optimization on orthogonal group. Its weak convergence was shown in [32] and global convergence for real matrix and 3rd order tensor case was proved in [38]. In this subsection, we introduce a complex generalization of the Jacobi-G algorithm (Algorithm 2.1). The main idea behind the algorithm is to choose Givens transformations that are well aligned with the Riemannian gradient<sup>7</sup> of  $f$  denoted by  $\text{grad } f(\cdot)$ .

---

**Algorithm 2.1** General Jacobi-G algorithm

---

**Input:** A differentiable  $f : \mathcal{U}_n \rightarrow \mathbb{R}$ , constant  $0 < \delta \leq \sqrt{2}/n$ , starting point  $\mathbf{U}_0$ .

**Output:** Sequence of iterations  $\mathbf{U}_k$ .

- **For**  $k = 1, 2, \dots$  until a stopping criterion is satisfied do
- Choose an index pair  $(i_k, j_k)$  satisfying

$$(2.13) \quad \|\text{grad } h_{(i_k, j_k), \mathbf{U}_{k-1}}(\mathbf{I}_2)\| \geq \delta \|\text{grad } f(\mathbf{U}_{k-1})\|.$$

- Find  $\Psi_k$  that maximizes  $h_k(\Psi) \stackrel{\text{def}}{=} h_{(i_k, j_k), \mathbf{U}_{k-1}}(\Psi)$ .
  - Update  $\mathbf{U}_k = \mathbf{U}_{k-1} \mathbf{G}^{(i_k, j_k, \Psi_k)}$ .
  - **End for**
- 

It is shown in Section 4 that it is always possible to find  $(i_k, j_k)$  satisfying (2.13), provided  $\delta \leq \sqrt{2}/n$  (the meaning of  $\delta$  will be also explained). Next, we summarize main results on convergence of Algorithm 2.1 for (2.5) and (2.7),  $d \leq 3$ .

- **Proposition 5.5:** we show that, similarly to the algorithm of [32], the weak convergence takes place ( $\text{grad } f(\mathbf{U}_k) \rightarrow 0$ ), which implies that every accumulation point  $\bar{\mathbf{U}}$  of the sequence  $\{\mathbf{U}_k\}$  is a stationary point; moreover, we are able to retrieve global convergence rates along the lines of [9].
- **Theorem 7.4:** if an accumulation point  $\bar{\mathbf{U}}$  satisfies regularity conditions (i.e., restrictions  $h_{(i,j), \bar{\mathbf{U}}}$ , for all  $i < j$ , have semi-strict local maxima at  $\mathbf{I}_2$ ), then  $\bar{\mathbf{U}}$  is the only limit point of  $\{\mathbf{U}_k\}$ ; if in addition, the rank of the Hessian at  $\bar{\mathbf{U}}$  is maximal (i.e., equal to  $n(n-1)$ ), then the speed of convergence is linear.
- **Theorem 7.5:** if  $\mathbf{U}_*$  is a semi-strict local maximum of  $f$ , then Algorithm 2.1 converges linearly to  $\mathbf{U}_*$  (or an equivalent point) when started at any point

---

<sup>6</sup>or a similar case of a pair of commuting matrices [13]. These cases are special because, the matrices can be always diagonalized (the minimal value of the off-norm is zero).

<sup>7</sup>The definition of Riemannian gradient is postponed to Section 3.

in a neighborhood of  $\mathbf{U}_*$ .

We eventually provide in [Subsection 7.3](#) examples of tensor and matrix diagonalization problems where the regularity conditions are satisfied. In the results listed above, we use the notion of semi-strict local maximum due to invariance of the cost function with respect to (2.11). This makes the Riemannian Hessian rank-deficient (rank at most  $n(n-1)$ ) at any stationary point, hence the maxima cannot be strict. We use the following tools to overcome this issue:

- Morse-Bott property that generalizes Morse property at a stationary point;
- quotient manifold  $\tilde{\mathcal{U}}_n$ : factorizing  $\mathcal{U}_n$  by the equivalence relation in (2.11).

**3. Unitary group as a real manifold.** This section contains all necessary facts about the unitary group, derivatives of the cost functions, geodesics, etc.

**3.1. Wirtinger calculus.** First, we introduce the following real-valued inner product<sup>8</sup> on  $\mathbb{C}^{m \times n}$ . For  $\mathbf{X} = \mathbf{X}^{\Re} + i\mathbf{X}^{\Im}, \mathbf{Y} = \mathbf{Y}^{\Re} + i\mathbf{Y}^{\Im} \in \mathbb{C}^{m \times n}$ , we denote

$$(3.1) \quad \langle \mathbf{X}, \mathbf{Y} \rangle_{\Re} \stackrel{\text{def}}{=} \langle \mathbf{X}^{\Re}, \mathbf{Y}^{\Re} \rangle + \langle \mathbf{X}^{\Im}, \mathbf{Y}^{\Im} \rangle = \Re(\text{tr}\{\mathbf{X}^H \mathbf{Y}\}).$$

This makes  $\mathbb{C}^{m \times n}$  a real Euclidean space of dimension  $2mn$ .

Note that a function  $f : \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$  is never holomorphic unless it is constant; therefore we do not require  $f$  to be complex differentiable, but differentiable with respect to the real and imaginary parts. We use a shorthand notation  $\nabla_{\mathbf{X}^{\Re}} f, \nabla_{\mathbf{X}^{\Im}} f \in \mathbb{R}^{m \times n}$  for matrix derivatives with respect to real and imaginary parts of  $\mathbf{X} \in \mathbb{C}^{m \times n}$ . The Wirtinger derivatives  $\nabla_{\mathbf{X}} f, \nabla_{\mathbf{X}^*} f \in \mathbb{C}^{m \times n}$  are standardly defined [1, 12, 36] as

$$\nabla_{\mathbf{X}} f := \frac{1}{2} (\nabla_{\mathbf{X}^{\Re}} f - i \nabla_{\mathbf{X}^{\Im}} f), \quad \nabla_{\mathbf{X}^*} f := \frac{1}{2} (\nabla_{\mathbf{X}^{\Re}} f + i \nabla_{\mathbf{X}^{\Im}} f).$$

The matrix Euclidean gradient of  $f$  with respect to the inner product (3.1) becomes

$$\nabla^{(\Re)} f(\mathbf{X}) = \nabla_{\mathbf{X}^{\Re}} f + i \nabla_{\mathbf{X}^{\Im}} f = 2 \nabla_{\mathbf{X}^*} f(\mathbf{X}).$$

**3.2. Riemannian gradient.**  $\mathcal{U}_n$  can be viewed as an embedded real submanifold of  $\mathbb{C}^{n \times n}$  (see also [25, Appendix C.2.6]). By [3, §3.5.7], the tangent space to  $\mathcal{U}_n$  is associated with an  $n^2$ -dimensional  $\mathbb{R}$ -linear subspace of  $\mathbb{C}^{n \times n}$ :

$$\mathbf{T}_U \mathcal{U}_n = \{\mathbf{X} \in \mathbb{C}^{n \times n} : \mathbf{X}^H \mathbf{U} + \mathbf{U}^H \mathbf{X} = 0\} = \{\mathbf{X} \in \mathbb{C}^{n \times n} : \mathbf{X} = \mathbf{U} \mathbf{Z}, \quad \mathbf{Z} + \mathbf{Z}^H = 0\}.$$

Recall that  $\mathcal{U}_n$  is a matrix Lie group, with the Lie algebra of skew-Hermitian matrices  $\mathfrak{u}(n) = \{\mathbf{Z} \in \mathbb{C}^{n \times n} : \mathbf{Z} + \mathbf{Z}^H = 0\}$  (which coincides with  $\mathbf{T}_{\mathbf{I}_n} \mathcal{U}_n$  in our notation). Then for  $f : \mathbb{C}^{n \times n} \rightarrow \mathbb{R}$  differentiable in a neighborhood of  $\mathcal{U}_n$ , the Riemannian gradient is just the orthogonal projection of  $\nabla^{(\Re)} f(\mathbf{U})$  on  $\mathbf{T}_U \mathcal{U}_n$ :

$$(3.2) \quad \text{grad } f(\mathbf{U}) = \mathbf{U} \mathbf{\Lambda}(\mathbf{U}) \in \mathbf{T}_U \mathcal{U}_n, \quad \text{where}$$

$$(3.3) \quad \mathbf{\Lambda}(\mathbf{U}) = \frac{\mathbf{U}^H \nabla^{(\Re)} f(\mathbf{U}) - (\nabla^{(\Re)} f(\mathbf{U}))^H \mathbf{U}}{2} = \mathbf{U}^H \nabla_{\mathbf{U}^*} f(\mathbf{U}) - (\nabla_{\mathbf{U}^*} f(\mathbf{U}))^H \mathbf{U}.$$

Note that  $\mathbf{\Lambda}(\mathbf{U})$  is a skew-Hermitian matrix, i.e.,

$$(3.4) \quad \Lambda_{ij}(\mathbf{U}) = -(\Lambda_{ji}(\mathbf{U}))^*, \quad 1 \leq i, j \leq n.$$

<sup>8</sup>In some literature [1], a different inner product  $\frac{1}{2} \Re(\text{tr}\{\mathbf{X}^H \mathbf{Y}\})$  is adopted. We prefer a definition that is compatible with the Frobenius norm  $\langle \mathbf{X}, \mathbf{X} \rangle_{\Re} = \|\mathbf{X}\|^2$

In what follows, we will use the exponential map [3, p.102]  $\text{Exp}_{\mathbf{U}} : \mathbf{T}_{\mathbf{U}}\mathcal{U}_n \rightarrow \mathcal{U}_n$ , which maps 1-dimensional lines in the tangent space to geodesics and is given by

$$(3.5) \quad \text{Exp}_{\mathbf{U}}(\mathbf{U}\Omega) = \mathbf{U} \exp(\Omega),$$

where  $\exp(\cdot)$  is the matrix exponential. We will frequently use the following relation between  $\text{Exp}_{\mathbf{U}}$  and the Riemannian gradient. For any  $\Delta \in \mathbf{T}_{\mathbf{U}}\mathcal{U}_n$ , we have

$$(3.6) \quad \langle \Delta, \text{grad } f(\mathbf{U}) \rangle_{\mathbb{R}} = \left( \frac{d}{dt} f(\text{Exp}_{\mathbf{U}}(t\Delta)) \right) \Big|_{t=0}.$$

We also need the following fact about the case of scale-invariant functions.

LEMMA 3.1. *Assume that  $f : \mathcal{U}_n \rightarrow \mathbb{R}$  satisfies the invariance property (2.11). Then for any  $\mathbf{U} \in \mathcal{U}_n$  and  $\mathbf{S}$  as in (2.11) it holds that*

$$\text{grad } f(\mathbf{U}\mathbf{S}) = \text{grad } f(\mathbf{U})\mathbf{S}.$$

*Proof.* By the chain rule, we have  $\nabla_{\mathbf{U}^*} f(\mathbf{U}\mathbf{S}) = (\nabla_{\mathbf{U}^*} f(\mathbf{U})) \mathbf{S}$ . Therefore,

$$\text{grad } f(\mathbf{U}\mathbf{S}) = \mathbf{U}\mathbf{S} \left( \mathbf{S}^H \mathbf{U}^H \nabla_{\mathbf{U}^*} f(\mathbf{U}) \mathbf{S} - (\nabla_{\mathbf{U}^*} f(\mathbf{U}) \mathbf{S})^H \mathbf{U} \mathbf{S} \right) = \mathbf{U} \Lambda(\mathbf{U}) \mathbf{S},$$

where the last equality follows from  $\mathbf{S}\mathbf{S}^H = \mathbf{I}$ . □

**3.3. Derivatives for elementary rotations.** This section contains general facts about derivatives of  $h_{(i,j),\mathbf{U}}$ . First, for  $i \neq j$  we introduce a useful projection operator  $\mathcal{P}_{i,j} : \mathbb{C}^{n \times n} \rightarrow \mathbb{C}^{2 \times 2}$  that extracts a submatrix of  $\mathbf{X} \in \mathbb{C}^{n \times n}$  as follows:

$$(3.7) \quad \mathcal{P}_{i,j}(\mathbf{X}) = \begin{bmatrix} X_{ii} & X_{ij} \\ X_{ji} & X_{jj} \end{bmatrix}.$$

Its adjoint operator is  $\mathcal{P}_{i,j}^T : \mathbb{C}^{2 \times 2} \rightarrow \mathbb{C}^{n \times n}$ , i.e.,

$$(3.8) \quad \mathcal{P}_{i,j}^T \left( \begin{bmatrix} a & c \\ b & d \end{bmatrix} \right) = \begin{matrix} & & & i & j & & \\ & & & \vdots & \vdots & & \\ & & & \mathbf{0} & \mathbf{0} & & \\ i & \left[ \begin{array}{cccc} \cdots & a & c & \mathbf{0} \\ \cdots & b & d & \mathbf{0} \end{array} \right] & & & & \\ j & & & & & & \\ & & & \mathbf{0} & \mathbf{0} & & \end{matrix}.$$

Note that for the Givens transformation in (2.8) we have

$$\mathcal{P}_{i,j}(\mathbf{G}^{(i,j,\Psi)}) = \Psi,$$

which makes it easy to express the Riemannian gradient of  $h_{(i,j),\mathbf{U}}$  through that of  $f$ .

LEMMA 3.2. *The Riemannian gradient of  $h_{(i,j),\mathbf{U}}$  defined in (2.9) at the identity matrix  $\mathbf{I}_2$  is a submatrix of the matrix  $\Lambda(\mathbf{U})$  defined in (3.3):*

$$(3.9) \quad \text{grad } h_{(i,j),\mathbf{U}}(\mathbf{I}_2) = \mathcal{P}_{i,j}(\Lambda(\mathbf{U})) = \begin{bmatrix} \Lambda_{ii}(\mathbf{U}) & \Lambda_{ij}(\mathbf{U}) \\ \Lambda_{ji}(\mathbf{U}) & \Lambda_{jj}(\mathbf{U}) \end{bmatrix}.$$

*Proof.* Denote  $h = h_{(i,j),\mathbf{U}}$  for simplicity. For any  $\Delta \in \mathbf{T}_{\mathbf{I}_2}\mathcal{U}_2$ , by (3.6)

$$\begin{aligned} \langle \Delta, \text{grad } h(\mathbf{I}_2) \rangle_{\mathfrak{R}} &= \left( \frac{d}{dt} h(\text{Exp}_{\mathbf{I}_2}(t\Delta)) \right) \Big|_{t=0} = \left( \frac{d}{dt} f(\mathbf{U} \mathbf{G}^{(i,j), \text{Exp}_{\mathbf{I}_2}(\Delta t)}) \right) \Big|_{t=0} \\ &= \left( \frac{d}{dt} f(\text{Exp}_{\mathbf{U}}(\mathbf{U} \mathcal{P}_{i,j}^{\top}(\Delta)t)) \right) \Big|_{t=0} = \langle \mathbf{U} \mathcal{P}_{i,j}^{\top}(\Delta), \text{grad } f(\mathbf{U}) \rangle_{\mathfrak{R}} = \langle \Delta, \mathcal{P}_{i,j}(\Lambda(\mathbf{U})) \rangle_{\mathfrak{R}}, \end{aligned}$$

which completes the proof.  $\square$

**3.4. Quotient manifold.** In order to handle scale invariance, it is often convenient to work on the quotient manifold. We define the action of  $\mathbb{T}^n$  on  $\mathcal{U}_n$  as

$$\mathbf{U} \cdot (z_1, \dots, z_n) = \mathbf{U} \begin{bmatrix} z_1 & & 0 \\ & \ddots & \\ 0 & & z_n \end{bmatrix}.$$

Since the action of  $\mathbb{T}^n$  on  $\mathcal{U}_n$  is free and proper, the quotient manifold  $\tilde{\mathcal{U}}_n = \mathcal{U}_n/\mathbb{T}^n$  is well-defined. In order to define the gradient and Hessians on  $\tilde{\mathcal{U}}_n$ , we use the standard splitting into vertical and horizontal space

$$\mathbf{T}_{\mathbf{U}}\mathcal{U}_n = \mathcal{V}_{\mathbf{U}}\mathcal{U}_n \oplus \mathcal{H}_{\mathbf{U}}\mathcal{U}_n,$$

where  $\mathcal{H}_{\mathbf{U}}\mathcal{U}_n$  contains the skew-symmetric matrices with zero diagonal:

$$\mathcal{H}_{\mathbf{U}}\mathcal{U}_n = \{ \mathbf{X} \in \mathbb{C}^{n \times n} : \mathbf{X} = \mathbf{U} \mathbf{Z}, \quad \mathbf{Z} + \mathbf{Z}^{\text{H}} = 0, \quad \text{diag}\{\mathbf{Z}\} = \mathbf{0} \}.$$

An element  $\tilde{\mathbf{U}} \in \tilde{\mathcal{U}}_n$  is then represented by  $\mathbf{U}$  and the tangent space  $\mathbf{T}_{\tilde{\mathbf{U}}}\tilde{\mathcal{U}}_n$  is identified with  $\mathcal{H}_{\mathbf{U}}\mathcal{U}_n$ , see [3, §3.5.8]. Moreover, the Riemannian metric on  $\tilde{\mathcal{U}}_n$  is defined as

$$\langle \tilde{\xi}, \tilde{\eta} \rangle_{\mathbf{T}_{\tilde{\mathbf{U}}}\tilde{\mathcal{U}}_n} = \langle \xi, \eta \rangle_{\mathbf{T}_{\mathbf{U}}\mathcal{U}_n},$$

because the inner product is invariant with respect to the choice of representative  $\mathbf{U}$ , see [3, Section 3.6.2]. This makes  $\tilde{\mathcal{U}}_n$  a Riemannian manifold; the natural projection  $\pi : \mathbf{U} \mapsto \tilde{\mathbf{U}}$  then becomes a Riemannian submersion.

Due to the invariance property (2.11), the function  $f$  is, in fact, defined on  $\tilde{\mathcal{U}}_n$  (we will denote the corresponding function by  $\tilde{f} : \tilde{\mathcal{U}}_n \rightarrow \mathbb{R}$ ).

*Remark 3.3.* As shown in [3, eqn. (3.39)], for any  $f$  satisfying the scale invariance property, we have (2.11),  $\text{grad } f(\mathbf{U}) \in \mathcal{H}_{\mathbf{U}}\mathcal{U}_n$  (which naturally represents the gradient of  $\tilde{f}$  in  $\tilde{\mathcal{U}}_n$ ). Therefore, in particular, the main diagonal of  $\Lambda(\mathbf{U})$  is zero.

*Remark 3.4.* Note that as in [41, Thm. A.15], for any  $\mathbf{Z} \in \mathcal{H}_{\mathbf{U}}\mathcal{U}_n$  the geodesic

$$(3.10) \quad \gamma(t) = \text{Exp}_{\mathbf{U}}(\mathbf{Z}t)$$

is horizontal (i.e. its derivative stays in the horizontal space  $\dot{\gamma}(t) \in \mathcal{H}_{\gamma(t)}\mathcal{U}_n$ ). Thus the exponential map in the quotient manifold  $\tilde{\mathcal{U}}_n$  is also defined by (3.10).

Finally, we make remarks about the two-dimensional manifold of  $2 \times 2$  rotations  $\tilde{\mathcal{U}}_2$ .

*Remark 3.5.* The matrices  $\Psi(c, s_1, s_2)$  defined in (2.10), in fact, parametrize  $\tilde{\mathcal{U}}_2$ .

*Remark 3.6.* Since all the elements on the diagonals are zero, the tangent space  $\mathbf{T}_{\tilde{\mathbf{U}}}\tilde{\mathcal{U}}_n$  to the  $n(n-1)$ -dimensional manifold  $\tilde{\mathcal{U}}_n$  can be decomposed as a direct sum of  $\frac{n(n-1)}{2}$  copies of  $\mathbf{T}_{\tilde{\mathbf{I}}_2}\tilde{\mathcal{U}}_2$  (spaces of  $2 \times 2$  skew-symmetric matrices with zero diagonal corresponding to different pairs  $(i, j)$ ); this can be also seen from Lemma 3.2.



**3.5. Riemannian Hessian and stationary points.** For a Riemannian manifold  $\mathcal{M}$  and a  $C^2$  function  $f : \mathcal{M} \rightarrow \mathbb{R}$ , the Riemannian Hessian at  $x \in \mathcal{M}$  is either defined as a linear map  $\mathbf{T}_x\mathcal{M} \rightarrow \mathbf{T}_x\mathcal{M}$  or as a bilinear form on  $\mathbf{T}_x\mathcal{M}$ ; the usual definition is based on the Riemannian connection [3, p.105].

For our purposes, for simplicity, we assume that the exponential map  $\text{Exp}_x : \mathbf{T}_x\mathcal{M} \rightarrow \mathcal{M}$  is given, and adopt the following definition based on [3, Proposition 5.5.4]. The Riemannian Hessian  $\text{Hess}_x f$  is the linear map  $\mathbf{T}_x\mathcal{M} \rightarrow \mathbf{T}_x\mathcal{M}$  defined by

$$\text{Hess}_x f = \text{Hess}_{\mathbf{0}_x} (f \circ \text{Exp}_x),$$

where  $\mathbf{0}_x$  is the origin in the tangent space, and  $\text{Hess}_{\mathbf{0}_x} g$  is the Euclidean Hessian of  $g : \mathbf{T}_x\mathcal{M} \rightarrow \mathbb{R}$ . Hence, similarly to (3.6), there is the following expression for the values of Riemannian Hessian as a quadratic form at  $\Delta \in \mathbf{T}_x\mathcal{M}$ :

$$(3.11) \quad \langle \text{Hess}_x f[\Delta], \Delta \rangle_{\mathfrak{R}} = \left( \frac{d^2}{dt^2} f(\text{Exp}_x(t\Delta)) \right) \Big|_{t=0}.$$

The Riemannian Hessian gives necessary and sufficient conditions of local extrema (see, for example, [46, Theorem 4.1]).

- If  $x$  is a local maximum of  $f$  on  $\mathcal{M}$ , then  $\text{Hess}_x f \preceq 0$  (negative semidefinite);
- If  $\text{grad } f(x) = 0$  and  $\text{Hess}_x f \prec_{\mathbf{T}_x\mathcal{M}} 0$  (i.e.,  $\text{Hess}_x f \preceq 0$  and  $\text{rank}\{\text{Hess}_x f\} = \dim(\mathcal{M})$ ), then  $f$  has a strict local maximum at  $x$ .

Finally, we distinguish stationary points with nonsingular Riemannian Hessian.

**DEFINITION 3.7.** *A stationary point ( $x \in \mathcal{M}$ ,  $\text{grad } f(x) = 0$ ) is called non-degenerate if  $\text{Hess}_x f$  is nonsingular on  $\mathbf{T}_x\mathcal{M}$ .*

In our case, a stationary point is never non-degenerate, as shown below.

**LEMMA 3.8.** *Assume that  $f : \mathcal{U}_n \rightarrow \mathbb{R}$  satisfies the invariance property (2.11). Let  $U$  be a stationary point ( $\text{grad } f(U) = 0$ ) and*

$$(3.12) \quad \mathbf{Z}_k = [\mathbf{0} \ \cdots \ \mathbf{0} \ iu_k \ \mathbf{0} \ \cdots \ \mathbf{0}] = U\Omega_k \in \mathbf{T}_U\mathcal{U}_n,$$

where  $\Omega_k = ie_k e_k^\top$  (where  $e_k$  is the  $k$ -th unit vector). Then  $\text{Hess}_U f[\mathbf{Z}_k] = \mathbf{0}$  (i.e., all  $\mathbf{Z}_k$  are in the kernel of  $\text{Hess}_U f$ ). In particular,  $\text{rank}\{\text{Hess}_U f\} \leq n(n-1)$ .

*Proof.* Let  $\gamma : \mathbb{R} \rightarrow \mathcal{U}_n$  be a curve defined as  $\gamma(t) = \text{Exp}_U(t\mathbf{Z}_k)$ , with  $\gamma(0) = U$ ,  $\gamma'(0) = \mathbf{Z}_k$ . By [3, Def. 5.5.1], [3, (5.15)] and Lemma 3.1, we obtain

$$\begin{aligned} \text{Hess}_U f[\mathbf{Z}_k] &= \mathbf{P}_U \left( \frac{d}{dt} \text{grad } f(\gamma(t)) \Big|_{t=0} \right) = \mathbf{P}_U \left( \frac{d}{dt} \text{grad } f(U) \exp(t\Omega_k) \Big|_{t=0} \right) \\ &= \mathbf{P}_U (\text{grad } f(U)\Omega_k) = \frac{U}{2} (\Lambda(U)\Omega_k - \Omega_k\Lambda(U)). \end{aligned}$$

Note that  $U$  is a stationary point. Then  $\Lambda(U) = 0$ , and thus  $\text{Hess}_U f[\mathbf{Z}_k] = \mathbf{0}$ . Since  $\{\mathbf{Z}_k\}_{k=1}^n$  are linearly independent,  $\text{rank}\{\text{Hess}_U f\} \leq n(n-1)$ .  $\square$

#### 4. Finding Jacobi rotations and derivatives for complex forms.

**4.1. On correctness of Jacobi-G.** The following fact follows from Lemma 3.2.

**COROLLARY 4.1.** *Let  $f$  and  $h_{(i,j),U}$  be as in Lemma 3.2. Then*

$$\max_{1 \leq i < j \leq n} \|\text{grad } h_{(i,j),U}(\mathbf{I}_2)\| \geq \frac{\sqrt{2}}{n} \|\text{grad } f(U)\|.$$

*Proof.* By (3.2) and Lemma 3.2, we see that

$$\|\text{grad } f(\mathbf{U})\|^2 = \|\mathbf{\Lambda}(\mathbf{U})\|^2 = \sum_{i,j=1}^{n,n} |\Lambda(\mathbf{U})_{i,j}|^2 \leq \frac{n^2}{2} \max_{1 \leq i < j \leq n} \|\text{grad } h_{(i,j),\mathbf{U}}(\mathbf{I}_2)\|^2. \quad \square$$

*Remark 4.2.* Corollary 4.1 implies that for any differentiable  $f$  it is always possible to find  $(i_k, j_k)$  satisfying the inequality (2.13), provided  $\delta \leq \sqrt{2}/n$ .

In fact, it gives an explicit way to find such a pair, as shown by the following remark.

*Remark 4.3.* From Lemma 3.2 and Remark 3.3, the condition (2.13) becomes

$$(4.1) \quad \sqrt{2}|\Lambda_{i_k, j_k}| \geq \delta \|\mathbf{\Lambda}\|,$$

where  $\Lambda = \Lambda(\mathbf{U}_{k-1}) = \mathbf{U}_{k-1}^H \text{grad } f(\mathbf{U}_{k-1})$  is as in (3.3). Thus the pair can be selected by looking at the elements of  $\mathbf{\Lambda}$ , for example, according to one of the strategies: (a) choose the maximal modulus element of  $\mathbf{\Lambda}$ ; or (b) choose the first pair (e.g., in cyclic order) that satisfies (4.1); if  $\delta$  is small, then (4.1) is most of the time satisfied.

**4.2. Elementary rotations.** First of all, our cost functions that satisfy the invariance property (2.11); hence the restriction (2.9) is also scale-invariant

$$(4.2) \quad h_{(i,j),\mathbf{U}}(\mathbf{\Psi}) = h_{(i,j),\mathbf{U}}(\mathbf{\Psi} \begin{bmatrix} z_1 & 0 \\ 0 & z_2 \end{bmatrix}), \quad \text{for all } z_1, z_2 \in \mathbb{T}.$$

Hence, we can restrict to matrices  $\mathbf{\Psi} = \mathbf{\Psi}(c, s_1, s_2)$  defined in (2.10), and maximize

$$h_{(i,j),\mathbf{U}}(c, s_1, s_2) \stackrel{\text{def}}{=} h_{(i,j),\mathbf{U}}(\mathbf{\Psi}(c, s_1, s_2)) = h_{(i,j),\mathbf{U}}\left(\begin{bmatrix} c & -(s_1 + is_2) \\ s_1 - is_2 & c \end{bmatrix}\right);$$

next, we show how to maximize  $h_{(i,j),\mathbf{U}}(c, s_1, s_2)$  for cost functions (2.5) and (2.7).

**PROPOSITION 4.4.** *For a cost function  $f$  of the form (2.5) and (2.7) with  $d \leq 3$ , its restriction for any pair  $(i, j)$  and  $\mathbf{U} \in \mathcal{U}_n$  can be expressed as a quadratic form*

$$(4.3) \quad h_{(i,j),\mathbf{U}}(c, s_1, s_2) = \mathbf{r}^T \mathbf{\Gamma}^{(i,j),\mathbf{U}} \mathbf{r} + C, \quad \text{where}$$

$$(4.4) \quad \mathbf{r} = \mathbf{r}(c, s_1, s_2) \stackrel{\text{def}}{=} \begin{bmatrix} 2c^2 - 1 & -2cs_1 & -2cs_2 \end{bmatrix}^T = \begin{bmatrix} \cos 2\theta & -\sin 2\theta \cos \phi & -\sin 2\theta \sin \phi \end{bmatrix}^T,$$

$\mathbf{\Gamma}^{(i,j),\mathbf{U}} \in \mathbb{R}^{3 \times 3}$  is a symmetric matrix and  $C$  is a constant, whose entries depend polynomially on the real and imaginary parts of  $\mathbf{U}$  and of tensors  $\mathbf{B}$  or  $\mathbf{A}^{(\ell)}$ .

In fact, Proposition 4.4 was already known for special cases of problems (2.1)–(2.3) (see [18, Ch. 5] for an overview); in its general form, Proposition 4.4 is a special case of a general result (Theorem 4.16 in subsection 4.6) that establishes the form of  $h_{(i,j),\mathbf{U}}$  for any order  $d$ .

To illustrate the idea, we give an example for joint matrix diagonalization.

*Example 4.5.* For the function (2.1), denote  $\mathbf{W}^{(\ell)} = \mathbf{U}^H \mathbf{A}^{(\ell)} \mathbf{U}$ , so that  $f(\mathbf{U}) = \sum_{\ell=1}^L \|\text{diag}\{\mathbf{W}^{(\ell)}\}\|^2$ . Then it is known [15] that<sup>9</sup>

$$\mathbf{\Gamma}^{(i,j),\mathbf{U}} = \frac{1}{2} \sum_{\ell=1}^L \left( |W_{jj}^{(\ell)} + W_{ii}^{(\ell)}|^2 \mathbf{I}_3 + \Re \left( \mathbf{z}(\mathbf{W}^{(\ell)}) \mathbf{z}^H(\mathbf{W}^{(\ell)}) \right) \right), \quad \text{where}$$

$$\mathbf{z}(\mathbf{W}) \stackrel{\text{def}}{=} \begin{bmatrix} W_{jj} - W_{ii} & W_{ij} + W_{ji} & -i(W_{ij} - W_{ji}) \end{bmatrix}^T.$$

<sup>9</sup>Note that these expressions can be simplified for Hermitian matrices  $\mathbf{W}^{(\ell)}$ , because in this case  $W_{ij} + W_{ji} = 2W_{ij}^{\Re}$  and  $-i(W_{ij} - W_{ji}) = 2W_{ij}^{\Im}$ .

Similar expressions exist for the cost functions (2.2) (see [19, (9.29)] and [18, Section 5.3.2]) and (2.3) (see [16]), but we omit them due to space limitations, and also because Proposition 4.4 supersedes all these results.

*Remark 4.6.* By Proposition 4.4, the maximization of  $h_{(i,j),\mathbf{U}}(c, s_1, s_2)$  is equivalent to maximization of  $\mathbf{r}^\top \mathbf{\Gamma}^{(i,j),\mathbf{U}} \mathbf{r}$  subject to  $\|\mathbf{r}\| = 1$ . Thus a maximizer of  $h_{(i,j),\mathbf{U}}(c, s_1, s_2)$  can be obtained from an eigenvector of  $\mathbf{\Gamma}^{(i,j),\mathbf{U}}$ , which we summarize in Algorithm 4.1. Note that we can choose the maximizer such that  $c \geq \frac{\sqrt{2}}{2}$ .

---

**Algorithm 4.1** Finding Jacobi rotations

---

**Input:** Point  $\mathbf{U}$ , pair  $(i, j)$ .

**Output:** A maximizer  $(c, s_1, s_2)$  of  $h_{(i,j),\mathbf{U}}(c, s_1, s_2)$ .

- Build  $\mathbf{\Gamma} = \mathbf{\Gamma}^{(i,j),\mathbf{U}}$  according to Proposition 4.4.
  - Find a leading eigenvector  $\mathbf{w}$  corresponding to the maximal eigenvalue of  $\mathbf{\Gamma}$  (with normalization  $\|\mathbf{w}\| = 1, w_1 \geq 0$ ).
  - Choose  $\theta \in [0, \frac{\pi}{4}]$  and  $s_1, s_2$  by setting  $\theta = \frac{\arccos(w_1)}{2} \in [0, \frac{\pi}{4}]$ ,  $c = \cos(\theta) = \sqrt{\frac{w_1+1}{2}} \geq \frac{\sqrt{2}}{2}$ ,  $s_1 = -\frac{w_2}{2c}$ ,  $s_2 = -\frac{w_3}{2c}$ .
- 

**4.3. Riemannian derivatives for the cost functions.** In this subsection, we link the Riemannian derivatives of  $h_{(i,j),\mathbf{U}}$  with the entries of the matrix  $\mathbf{\Gamma}^{(i,j),\mathbf{U}}$ .

LEMMA 4.7. *Let  $h_{(i,j),\mathbf{U}}$  satisfy (4.2) and be expressed as in (4.3). Then*

$$\text{grad } h_{(i,j),\mathbf{U}}(\mathbf{I}_2) = 2 \begin{bmatrix} 0 & \mathbf{\Gamma}_{12}^{(i,j),\mathbf{U}} + i\mathbf{\Gamma}_{13}^{(i,j),\mathbf{U}} \\ -\mathbf{\Gamma}_{12}^{(i,j),\mathbf{U}} + i\mathbf{\Gamma}_{13}^{(i,j),\mathbf{U}} & 0 \end{bmatrix},$$

hence, in particular,  $\Lambda_{ij}(\mathbf{U}) = 2(\mathbf{\Gamma}_{12}^{(i,j),\mathbf{U}} + i\mathbf{\Gamma}_{13}^{(i,j),\mathbf{U}})$  by Lemma 3.2.

*Proof.* Denote  $h = h_{(i,j),\mathbf{U}}$  and  $\mathbf{\Gamma} = \mathbf{\Gamma}^{(i,j),\mathbf{U}}$ . By (3.4) and Remark 3.3, we see that  $\text{grad } h(\mathbf{I}_2)$  is skew-Hermitian, and is decomposable as  $\text{grad } h(\mathbf{I}_2) = 2\omega_1 \mathbf{\Delta}_1 + 2\omega_2 \mathbf{\Delta}_2$ ,

$$(4.5) \quad \text{where } \mathbf{\Delta}_1 = \begin{bmatrix} 0 & -\frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix}, \quad \mathbf{\Delta}_2 = \begin{bmatrix} 0 & -\frac{i}{2} \\ -\frac{i}{2} & 0 \end{bmatrix}.$$

Note that  $\{\mathbf{\Delta}_1, \mathbf{\Delta}_2\}$  is an orthogonal basis of  $\mathbf{T}_{\mathbf{I}_2} \tilde{\mathcal{U}}_2$ . Since  $\|\mathbf{\Delta}_1\|^2 = \|\mathbf{\Delta}_2\|^2 = 1/2$

$$\omega_k = \langle \mathbf{\Delta}_k, \text{grad } h(\mathbf{I}_2) \rangle_{\mathfrak{R}} = \left( \frac{d}{dt} h(e^{t\mathbf{\Delta}_k}) \right) \Big|_{t=0}$$

for  $k = 1, 2$ . On the other hand, we have

$$\begin{aligned} h(e^{t\mathbf{\Delta}_1}) &= h \left( \begin{bmatrix} \cos \frac{t}{2} & -\sin \frac{t}{2} \\ \sin \frac{t}{2} & \cos \frac{t}{2} \end{bmatrix} \right) = h \left( \cos \frac{t}{2}, \sin \frac{t}{2}, 0 \right) = \bar{h}(\cos t, -\sin t, 0), \\ h(e^{t\mathbf{\Delta}_2}) &= h \left( \begin{bmatrix} \cos \frac{t}{2} & -i \sin \frac{t}{2} \\ -i \sin \frac{t}{2} & \cos \frac{t}{2} \end{bmatrix} \right) = h \left( \cos \frac{t}{2}, 0, \sin \frac{t}{2} \right) = \bar{h}(\cos t, 0, -\sin t), \end{aligned}$$

where  $\bar{h}(\mathbf{r}) = \mathbf{r}^\top \mathbf{\Gamma} \mathbf{r}$ . Since  $\nabla \bar{h}(\mathbf{r}) = 2\mathbf{\Gamma} \mathbf{r}$ , we have

$$\omega_1 = -\frac{\partial \bar{h}}{\partial r_2}(1, 0, 0) = -2\Gamma_{21}, \quad \omega_2 = -\frac{\partial \bar{h}}{\partial r_3}(1, 0, 0) = -2\Gamma_{31},$$

which completes the proof. □

LEMMA 4.8. For  $h_{(i,j),\mathbf{U}}$  as in Lemma 4.7, and the basis of  $\mathbf{T}_{\mathbf{I}_2}\tilde{\mathcal{U}}_2$  as in (4.5), the Riemannian Hessian of  $\tilde{h}$  ( $h_{(i,j),\mathbf{U}}$  on  $\tilde{\mathcal{U}}_2$ ) is

$$(4.6) \quad \text{Hess}_{\mathbf{I}_2}\tilde{h} = \mathfrak{D}_{\mathbf{U}}^{(i,j)} \stackrel{\text{def}}{=} 2 \left( \begin{bmatrix} \Gamma_{2,2}^{(i,j),\mathbf{U}} & \Gamma_{2,3}^{(i,j),\mathbf{U}} \\ \Gamma_{3,2}^{(i,j),\mathbf{U}} & \Gamma_{3,3}^{(i,j),\mathbf{U}} \end{bmatrix} - \Gamma_{1,1}^{(i,j),\mathbf{U}} \mathbf{I}_2 \right).$$

*Proof.* We denote  $h = h_{(i,j),\mathbf{U}}$  and  $\Gamma = \Gamma^{(i,j),\mathbf{U}}$  for simplicity, and take

$$(4.7) \quad \Omega = \alpha_1 \Delta_1 + \alpha_2 \Delta_2, \quad \text{where } \alpha_1, \alpha_2 \in \mathbb{R}, \alpha_1^2 + \alpha_2^2 = 1,$$

and  $\{\Delta_1, \Delta_2\}$  are as in (4.5). Then  $h(e^{t\Omega})$  and its derivative can be expressed as

$$(4.8) \quad h(e^{t\Omega}) = h \left( \begin{bmatrix} \cos \frac{t}{2} & -(\alpha_1 + i\alpha_2) \sin \frac{t}{2} \\ (\alpha_1 - i\alpha_2) \sin \frac{t}{2} & \cos \frac{t}{2} \end{bmatrix} \right) = \bar{h}(\cos t, -\alpha_1 \sin t, -\alpha_2 \sin t),$$

$$\frac{d}{dt} h(e^{t\Omega}) = -2 \begin{bmatrix} \sin t & \alpha_1 \cos t & \alpha_2 \cos t \end{bmatrix} \Gamma \begin{bmatrix} \cos t & -\alpha_1 \sin t & -\alpha_2 \sin t \end{bmatrix}^\top,$$

and thus by [3, (5.32)]

$$(4.9) \quad \langle \Omega, \text{Hess}_{\mathbf{I}_2} h[\Omega] \rangle_{\mathbb{R}} = \left( \frac{d^2}{dt^2} h(e^{t\Omega}) \right) \Big|_{t=0} = \begin{bmatrix} \alpha_1 & \alpha_2 \end{bmatrix} \mathfrak{D}_{\mathbf{U}}^{(i,j)} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix},$$

Finally, note that the geodesic is horizontal (its derivative stays in the horizontal space), hence (4.9) is valid for the Hessian of  $\tilde{h}$ .  $\square$

COROLLARY 4.9. If  $\mathbf{I}_2$  is a local maximizer of  $h_{(i,j),\mathbf{U}}$ , then  $\mathfrak{D}_{\mathbf{U}}^{(i,j)} \preceq 0$ .

Remark 4.10. Denote  $\Gamma = \Gamma^{(i,j),\mathbf{U}}$ . Then  $\mathfrak{D}_{\mathbf{U}}^{(i,j)}$  is negative definite if and only if

$$\Gamma_{11} > \lambda_{\max} \left( \begin{bmatrix} \Gamma_{22} & \Gamma_{23} \\ \Gamma_{23} & \Gamma_{33} \end{bmatrix} \right).$$

If, in addition,  $\text{grad } h_{(i,j),\mathbf{U}}(\mathbf{I}_2) = 0$ , this is equivalent to saying that  $\lambda_1(\Gamma) > \lambda_2(\Gamma)$  (i.e., the first two eigenvalues are separated) and  $\Gamma_{11} = \lambda_1(\Gamma)$ .

#### 4.4. Complex conjugate forms and equivalence of the cost functions.

For  $\mathcal{A} \in \mathbb{C}^{n \times \dots \times n}$  of order  $d$  and an integer  $t$ ,  $0 \leq t \leq d$ , we define the corresponding homogeneous conjugate form [33] (a generalization of a homogeneous polynomial) as

$$(4.10) \quad g_{\mathcal{A},t}(\mathbf{u}) = \mathcal{A} \bullet_1 \mathbf{u}^* \cdots \bullet_t \mathbf{u}^* \bullet_{t+1} \mathbf{u} \cdots \bullet_d \mathbf{u},$$

i.e., the tensor contracted  $t$  times with  $\mathbf{u}^*$  and the remaining  $d-t$  times with  $\mathbf{u}$ . Then it is easy to see that the cost functions (2.5) and (2.7) can be written as<sup>10</sup>

$$(4.11) \quad f(\mathbf{U}) = \sum_{k=1}^n \gamma(\mathbf{u}_k),$$

where  $\gamma(\mathbf{u})$  is one of the following options depending on the cost function:

$$(4.12) \quad \gamma(\mathbf{u}) = \sum_{\ell=1}^L \alpha_\ell |g_{\mathcal{A}^{(\ell)},t_\ell}(\mathbf{u})|^2, \quad \text{or}$$

$$(4.13) \quad \gamma(\mathbf{u}) = g_{\mathcal{B},d}(\mathbf{u}), \quad \text{where } \mathcal{B} \text{ is Hermitian in the sense of (2.6).}$$

Note that we call forms of type (4.13) *Hermitian forms*. The equivalence of (2.5) and (2.7) is established by the following result.

<sup>10</sup>similarly to contrast functions [16, 17]

LEMMA 4.11. *When restricted to norm-one vectors  $\mathbf{u}$ ,  $\|\mathbf{u}\| = 1$ , a function  $\gamma(\mathbf{u})$  is a Hermitian form (4.13) of order  $2d$  if and only if it can be represented as (4.12) for tensors  $\mathcal{A}^{(1)}, \dots, \mathcal{A}^{(L)}$  of orders  $d_1, \dots, d_L \leq d$ .*

Lemma 4.11 is a rather straightforward generalization of the results of [33, Proposition 3.5]; still, we provide a proof in Appendix A for completeness, and also because our notation is slightly different from that of [33].

We conclude this subsection by showing how to find Wirtinger derivatives for forms (4.10).

LEMMA 4.12. *For a form  $g(\mathbf{u}) = g_{\mathcal{A},t}(\mathbf{u})$  defined in (4.10), it holds that*

$$\begin{aligned}\nabla_{\mathbf{u}^*} g(\mathbf{u}) &= \sum_{k=1}^t \mathcal{A} \bullet_1 \mathbf{u}^* \cdots \cancel{\bullet_k \mathbf{u}^*} \cdots \bullet_t \mathbf{u}^* \bullet_{t+1} \mathbf{u} \cdots \cdots \bullet_d \mathbf{u}, \\ \nabla_{\mathbf{u}} g(\mathbf{u}) &= \sum_{k=1}^{d-t} \mathcal{A} \bullet_1 \mathbf{u}^* \cdots \cdots \bullet_t \mathbf{u}^* \bullet_{t+1} \mathbf{u} \cdots \cancel{\bullet_{t+k} \mathbf{u}} \cdots \bullet_d \mathbf{u}, \\ \nabla_{\mathbf{u}^*} |g(\mathbf{u})|^2 &= (g(\mathbf{u}))^* \nabla_{\mathbf{u}^*} g(\mathbf{u}) + (g(\mathbf{u})) (\nabla_{\mathbf{u}} g(\mathbf{u}))^*.\end{aligned}$$

*Proof.* The first two equations follow by the rule of product differentiation and the following identities [30, Table IV]

$$\nabla_{\mathbf{u}} (\mathbf{u}^H \mathbf{a})(\mathbf{u}) = \nabla_{\mathbf{u}^*} (\mathbf{u}^T \mathbf{a})(\mathbf{u}) = \mathbf{0}, \quad \nabla_{\mathbf{u}^*} (\mathbf{u}^H \mathbf{a})(\mathbf{u}) = \nabla_{\mathbf{u}} (\mathbf{u}^T \mathbf{a})(\mathbf{u}) = \mathbf{a}.$$

The last equation follows<sup>11</sup> from the rule of differentiation of composition [30, Theorem 1], and the fact that  $d|z|^2 = z^* dz + z dz^*$ .  $\square$

**4.5. Riemannian gradients for cost functions of interest.** Before computing derivatives for (2.5) and (2.7), we make a remark about symmetries in these functions.

*Remark 4.13.* For any form  $g_{\mathcal{A},t}(\mathbf{u})$  (4.10) we can assume without loss of generality that the tensor  $\mathcal{A}$  is *t-semi-symmetric*, i.e., satisfies the following symmetries:

$$\mathcal{A}_{i_1 \dots i_t i_{t+1} \dots i_d} = \mathcal{A}_{\pi_1(i_1 \dots i_t) \pi_2(i_{t+1} \dots i_d)}$$

for any index  $(i_1 \dots i_t i_{t+1} \dots i_d)$  and any pair of permutations  $\pi_1(\cdot)$  and  $\pi_2(\cdot)$  of indices corresponding to the same group of contractions in (4.10). For example,

- for any  $\mathcal{A} \in \mathbb{C}^{n \times n \times n}$  we can define  $\mathcal{T}_{ijk} = \frac{\mathcal{A}_{ijk} + \mathcal{A}_{ikj}}{2}$  such that

$$\mathcal{A} \bullet_1 \mathbf{u}^* \bullet_2 \mathbf{u} \bullet_3 \mathbf{u} = \mathcal{T} \bullet_1 \mathbf{u}^* \bullet_2 \mathbf{u} \bullet_3 \mathbf{u};$$

- similarly, for any  $\mathcal{B} \in \mathbb{C}^{n \times n \times n \times n}$  and  $\mathcal{S}_{ijkl} = \frac{\mathcal{B}_{ijkl} + \mathcal{B}_{ijlk} + \mathcal{B}_{jikl} + \mathcal{B}_{jilk}}{4}$  we have

$$\mathcal{B} \bullet_1 \mathbf{u}^* \bullet_2 \mathbf{u}^* \bullet_3 \mathbf{u} \bullet_4 \mathbf{u} = \mathcal{S} \bullet_1 \mathbf{u}^* \bullet_2 \mathbf{u}^* \bullet_3 \mathbf{u} \bullet_4 \mathbf{u}.$$

Thus the tensors can be assumed to be semi-symmetric in (2.5) and (2.7).

Next, we are going to find Riemannian gradients for cost functions (2.5) and (2.7). Since the cost functions can be written as (4.11), we have

$$(4.14) \quad \nabla_{\mathbf{U}^*} f(\mathbf{U}) = [\nabla_{\mathbf{u}^*} \gamma(\mathbf{u}_1) \quad \cdots \quad \nabla_{\mathbf{u}^*} \gamma(\mathbf{u}_n)],$$

hence Lemma 4.12 can be used to prove the following result.

<sup>11</sup>An alternative proof can be derived by using the representation of  $|g(\mathbf{u})|^2$  as a Hermitian form, contained in the proof of Lemma 4.12.

PROPOSITION 4.14. (i) Let  $\mathcal{B}$  be a Hermitian  $d$ -semi-symmetric (as in Remark 4.13) tensor. Then for the cost function (2.7),

$$(4.15) \quad \left( \mathbf{U}^H \nabla_{\mathbf{U}^*} f(\mathbf{U}) \right)_{ij} = d\mathcal{V}_{ij\dots j}, \quad \Lambda(\mathbf{U})_{ij} = d(\mathcal{V}_{ij\dots j} - \mathcal{V}_{i\dots ij}),$$

$$(4.16) \quad \text{where } \mathcal{V} = \mathcal{B} \bullet_1 \mathbf{U}^H \dots \bullet_d \mathbf{U}^H \bullet_{d+1} \mathbf{U}^T \dots \bullet_{2d} \mathbf{U}^T$$

is the rotated Hermitian tensor.

(ii) Let  $\mathcal{A}$  be a  $t$ -semi-symmetric tensor and  $\gamma(\mathbf{u}) = |g_{\mathcal{A},t}(\mathbf{u})|^2$ . For  $f$  defined as (4.11) the gradients can be expressed as

$$(4.17) \quad \left( \mathbf{U}^H \nabla_{\mathbf{U}^*} f(\mathbf{U}) \right)_{ij} = t\mathcal{W}_{j\dots j}^* \mathcal{W}_{ij\dots j} + (d-t)\mathcal{W}_{j\dots j} \mathcal{W}_{j\dots ji}^*;$$

$$\Lambda_{ij}(\mathbf{U}) = t(\mathcal{W}_{j\dots j}^* \mathcal{W}_{ij\dots j} - \mathcal{W}_{i\dots i} \mathcal{W}_{ji\dots i}^*) + (d-t)(\mathcal{W}_{j\dots j} \mathcal{W}_{j\dots ji}^* - \mathcal{W}_{i\dots i}^* \mathcal{W}_{i\dots ij}),$$

where  $\mathcal{W} = \mathcal{A} \bullet_1 \mathbf{U}^H \dots \bullet_t \mathbf{U}^H \bullet_{t+1} \mathbf{U}^T \dots \bullet_d \mathbf{U}^T$  is the rotated tensor.

*Proof.* (i) By (4.14) and Lemma 4.12, we get

$$\begin{aligned} \left( \mathbf{U}^H \nabla_{\mathbf{U}^*} f(\mathbf{U}) \right)_{ij} &= \mathbf{u}_i^H \nabla_{\mathbf{u}^*} \gamma(\mathbf{u}_j) \\ &= \sum_{k=1}^d \mathcal{B} \bullet_1 \mathbf{u}_j^* \dots \bullet_k \mathbf{u}_j^* \dots \bullet_d \mathbf{u}_j^* \bullet_{d+1} \mathbf{u}_j \dots \bullet_{2d} \mathbf{u}_j \bullet_k \mathbf{u}_i^* = d\mathcal{V}_{ij\dots j}, \end{aligned}$$

where the last equality is due to symmetries. The form of  $\Lambda$  follows from (3.3).

(ii) The proof is similar<sup>12</sup> to (i), and follows from Lemma 4.12 and the equalities

$$g(\mathbf{u}_j) = g_{\mathcal{A},t}(\mathbf{u}_j) = \mathcal{W}_{j\dots j}, \quad \mathbf{u}_i^H \nabla_{\mathbf{u}^*} g(\mathbf{u}_j) = \mathcal{W}_{ij\dots j}, \quad \mathbf{u}_i^H (\nabla_{\mathbf{u}} g(\mathbf{u}))^* = \mathcal{W}_{ji\dots i}^* \quad \square$$

Remark 4.15. Part ii of Proposition 4.14 also allows us to find the Riemannian gradient for all functions of the form (2.5), by summing individual gradients for each  $\mathcal{A}^{(\ell)}$ . For example, the Riemannian gradient of the cost function (2.1) simplifies to

$$(4.18) \quad \Lambda_{ij}(\mathbf{U}) = \sum_{\ell=1}^L (W_{jj}^{(\ell)} - W_{ii}^{(\ell)})^* W_{ij}^{(\ell)} + (W_{jj}^{(\ell)} - W_{ii}^{(\ell)})(W_{ji}^{(\ell)})^*,$$

where  $W^{(\ell)}$  is as in Example 4.5. Note that (4.18) agrees with Lemma 4.7.

**4.6. Elementary update for Hermitian forms.** In this subsection, for simplicity, we only consider Hermitian tensors (2.6) of order  $2d$  which we assume to be  $d$ -semi-symmetric; we also take  $\mathcal{V}$  as in (4.16). Then  $h_{(i,j),\mathbf{U}}(\Psi)$  has the form

$$h_{(i,j),\mathbf{U}}(\Psi) = \text{tr}\{\mathcal{V} \bullet_1 \mathbf{G}^H \dots \bullet_d \mathbf{G}^H \bullet_{d+1} \mathbf{G}^T \dots \bullet_{2d} \mathbf{G}^T\}$$

where  $\mathbf{G} = \mathbf{G}^{(i,j,k,\Psi)}$  is the Givens transformation. Note that the Givens transformations change only elements of  $\mathcal{V}$  with at least one of indices equal to  $i$  or  $j$ , hence

$$(4.19) \quad h_{(i,j),\mathbf{U}}(\Psi) = \underbrace{\sum_{k \neq i,j} \mathcal{V}_{k\dots k}}_{\text{constant}} + \text{tr}\{\mathcal{T} \bullet_1 \Psi^H \dots \bullet_d \Psi^H \bullet_{d+1} \Psi^T \dots \bullet_{2d} \Psi^T\},$$

where  $\mathcal{T} = \mathcal{V}_{(i,j),\dots,(i,j)}$  is the  $2 \times \dots \times 2$  subtensor of  $\mathcal{V}$  corresponding to indices  $i, j$ . Then the following result characterizes the elementary rotations.

<sup>12</sup>The proof can be also directly obtained from (A.2) and (i); the tensor needs to be semi-symmetrized before applying (i), hence the second term appears in (4.15) compared with (4.17).

**THEOREM 4.16.** *Let  $\mathcal{T}$  be a Hermitian  $2d$ -order  $d$ -semi-symmetric  $2 \times \dots \times 2$  tensor. Then there exists a  $3 \times \dots \times 3$  real symmetric tensor  $\mathcal{F}$  of order  $2m$  for  $m = \lfloor \frac{d}{2} \rfloor$  such that*

$$\tilde{h}(c, s_1, s_2) \stackrel{\text{def}}{=} \text{tr}\{\mathcal{T} \bullet_1 \Psi^H \dots \bullet_d \Psi^H \bullet_{d+1} \Psi^T \dots \bullet_{2d} \Psi^T\} = \mathcal{F} \bullet_1 \mathbf{r} \dots \bullet_{2m} \mathbf{r},$$

where  $\Psi = \Psi(c, s_1, s_2)$  and  $\mathbf{r} = \mathbf{r}(c, s_1, s_2)$  are as in (2.10) and (4.4).

The proof of [Theorem 4.16](#) is contained in [Appendix A](#).

*Remark 4.17.* [Theorem 4.16](#) implies that:

- $m = 1$  for  $d \leq 3$ , i.e.,  $\mathcal{F}$  is a symmetric  $3 \times 3$  matrix (called  $\Gamma$  in [Proposition 4.4](#)). Thus, [Theorem 4.16](#) provides a proof for [Proposition 4.4](#).
- $m = 2$  for  $d = 4$ , in particular, the elementary update for the 4-th order complex tensor diagonalization requires maximizing a 4-th order ternary form (which was established in [\[19\]](#) for this particular case).
- For  $d > 3$  (unlike  $d \leq 3$ ), the update cannot be computed in a closed form.

*Remark 4.18.* The proof of [Theorem 4.16](#) gives a systematic way to find the coefficients of  $\mathcal{F}$  for any instance of (2.5) or (2.7), and thus generalizes existing expressions derived for special cases (see [\[18, Ch. 5\]](#)).

## 5. Weak convergence results.

### 5.1. Global rates of convergence of descent algorithms on manifolds.

We first recall a simplified version of result presented in [\[9, Thm. 2.5\]](#) on convergence of ascent algorithms (originally proposed in [\[9\]](#) for retraction-based algorithms).

**LEMMA 5.1** ([\[9, Theorem 2.5\]](#)). *Let  $f : \mathcal{M} \rightarrow \mathbb{R}$  be bounded from above by  $f^*$ . Suppose that, for a sequence<sup>13</sup> of  $x_k$ , there exists  $c > 0$  such that*

$$(5.1) \quad f(x_{k+1}) - f(x_k) \geq c \|\text{grad } f(x_k)\|^2.$$

Then

- (i)  $\|\text{grad } f(x_k)\| \rightarrow 0$  as  $k \rightarrow \infty$ ;
- (ii) We can find an  $x_k$  with  $\|\text{grad } f(x_k)\| \leq \varepsilon$  and  $f(x_k) \geq f(x_0)$  in at most

$$K_\varepsilon = \left\lceil \frac{f^* - f(x_0)}{c} \frac{1}{\varepsilon^2} \right\rceil$$

iterations; i.e., there exists  $k \leq K_\varepsilon$  such that  $\|\text{grad } f(x_k)\| < \varepsilon$ .

*Proof.* (i) We use the classic telescopic sums argument to obtain

$$f^* - f(x_0) \geq f(x_K) - f(x_0) = \sum_{k=0}^{K-1} (f(x_{k+1}) - f(x_k)) \geq c \sum_{k=0}^{K-1} \|\text{grad } f(x_k)\|^2.$$

Then we have that  $\sum_{k=0}^{\infty} \|\text{grad } f(x_k)\|^2$  is convergent, thus  $\|\text{grad } f(x_k)\| \rightarrow 0$ .

- (ii) Assume that  $\|\text{grad } f(x_k)\| > \varepsilon$  for all  $K - 1$  iterations. Then, in a similar way,

$$f^* - f(x_0) \geq cK \min_{0 \leq k \leq K-1} \|\text{grad } f(x_k)\|^2 > cK\varepsilon^2,$$

which can only hold if  $K < K_\varepsilon$ . □

<sup>13</sup>Note that in the original formulation of [\[9, Theorem 2.5\]](#)  $x_k$  were chosen as retractions of some vectors in  $\mathbf{T}_{x_{k-1}}$ . However, it is easy to see that this condition is not needed in the proof.

For checking the ascent condition (5.1), we recall a lemma on retractions.

DEFINITION 5.2. ([3, Definition 4.4.1]) A retraction on a manifold  $\mathcal{M}$  is a smooth mapping  $\text{Retr}$  from the tangent bundle  $\mathbf{T}\mathcal{M}$  to  $\mathcal{M}$  with the following properties. Let  $\text{Retr}_x : \mathbf{T}_x\mathcal{M} \rightarrow \mathcal{M}$  denote the restriction of  $\text{Retr}$  to the tangent vector space  $\mathbf{T}_x\mathcal{M}$ .

- (i)  $\text{Retr}_x(\mathbf{0}_x) = x$ , where  $\mathbf{0}_x$  is the zero vector in  $\mathbf{T}_x\mathcal{M}$ ;
- (ii) The differential of  $\text{Retr}_x$  at  $\mathbf{0}_x$ ,  $D\text{Retr}_x(\mathbf{0}_x)$ , is the identity map.

LEMMA 5.3 ([9, Lemma 2.7]). Let  $\mathcal{M} \subseteq \mathbb{R}^n$  be a compact Riemannian submanifold. Let  $\text{Retr}$  be a retraction on  $\mathcal{M}$ . Suppose that  $f : \mathcal{M} \rightarrow \mathbb{R}$  has Lipschitz continuous gradient in the convex hull of  $\mathcal{M}$ . Then there exists  $L \geq 0$  such that for all  $x \in \mathcal{M}$  and  $\eta \in \mathbf{T}_x\mathcal{M}$ , it holds that

$$(5.2) \quad |f(\text{Retr}_x(\eta)) - (f(x) + \langle \eta, \text{grad } f(x) \rangle)| \leq \frac{L}{2} \|\eta\|^2,$$

i.e.,  $f(\text{Retr}_x(\eta))$  is uniformly well approximated by its first order approximation.

COROLLARY 5.4. Let  $f$  be any of the functions (2.5) or (2.7). Then there exists a constant  $L \geq 0$  such that the uniform (on  $\mathcal{U}_n$ ) approximation bound holds true.

*Proof.* Note that we can view  $\mathcal{U}_n$  as a real submanifold of  $\mathbb{C}^{n \times n}$ , and its convex hull is compact. The cost functions (2.5) and (2.7) are defined on  $\mathbb{C}^{n \times n}$  and are polynomial in the real and imaginary parts of  $\mathbf{U}$ . This implies Lipschitz continuity of  $f$  on the convex hull of  $\mathcal{U}_n$ , hence Lemma 5.3 can be applied.  $\square$

**5.2. Convergence of Jacobi-G algorithm to stationary points.** We will show in this subsection that the iterations in Algorithm 2.1 are a special case of the iterations in Lemma 5.1, and the convergence results of Lemma 5.1 apply.

PROPOSITION 5.5. Let  $f : \mathcal{U}_n \rightarrow \mathbb{R}$  be one of the functions (2.5) or (2.7), and  $L \geq 0$  be from Corollary 5.4. For Algorithm 2.1, we have:

- (i)  $\|\text{grad } f(\mathbf{U}_k)\| \rightarrow 0$  in Algorithm 2.1; in particular, every accumulation point in Algorithm 2.1 is a stationary point.
- (ii) For  $\delta$  as in (2.13), Algorithm 2.1 needs at most

$$\left\lceil \frac{2L(f^* - f(x_0))}{\delta^2} \frac{1}{\varepsilon^2} \right\rceil$$

iterations to reach an  $\varepsilon$ -optimal solution ( $\|\text{grad } f(\mathbf{U}_k)\| \leq \varepsilon$ ).

*Proof.* We need to show that the ascent conditions are satisfied. Let  $h = h_{(i,j),\mathbf{U}}$  be as in (2.9) and  $\Psi_{\text{opt}}$  be its maximizer. We set

$$\Delta = \mathbf{U} \mathcal{P}_{i,j}^T \mathcal{P}_{i,j}(\Lambda(\mathbf{U})) \in \mathbf{T}_{\mathbf{U}}\mathcal{U}_n,$$

which is a projection of  $\text{grad } f(\mathbf{U})$  onto the tangent space to the submanifold of the matrices of type  $\mathbf{U}\mathbf{G}^{(i,j,\Psi)}$ . Next, denote  $\Psi_1 = \text{Exp}_{\mathbf{I}_2}(\frac{1}{L} \text{grad } h(\mathbf{I}_2))$ . Then, by Lemma 5.3 and Corollary 5.4, we have<sup>14</sup> that

$$\begin{aligned} h(\Psi_{\text{opt}}) - h(\mathbf{I}_2) &\geq h(\Psi_1) - h(\mathbf{I}_2) = f\left(\text{Exp}_{\mathbf{U}}\left(\frac{\Delta}{L}\right)\right) - f(\mathbf{U}) \\ &\geq \left\langle \frac{\Delta}{L}, \text{grad } f(\mathbf{U}) \right\rangle_{\mathbb{R}} - \frac{L}{2} \left\| \frac{\Delta}{L} \right\|^2 = \frac{\|\text{grad } h(\mathbf{I}_2)\|^2}{2}, \end{aligned}$$

<sup>14</sup>Note that the exponential map (3.5) is a retraction (see [3, Proposition 5.4.1]).



where the last equality is from (3.2) and (3.9). Finally, we note that

$$f(\mathbf{U}_k) - f(\mathbf{U}_{k-1}) = h_k(\Psi_k) - h_k(\mathbf{I}_2) \geq \frac{1}{2L} \|\text{grad } h_k(\mathbf{I}_2)\|^2 \geq \frac{\delta^2}{2L} \|\text{grad } f(\mathbf{U}_{k-1})\|^2,$$

and thus the descent condition (5.1) holds with the constant  $\frac{\delta^2}{2L}$ .  $\square$

**6. Łojasiewicz inequality.** In this section, we recall known results and preliminaries that are needed for the main results in Section 7.

**6.1. Łojasiewicz gradient inequality and speed of convergence.** Here we recall the results on convergence of descent algorithms on analytic submanifolds that use Łojasiewicz gradient inequality [39], as presented in [48]. These results were used in [38] to prove the global convergence of Jacobi-G on the orthogonal group.

DEFINITION 6.1 (Łojasiewicz gradient inequality, [47, Definition 2.1]). *Let  $\mathcal{M} \subseteq \mathbb{R}^n$  be a Riemannian submanifold of  $\mathbb{R}^n$ . The function  $f : \mathcal{M} \rightarrow \mathbb{R}$  satisfies a Łojasiewicz gradient inequality at a point  $x \in \mathcal{M}$ , if there exist  $\delta > 0$ ,  $\sigma > 0$  and  $\zeta \in (0, \frac{1}{2}]$  such that for all  $y \in \mathcal{M}$  with  $\|y - x\| < \delta$ , it holds that*

$$(6.1) \quad |f(x) - f(y)|^{1-\zeta} \leq \sigma \|\text{grad } f(y)\|.$$

The following lemma guarantees that (6.1) is satisfied for the real analytic functions defined on an analytic manifold.

LEMMA 6.2 ([47, Proposition 2.2 and Remark 1]). *Let  $\mathcal{M} \subseteq \mathbb{R}^n$  be an analytic submanifold<sup>15</sup> and  $f : \mathcal{M} \rightarrow \mathbb{R}$  be a real analytic function. Then for any  $x \in \mathcal{M}$ ,  $f$  satisfies a Łojasiewicz gradient inequality (6.1) for some<sup>16</sup>  $\delta, \sigma > 0$  and  $\zeta \in (0, \frac{1}{2}]$ .*

Łojasiewicz gradient inequality allows for proving convergence of optimization algorithms to a single limit point.

THEOREM 6.3 ([47, Theorem 2.3]). *Let  $\mathcal{M} \subseteq \mathbb{R}^n$  be an analytic submanifold and  $\{x_k : k \in \mathbb{N}\} \subseteq \mathcal{M}$ . Suppose that  $f$  is real analytic and, for large enough  $k$ ,*

$$(6.2) \quad |f(x_{k+1}) - f(x_k)| \geq \sigma \|\text{grad } f(x_k)\| \|x_{k+1} - x_k\|;$$

(ii)  $\text{grad } f(x_k) = 0$  implies that  $x_{k+1} = x_k$ .

Then any accumulation point  $x_*$  of  $\{x_k : k \in \mathbb{N}\} \subseteq \mathcal{M}$  is the only limit point.

If, in addition, for some  $\kappa > 0$  and for large enough  $k$  it holds that

$$(6.3) \quad \|x_{k+1} - x_k\| \geq \kappa \|\text{grad } f(x_k)\|,$$

then the following convergence rates apply

$$\|x_k - x_*\| \leq C \begin{cases} e^{-ck}, & \text{if } \zeta = \frac{1}{2} \text{ (for some } c > 0), \\ k^{-\frac{\zeta}{1-2\zeta}}, & \text{if } 0 < \zeta < \frac{1}{2}, \end{cases}$$

where  $\zeta$  is the parameter in (6.1) at the limit point  $x_*$ .

Remark 6.4. We can relax the conditions of Theorem 6.3 as follows. We can require just that (6.2) holds for all  $k$  such that  $\|x_k - x_*\| < \varepsilon$ , where  $x_*$  is an accumulation point of the sequence and  $\varepsilon > 0$  is some radius. This can be verified by inspecting the proof of Theorem 6.3 (see also the proof of [2, Theorem 3.2])

<sup>15</sup>See [35, Definition 2.7.1] or [38, Definition 5.1] for a definition of an analytic submanifold.

<sup>16</sup>The values of  $\delta, \sigma, \zeta$  depend on a specific point.

In the case  $\zeta = \frac{1}{2}$ , according to [Theorem 6.3](#), the convergence is linear (similarly to the classic results on local convergence of the gradient descent algorithm [\[45, 11\]](#)). In the optimization literature, the inequality [\(6.1\)](#) with  $\zeta = \frac{1}{2}$  is often called *Polyak-Lojasiewicz inequality*<sup>17</sup>. In the next subsection, we recall some sufficient conditions for Polyak-Lojasiewicz inequality to hold.

**6.2. Lojasiewicz inequality at stationary points.** It is known, and widely used in optimization (especially in the Euclidean case), that around a strong local maximum the function satisfies the Polyak-Lojasiewicz inequality. In fact, it is also valid for non-degenerate stationary points, as shown in [\[31\]](#). Here we recall the most general recent result on possibly degenerate stationary points that satisfy the so-called Morse-Bott property (see also [\[8, p.248\]](#)).

**DEFINITION 6.5** ([\[23, Definition 1.5\]](#)). *Let  $\mathcal{M}$  be a  $C^\infty$  submanifold and  $f : \mathcal{M} \rightarrow \mathbb{R}$  be a  $C^2$  function. Denote the set of stationary points as*

$$\text{Crit} f = \{x \in \mathcal{M} : \text{grad} f(x) = 0\}.$$

*The function  $f$  is said to be Morse-Bott at  $x_0 \in \mathcal{M}$  if there exists an open neighborhood  $\mathcal{U} \subseteq \mathcal{M}$  of  $x_0$  such that*

- (i)  $\mathcal{C} = \mathcal{U} \cap \text{Crit} f$  is a relatively open, smooth submanifold of  $\mathcal{M}$ ;
- (ii)  $\mathbf{T}_{x_0} \mathcal{C} = \text{Ker Hess}_{x_0} f$ .

*Remark 6.6.* (i) If  $x_0 \in \mathcal{M}$  is a non-degenerate stationary point, then  $f$  is Morse-Bott at  $x_0$ , since  $\{x_0\}$  is a zero-dimensional manifold in this case.

(ii) If  $x_0 \in \mathcal{M}$  is a degenerate stationary point, then condition (ii) in [Definition 6.5](#) can be rephrased<sup>18</sup> as

$$(6.4) \quad \text{rank}\{\text{Hess}_{x_0} f\} = \dim \mathcal{M} - \dim \mathcal{C}.$$

For the functions that satisfy the Morse-Bott property, it was recently shown that the Polyak-Lojasiewicz inequality holds true.

**THEOREM 6.7** ([\[23, Theorem 3, Corollary 5\]](#)). *If  $\mathcal{U} \subseteq \mathbb{R}^n$  is an open subset and  $f : \mathcal{U} \rightarrow \mathbb{R}$  is Morse-Bott at a stationary point  $x$ , then there exist  $\delta, \sigma > 0$  such that*

$$|f(y) - f(x)| \leq \sigma \|\nabla f(y)\|^2,$$

*for any  $y \in \mathcal{U}$  satisfying  $\|y - x\| \leq \delta$ .*

We can also easily deduce the same result on a smooth manifold  $\mathcal{M}$ .

**PROPOSITION 6.8.** *If  $\mathcal{U} \subseteq \mathcal{M}$  is an open subset and a  $C^2$  function  $f : \mathcal{U} \rightarrow \mathbb{R}$  is Morse-Bott at a stationary point  $x$ , then there exist an open neighborhood  $\mathcal{V} \subseteq \mathcal{U}$  of  $x$  and  $\sigma > 0$  such that for all  $y \in \mathcal{V}$  it holds that*

$$|f(y) - f(x)| \leq \sigma \|\text{grad} f(y)\|^2.$$

*Proof.* Consider the exponential map  $\text{Exp}_x : \mathbf{T}_x \mathcal{M} \rightarrow \mathcal{M}$ , which is a local diffeomorphism. Let  $\mathcal{W} \subseteq \mathbf{T}_x \mathcal{M}$  be an open subset such that  $\text{Exp}_x(\mathcal{W}) = \mathcal{U}$ . Let  $\hat{f} = f \circ \text{Exp}_x$  be the composite map from  $\mathcal{W}$  to  $\mathbb{R}$ . Then

$$(6.5) \quad \nabla \hat{f}(y') = \mathbf{J}_{\text{Exp}_x}^\top(y') \text{grad} f(y),$$

<sup>17</sup>The inequality [\(6.1\)](#) with  $\zeta = \frac{1}{2}$  goes back to Polyak [\[45\]](#), who used it for proving linear convergence of the gradient descent algorithm.

<sup>18</sup>due to the fact that  $\mathbf{T}_{x_0} \mathcal{C} \subseteq \text{Ker Hess}_{x_0} f$ .

where  $y' \in \mathcal{W}$  and  $y = \text{Exp}_x(y')$ . It follows that  $\text{Exp}_x$  gives a diffeomorphism between  $\text{Crit}f$  and  $\text{Crit}\hat{f}$ . Since  $\text{Hess}_x f = \text{H}_{\hat{f}}(0)$  by [3, Proposition 5.5.5], we have that  $\hat{f}$  is Morse-Bott at 0. Therefore, by Theorem 6.7, there exist  $\sigma' > 0$ ,  $\sigma > 0$  and an open neighborhood  $\mathcal{V} \subseteq \mathcal{U}$  of  $x$  such that

$$|f(y) - f(x)| = |\hat{f}(y') - \hat{f}(0)| \leq \sigma' \|\nabla \hat{f}(y')\|^2 \leq \sigma \|\text{grad} f(y)\|^2,$$

for any  $y \in \mathcal{V}$ , where the last inequality holds because  $\mathbf{J}_{\text{Exp}_x}$  is nonsingular in a neighborhood of  $x$ .  $\square$

*Remark 6.9.* For the case of non-degenerate stationary points and  $C^\infty$  functions, Proposition 6.8 is proved in [31, Lemma 4.1], which is a simple corollary of Morse Lemma [43, Lemma 2.2]. For  $C^\infty$  functions and Morse-Bott functions, Proposition 6.8 (as noted in [23]) is also a simple corollary of Morse-Bott Lemma [7].

*Remark 6.10.* Morse-Bott property is known to be useful for studying convergence properties. For example, it is shown in [29, Appendix C] that if the cost function is (globally) Morse-Bott, i.e., satisfies the Morse-Bott property at all the stationary point, then the continuous gradient flow converges to a single point.

Finally, we recall an important property of non-degenerate local maxima, which follows from the classic Morse Lemma [43].

LEMMA 6.11. *Let  $x$  be a non-degenerate local maximum (according to Definition 3.7) of a smooth function  $f$  such that  $f(x) = c$ . Then there exists a simply connected open neighborhood  $\mathcal{W}$  of  $x$  such that*

- $x$  is the only critical point in  $\mathcal{W}$ ;
- its boundary is a level curve (i.e  $f(y) = a < c$ , for all  $y \in \delta(\mathcal{W})$ );
- the superlevel sets  $\mathcal{W}_b = \{x \in \mathcal{W}, f(x) \geq b > a\}$  are simply connected and nested.

*Remark 6.12.* In Lemma 6.11, we can also select the neighborhood in such a way that Hessian is negative definite at each point  $y$ , which implies that for any geodesic<sup>19</sup>  $\gamma(t)$  passing through  $y$ ,  $\gamma(0) = y$ , the second derivative of  $f(\gamma(t))$  at 0 is negative.

## 7. Convergence results based on Łojasiewicz inequality.

**7.1. Preliminary lemmas: checking the decrease conditions.** In this subsection, we are going to find some sufficient conditions for (6.2) and (6.3) to hold in Algorithm 2.1, which will allow us to use Theorem 6.3.

Let  $\mathbf{U}_k = \mathbf{U}_{k-1} \mathbf{G}^{(i_k, j_k, \Psi_k)}$  be the iterations in Algorithm 2.1. Obviously,

$$\|\mathbf{U}_k - \mathbf{U}_{k-1}\| = \|\Psi_k - \mathbf{I}_2\|.$$

Assume that  $\Psi_k$  is obtained as in Proposition 4.4, i.e., by taking  $\mathbf{w}$  as the leading eigenvector of  $\Gamma^{(i_k, j_k, \mathbf{U}_{k-1})}$  (normalized so that  $w_1 = \cos 2\theta = 2c^2 - 1 > 0$  in (4.4)) as in Remark 4.6, and retrieving  $\Psi_k$  from  $\mathbf{w}$  according to (2.10) and (4.4). We first express  $\|\Psi_k - \mathbf{I}_2\|$  through  $w_1$ .

LEMMA 7.1. *For the iterations  $\Psi_k$  obtained as in Proposition 4.4, it holds that*

$$(7.1) \quad \sqrt{2} \|\Psi_k - \mathbf{I}_2\| \geq \sqrt{1 - w_1^2} \geq \frac{\sqrt{\sqrt{2} + 2}}{2} \|\Psi_k - \mathbf{I}_2\|$$

<sup>19</sup>A related discussion on geodesic convexity of functions can be found in [46].

*Proof.* Note that

$$\|\Psi_k - \mathbf{I}_2\| = \left\| \begin{bmatrix} c-1 & -s \\ s^* & c-1 \end{bmatrix} \right\| = \sqrt{2(1-c)^2 + 2(1-c^2)} = 2\sqrt{1-c}.$$

Next, we note that  $\sqrt{1-w_1^2} = 2c\sqrt{1-c^2}$  and

$$\frac{\sqrt{1-w_1^2}}{\|\Psi_k - \mathbf{I}_2\|} = \frac{2c\sqrt{1-c^2}}{2\sqrt{1-c}} = c\sqrt{1+c}.$$

By [Remark 4.6](#), we have  $c \in [\frac{1}{\sqrt{2}}; 1]$ , hence the ratio can be bounded from above by its values at the endpoints of the interval.  $\square$

Since we are looking at [Algorithm 2.1](#), we can replace in both inequalities of (7.1)  $\text{grad } f(\mathbf{U}_{k-1})$  with  $\text{grad } h_{(i,j),\mathbf{U}}(\mathbf{I}_2)$ . Next, we prove a result for condition (6.3).

**LEMMA 7.2.** *Let  $f : \mathcal{U}_n \rightarrow \mathbb{R}$  be as in [Proposition 4.4](#). Then there exists a universal constant  $\kappa > 0$  such that*

$$\|\Psi_k - \mathbf{I}_2\| \geq \kappa \|\text{grad } h_k(\mathbf{I}_2)\|.$$

*Proof.* We denote  $\mathbf{\Gamma} = \mathbf{\Gamma}^{(i_k, j_k, \mathbf{U}_{k-1})}$  as in (4.3). By [Lemma 4.7](#), we have that

$$\|\text{grad } h_k(\mathbf{I}_2)\| = 2\sqrt{2}\sqrt{\Gamma_{12}^2 + \Gamma_{13}^2}.$$

By [Lemma 7.1](#), it is sufficient to prove that

$$1 - w_1^2 \geq \kappa'(\Gamma_{12}^2 + \Gamma_{13}^2)$$

for a universal constant  $\kappa' > 0$ . Let  $\lambda_1 \geq \lambda_2 \geq \lambda_3$  be the eigenvalues of  $\mathbf{\Gamma}$ . Without loss of generality, we set  $\mathbf{\Gamma}' = \mathbf{\Gamma} - \lambda_3 \mathbf{I}_3$ ,  $\mu_1 = \lambda_1 - \lambda_3$  and  $\mu_2 = \lambda_2 - \lambda_3$ . Then

$$(7.2) \quad \mathbf{\Gamma}' = \mu_1 \mathbf{w} \mathbf{w}^\top + \mu_2 \mathbf{v} \mathbf{v}^\top,$$

where  $\mathbf{v}$  is the second eigenvector of  $\mathbf{\Gamma}$ . It follows that

$$(7.3) \quad \begin{aligned} \Gamma_{12}^2 + \Gamma_{13}^2 &= (\Gamma'_{12})^2 + (\Gamma'_{13})^2 = (\mu_1 w_1 w_2 + \mu_2 v_1 v_2)^2 + (\mu_1 w_1 w_3 + \mu_2 v_1 v_3)^2 \\ &= \mu_1^2 w_1^2 (w_2^2 + w_3^2) + 2\mu_1 \mu_2 w_1 v_1 (w_2 v_2 + w_3 v_3) + \mu_2^2 v_1^2 (v_2^2 + v_3^2) \\ &= \mu_1^2 w_1^2 (1 - w_1^2) - 2\mu_1 \mu_2 w_1^2 v_1^2 + \mu_2^2 v_1^2 (1 - v_1^2) \\ &\leq (1 - w_1^2)(\mu_1^2 + \mu_2^2), \end{aligned}$$

where the last equality and inequality is due to orthonormality of  $\mathbf{v}$  and  $\mathbf{w}$  (which implies  $v_1^2 \leq 1 - w_1^2$ ). By expanding  $\mu_1$  and  $\mu_2$ , it is not difficult to verify that  $\mu_1^2 + \mu_2^2 \leq 2\|\mathbf{\Gamma}\|^2$ . Finally, by [Theorem 4.16](#), the elements of  $\mathbf{\Gamma}$  continuously depend on  $\mathbf{U} \in \mathcal{U}_n$ . Therefore,  $\|\mathbf{\Gamma}\|$  is bounded from above, and thus the proof is completed.  $\square$

We are ready to check the sufficient decrease condition (6.2).

**LEMMA 7.3.** *Let  $\mathbf{\Gamma} = \mathbf{\Gamma}^{(i_k, j_k, \mathbf{U}_{k-1})}$  be as in (4.3). Let  $\lambda_1 \geq \lambda_2 \geq \lambda_3$  be the eigenvalues of  $\mathbf{\Gamma}$ , and  $\eta = \frac{\lambda_2 - \lambda_3}{\lambda_1 - \lambda_3}$ . Suppose that  $1 - \eta \geq \varepsilon$  for some  $\varepsilon > 0$ . Then*

$$|h_k(\Psi_k) - h_k(\mathbf{I}_2)| \geq \frac{\varepsilon}{4} \|\text{grad } h_k(\mathbf{I}_2)\| \sqrt{1 - w_1^2}.$$

*Proof.* Define the ratio

$$(7.4) \quad q(\mathbf{\Gamma}, \mathbf{w}) = \frac{(\mathbf{w}^\top \mathbf{\Gamma} \mathbf{w} - \Gamma_{11})^2}{(\Gamma_{12}^2 + \Gamma_{13}^2)(1 - w_1^2)}.$$

It is sufficient to prove that  $q(\mathbf{\Gamma}, \mathbf{w}) \geq \varepsilon^2/2$ . Denote  $\rho \stackrel{\text{def}}{=} 1 - w_1^2 \geq v_1^2$ , where  $\mathbf{v}$  is as in the proof of [Lemma 7.2](#). From (7.2) and (7.3) we immediately have

$$(7.5) \quad \mathbf{w}^\top \mathbf{\Gamma} \mathbf{w} - \Gamma_{11} = \mu_1 - (\mu_1 w_1^2 + \mu_2 v_1^2) = \mu_1(\rho - \eta v_1^2) \geq \mu_1 \rho(1 - \eta),$$

$$(7.6) \quad \Gamma_{12}^2 + \Gamma_{13}^2 \leq \rho \mu_1^2(1 + \eta^2).$$

Using (7.5) and (7.6), we get

$$\frac{1}{q(\mathbf{\Gamma}, \mathbf{w})} = \frac{(\Gamma_{12}^2 + \Gamma_{13}^2)\rho}{(\mathbf{w}^\top \mathbf{\Gamma} \mathbf{w} - \Gamma_{11})^2} \leq \frac{\rho^2 \mu_1^2(1 + \eta^2)}{\rho^2 \mu_1^2(1 - \eta)^2} \leq \frac{2}{\varepsilon^2}.$$

The proof is complete.  $\square$

## 7.2. Main results.

**THEOREM 7.4.** *Let  $f : \mathcal{U}_n \rightarrow \mathbb{R}$  be as in [Proposition 4.4](#), and  $\overline{\mathbf{U}}$  be an accumulation point of [Algorithm 2.1](#) (and  $\text{grad } f(\overline{\mathbf{U}}) = 0$  by [Proposition 5.5](#)). Assume that  $\mathfrak{D}_{\overline{\mathbf{U}}}^{(i,j)}$  defined in (4.6) is negative definite for all  $i < j$ . Then*

- (i)  $\overline{\mathbf{U}}$  is the only limit point and convergence rates in [Theorem 6.3](#) apply.
- (ii) If the rank of Riemannian Hessian is maximal at  $\overline{\mathbf{U}}$  (i.e.,  $\text{rank}\{\text{Hess}_{\overline{\mathbf{U}}} f\} = n(n-1)$ ), then the speed of convergence is linear.

*Proof.* (i) Since  $\mathfrak{D}_{\overline{\mathbf{U}}}^{(i,j)}$  is negative definite for any  $i \neq j$ , the two top eigenvalues of  $\mathbf{\Gamma}^{(i,j;\overline{\mathbf{U}})}$  are separated by [Remark 4.10](#). Therefore, there exists  $\varepsilon > 0$  such that

$$\frac{\lambda_2(\mathbf{\Gamma}^{(i,j;\overline{\mathbf{U}})}) - \lambda_3(\mathbf{\Gamma}^{(i,j;\overline{\mathbf{U}})})}{\lambda_1(\mathbf{\Gamma}^{(i,j;\overline{\mathbf{U}})}) - \lambda_3(\mathbf{\Gamma}^{(i,j;\overline{\mathbf{U}})})} < 1 - \varepsilon.$$

By the continuity of  $\mathbf{\Gamma}^{(i,j;\mathbf{U})}$  with respect to  $\mathbf{U}$ , the conditions of [Lemma 7.3](#) are satisfied in a neighborhood of  $\overline{\mathbf{U}}$ . Therefore, there exists  $c > 0$  such that

$$\begin{aligned} |f(\mathbf{U}_k) - f(\mathbf{U}_{k-1})| &\geq c \|\text{grad } h_k(\mathbf{I}_2)\| \|\mathbf{U}_k - \mathbf{U}_{k-1}\| \\ &\geq c\delta \|\text{grad } f(\mathbf{U}_{k-1})\| \|\mathbf{U}_k - \mathbf{U}_{k-1}\|, \end{aligned} \quad \square$$

in a neighborhood of  $\overline{\mathbf{U}}$  by [Lemma 7.3](#), [Lemma 7.1](#) and (2.13). By [Remark 6.4](#), it is enough to use [Theorem 6.3](#), hence  $\overline{\mathbf{U}}$  is the only limit point. Moreover, by [Lemma 7.2](#) and (2.13), the convergence rates apply.

- (ii) Due to the scaling invariance,  $\overline{\mathbf{U}}$  belongs to an  $n$ -dimensional submanifold of stationary points defined by  $\overline{\mathbf{U}}\mathbf{S}$ , where  $\mathbf{S}$  is as in (2.11). Since  $\text{rank}\{\text{Hess}_{\overline{\mathbf{U}}} f\} = n(n-1)$ ,  $f$  is Morse-Bott at  $\overline{\mathbf{U}}$  by [Remark 6.6](#). Therefore, by [Proposition 6.8](#),  $\zeta = 1/2$  in (6.1) at  $\overline{\mathbf{U}}$ , and thus the convergence is linear by [Theorem 6.3](#).

**THEOREM 7.5.** *Let  $f$  be as in [Theorem 7.4](#), and  $\mathbf{U}_*$  be a semi-strict local maximum point of  $f$  (i.e.,  $\text{rank}\{\text{Hess}_{\mathbf{U}_*} f\} = n(n-1)$ ). Then there exists a neighborhood  $\mathcal{W}$  of  $\mathbf{U}_*$ , such that for any starting point  $\mathbf{U}_0 \in \mathcal{W}$ , [Algorithm 2.1](#) converges linearly to  $\mathbf{U}_*\mathbf{S}$ , where  $\mathbf{S}$  is of the form (2.11).*

*Proof.* Let  $\tilde{\mathcal{U}}_n$  be the quotient manifold defined in subsection 3.4. By Lemma 3.8 we have that  $\text{rank}\{\text{Hess}_{\tilde{\mathcal{U}}_*} f\} = \text{rank}\{\text{Hess}_{\mathcal{U}_*} f\} = n(n-1)$ , and therefore it is negative definite. Let us take the open neighborhood  $\tilde{\mathcal{W}}$  of  $\tilde{\mathcal{U}}_*$  as in Lemma 6.11. For simplicity assume that  $f(\mathcal{U}_*) = 0$ .

Next, assume that  $\tilde{\mathcal{U}}_{k-1} \in \tilde{\mathcal{W}}$ , and consider the  $\mathcal{U}_k = \mathcal{U}_{k-1} \mathbf{G}^{(i_k, j_k, \Psi_k)}$  with  $\Psi_k$  given as the maximizer of (4.3). Let  $b = f(\mathcal{U}_{k-1})$ . In what follows, we are going to prove that  $\tilde{\mathcal{U}}_k \in \tilde{\mathcal{W}}_b$  (defined as in Lemma 6.11), so that the sequence  $\tilde{\mathcal{U}}_k$  never leaves the set  $\tilde{\mathcal{W}}$ .

Recall that  $\Psi_k$  is computed as follows (see Remark 4.6): take the vector  $\mathbf{w}$  as in (4.4). Take  $\alpha_1 = -w_2/\sqrt{1-w_1^2}$ ,  $\alpha_2 = -w_3/\sqrt{1-w_1^2}$  (we can assume  $w_1 \neq 1$  because otherwise  $\Psi_k = \mathbf{I}_2$  and this case is trivial), and consider the following geodesic in  $\mathcal{U}_n$ :

$$\gamma(t) = \text{Exp}_{\mathcal{U}_{k-1}}(\mathcal{U}_{k-1} \mathcal{P}_{i,j}^\top(\Omega t)) = \mathcal{U}_{k-1} \mathbf{G}^{(i_k, j_k, \exp(\Omega t))},$$

where  $\Omega \in \mathbf{T}_{\mathcal{I}_2} \tilde{\mathcal{U}}_2$  is defined as in (4.7). The geodesic starts at  $\gamma(0) = \mathcal{U}_{k-1}$ , and reaches  $\gamma(t_*) = \mathcal{U}_k$  at  $t_* = \arccos(w_1) \in (0, \frac{\pi}{2}]$ . Note that by Remark 3.4, the corresponding curve  $\tilde{\gamma}$  is a geodesic in the quotient manifold  $\tilde{\mathcal{U}}_n$ .

Next, from (4.3) (applied to  $\Gamma = \Gamma^{(i_k, j_k, \mathcal{U}_{k-1})}$ ) we have that

$$f(\gamma(t)) = h(e^{\Omega t}) = [\cos t \quad -\alpha_1 \sin t \quad -\alpha_2 \sin t] \Gamma [\cos t \quad -\alpha_1 \sin t \quad -\alpha_2 \sin t]^\top + C,$$

hence  $f(\gamma(t))$  can be represented (for some constants  $A, C_1$ ) as

$$f(\gamma(t)) = A \cos(2(t - t_*)) + C_1;$$

note that  $A > 0$  since  $t_* > 0$  is the maximizer.

Next, by Remark 6.12, we should have  $\frac{d}{dt} f(\gamma(0)) = -4A \cos(-2t_*) < 0$ , which implies  $\cos(2t_*) > 0$ . Thus, we can further reduce the domain where  $t_*$  is located to  $t_* \in (0, \frac{\pi}{4}]$ . Hence we have that  $\frac{d}{dt} f(\gamma(t)) = -4A \sin(2(t - t_*)) > 0$  for any  $t \in [0, t_*)$ , and thus the cost function is increasing; note that  $\frac{d}{dt} f(\gamma(t_*)) = 0$  and there are no other stationary points in  $t \in [0, t_*)$ .

Next, by continuity and because  $\tilde{\mathcal{W}}$  is open, there exists a small  $\varepsilon > 0$  such that  $\tilde{\gamma}(\varepsilon)$  is in the interior of  $\tilde{\mathcal{W}}_b$ . By periodicity of  $f(\gamma(t))$  and continuity, we have that there exists  $t_2$  such that  $\tilde{\gamma}(t_2) \in \delta(\tilde{\mathcal{W}}_b)$  and  $\tilde{\gamma}(t) \in \tilde{\mathcal{W}}_b$  for all  $t \in [0, t_2]$ . By Rolle's theorem, there exists a local maximum of  $f(\gamma(t))$  in  $[0, t_2]$ . Note that by construction, the closest positive local maximum to 0 is at  $t_*$ . Therefore  $\tilde{\mathcal{U}}_k = \tilde{\gamma}(t_*) \in \tilde{\mathcal{W}}_b$ , hence we stay in the same neighborhood  $\tilde{\mathcal{W}}$ .

Finally, as a neighborhood of  $\mathcal{U}_* \in \mathcal{U}_n$ , we can take the preimage  $\mathcal{W} = \pi^{-1}(\tilde{\mathcal{W}})$ ; also linear convergence rate follows from Theorem 7.4. The proof is complete.  $\square$

**7.3. Examples of cost functions satisfying regularity conditions.** In this subsection, we provide examples when the regularity conditions of Theorems 7.4 and 7.5 are satisfied for diagonalizable tensors and matrices at the diagonalizing rotation. Recall that  $\mathcal{A} \in \mathbb{C}^{n \times \dots \times n}$  is a diagonal tensor if all the elements are zero except  $\text{diag}\{\mathcal{A}\}$ .

PROPOSITION 7.6. (i) For a set of jointly orthogonally diagonalizable matrices

$$\mathbf{A}^{(\ell)} = \mathcal{U}_* \begin{bmatrix} \mu_1^{(\ell)} & & 0 \\ & \ddots & \\ 0 & & \mu_n^{(\ell)} \end{bmatrix} \mathcal{U}_*^H,$$

such that for any pair  $i \neq j$ ,

$$\sum_{\ell=1}^L (\mu_i^{(\ell)} - \mu_j^{(\ell)})^2 > 0,$$

the matrix  $\mathbf{U}_*$  is a semi-strict (as in [Theorem 7.5](#)) local maximum point of the cost function (2.1).

(ii) For an orthogonally diagonalizable 3rd order tensor

$$\mathcal{A} = \mathcal{D} \bullet_1 \mathbf{U}_* \bullet_2 \mathbf{U}_*^* \bullet_3 \mathbf{U}_*^*,$$

where  $\mathcal{D}$  is a diagonal tensor with at most one zero element on the diagonal, the matrix  $\mathbf{U}_*$  is a semi-strict local maximum point of the cost function (2.2).

(iii) For an orthogonally diagonalizable 4th order tensor

$$\mathcal{A} = \mathcal{D} \bullet_1 \mathbf{U}_* \bullet_2 \mathbf{U}_* \bullet_3 \mathbf{U}_*^* \bullet_4 \mathbf{U}_*^*,$$

where  $\mathcal{D}$  is a diagonal tensor, the values on the diagonal are either (a) all positive or (b) there is at most one  $i$  with  $\mathcal{D}_{iiii} \leq 0$ , for which  $\mathcal{D}_{iiii} + \mathcal{D}_{jjjj} > 0$  for all  $j \neq i$ , the matrix  $\mathbf{U}_*$  is a semi-strict local maximum point of the function (2.3).

For proving [Proposition 7.6](#), we need a lemma about Hessians of multilinear forms.

LEMMA 7.7. Let  $\gamma(\mathbf{u})$  be a Hermitian form of order  $2d$   $\gamma(\mathbf{u}) = g_{\mathcal{B},t}(\mathbf{u})$ , where  $\mathcal{B}$  is diagonal tensor. Then for any distinct indices  $1 \leq i \neq j \neq k \leq n$  it holds that

$$\frac{\partial^2 \gamma}{\partial u_i^* \partial u_j^*}(\mathbf{e}_k) = \frac{\partial^2 \gamma}{\partial u_i^* \partial u_j}(\mathbf{e}_k) = 0.$$

*Proof.* By continuing differentiation as in [Lemma 4.12](#), we get that

$$\begin{aligned} \mathbf{e}_i^\top \left( \frac{\partial^2 \gamma}{\partial \mathbf{u}^* \partial \mathbf{u}^*}(\mathbf{e}_k) \right) \mathbf{e}_j &= \sum_{\substack{s \neq p \\ 1 \leq s, p \leq d}} (\mathcal{B} \bullet_s \mathbf{e}_i \bullet_p \mathbf{e}_j)_{k\dots k} = 0, \\ \mathbf{e}_i^\top \left( \frac{\partial^2 \gamma}{\partial \mathbf{u}^* \partial \mathbf{u}}(\mathbf{e}_k) \right) \mathbf{e}_j &= \sum_{s=1}^d \sum_{p=d+1}^{2d} (\mathcal{B} \bullet_s \mathbf{e}_i \bullet_p \mathbf{e}_j)_{k\dots k} = 0, \end{aligned}$$

which completes the proof.  $\square$

*Proof of [Proposition 7.6](#).* Without loss of generality, we can consider only the case  $\mathbf{U}_* = \mathbf{I}_n$ , so that all the matrices/tensors are diagonal. Due to diagonality of matrices/tensors (the off-diagonal elements are zero) from [Proposition 4.14](#) we have that  $\mathbf{I}_n$  is a stationary point and the Euclidean gradient  $\nabla^{(\mathbb{R})} f(\mathbf{I}_n)$  is a diagonal matrix that contains  $2d \text{diag}\{\mathcal{B}\}$  on its diagonal. Moreover, by [4, Eq. (8)–(10)] the Riemannian Hessian is a sum of the projection of the Euclidean Hessian on the tangent space and a second term given by the Weingarten operator

$$(7.7) \quad \text{Hess}_{\mathbf{I}_n} f[\eta] = \Pi_{\mathbf{T}_{\mathbf{I}_n} \mathcal{U}_n} \mathbf{H}_f(\mathbf{I}_n)[\eta] + \mathfrak{A}_{\mathbf{I}_n}(\eta, \Pi_{(\mathbf{T}_{\mathbf{I}_n} \mathcal{U}_n)^\perp} \nabla^{(\mathbb{R})} f(\mathbf{I}_n)),$$

where  $\mathbf{H}_f$  is the Euclidean Hessian of  $f$ , and the Weingarten operator for  $\mathcal{U}_n$  (similarly to the case of orthogonal group [4]) is given by

$$\mathfrak{A}_{\mathbf{U}}(\mathbf{Z}, \mathbf{V}) = \mathbf{U} \frac{1}{2} \left( \mathbf{Z}^H \mathbf{V} - \mathbf{V}^H \mathbf{Z} \right).$$

First, we show that the Euclidean Hessian does not contain off-diagonal blocks. From (7.7), we just need to look at the Euclidean Hessian. Take two pairs of indices  $(i, k)$  and  $(j, l)$  and look at the second-order Wirtinger derivatives

$$\frac{\partial^2 f}{\partial U_{i,k}^* \partial U_{j,l}} \quad \text{and} \quad \frac{\partial^2 f}{\partial U_{i,k}^* \partial U_{j,l}^*}.$$

Since by (4.14),  $(\nabla_{U^*} f)_{i,k}$  is a function of  $\mathbf{u}_k$  only, these terms can only be nonzero if  $j = k$  or  $l = k$ . Let us choose  $l = k$  (and  $i \neq j$ ). In that case, by Lemma 7.7,

$$\frac{\partial^2 f}{\partial U_{i,k}^* \partial U_{j,k}}(\mathbf{I}_n) = \frac{\partial^2 \gamma}{\partial u_i^* \partial u_j}(\mathbf{e}_k) = 0, \quad \frac{\partial^2 f}{\partial U_{i,k}^* \partial U_{j,k}^*}(\mathbf{I}_n) = \frac{\partial^2 \gamma}{\partial u_i^* \partial u_j^*}(\mathbf{e}_k) = 0.$$

Similarly, we can show that off-diagonal blocks in the second Hessian term is also equal to zero. Indeed, take  $\mathbf{Z} = \mathcal{P}_{i,k}^\top(\Psi_1)$ , where  $\Psi_1$  is a  $2 \times 2$  skew-Hermitian matrix. Recall that  $\mathbf{V} = \Pi_{(\mathbf{T}_{\mathbf{I}_n} u_n)_\perp} \nabla^{(\Re)} f(\mathbf{I}_n) = \nabla^{(\Re)} f(\mathbf{I}_n)$  is diagonal, hence  $\mathbf{A} = \frac{\mathbf{Z}^H \mathbf{V} - \mathbf{V}^H \mathbf{Z}}{2} = \mathcal{P}_{i,k}^\top(\Psi_2)$  for some  $2 \times 2$  skew-Hermitian matrix  $\Psi_2$ . In this case, if  $(j, l) \neq (i, k)$ , then  $\langle \mathbf{A}, \mathcal{P}_{j,l}^\top(\Psi_3) \rangle_{\Re} = 0$  for any  $2 \times 2$  skew-Hermitian  $\Psi_3$ . Thus the Riemannian Hessian is block-diagonal with the terms given in Lemma 4.8.

Finally, we apply Proposition 4.4 and get that

- (i)  $\mathfrak{D}_{\mathbf{I}_n}^{(i,j)} = -\mathbf{I}_2 \sum_{\ell=1}^L (\mu_i^{(\ell)} - \mu_j^{(\ell)})^2$  for the cost function (2.1);
- (ii)  $\mathfrak{D}_{\mathbf{I}_n}^{(i,j)} = -\frac{3}{2} \mathbf{I}_2 (|\mathcal{D}_{iii}|^2 + |\mathcal{D}_{jjj}|^2)$  for the cost function (2.2);
- (iii)  $\mathfrak{D}_{\mathbf{I}_n}^{(i,j)} = -\mathbf{I}_2 (\mathcal{D}_{iiii} + \mathcal{D}_{jjjj})$  for the cost function (2.3).

It is easy to check that, in all three cases,  $\mathfrak{D}_{\mathbf{I}_n}^{(i,j)}$  is negative definite for any  $i \neq j$  and only if the conditions of the proposition are satisfied. The proof is complete.  $\square$

**8. Implementation details and experiments.** In this section, we comment on implementation details for Algorithm 2.1 and Jacobi-type methods in general. Note that implementations of Jacobi-type methods [14, 15, 19, 16] for the cyclic order of pairs are widely available, but they are often tailored to source separation problems and use implicit calculations. The codes reproducing experiments in this section are publicly available at <https://github.com/kdu/jacobi-G-unitary-matlab> (implemented in MATLAB, version R2019b). Note that some experiments for the orthogonal group are available in [38].

**8.1. Implementation and computational complexity.** Consider the general problem of maximizing (2.5) (for  $d \leq 3$ ). Note that the Givens rotations (from Theorem 4.16), as well as the Riemannian gradient (from Proposition 4.14), are expressed in terms of the rotated tensors. This leads to the following practical modification of Algorithm 2.1: instead of updating  $\mathbf{U}_k = \mathbf{U}_{k-1} \mathbf{G}^{(i_k, j_k, \Psi_k)}$ , we can transform the tensors themselves. We summarize this idea in Algorithm 8.1 for the case  $d = 2$  (simultaneous diagonalization of matrices), and the cost function (2.1).

Let us comment on the complexities of the steps (in what follows, we only count numbers of complex multiplications). Some basic comments first:

- We can assume that the complexity of step 4 is constant  $O(1)$ : indeed, by Theorem 4.16, an eigenvector of a  $3 \times 3$  matrix needs to be found.
- In step 5, only a “cross” inside each of the matrices is updated (the elements with the one of the indices  $i$  or  $j$ ). This gives a total complexity (for naive implementation) of  $8Ln$  multiplications per update.



---

**Algorithm 8.1** Jacobi-type algorithm by rotating the tensors

---

**Input:** Matrices  $\mathbf{A}^{(\ell)}$ ,  $1 \leq \ell \leq L$ , starting point  $\mathbf{U}_0$ .

**Output:** Sequence of iterations  $\mathbf{U}_k$ , rotated matrices  $\mathbf{W}_k^{(\ell)}$ .

1. initialize  $\mathbf{W}_0^{(\ell)} = \mathbf{U}^H \mathbf{A}^{(\ell)} \mathbf{U}$ , for all  $\ell$ .
2. **For**  $k = 1, 2, \dots$  until a stopping criterion is satisfied do
3.   Choose an index pair  $(i_k, j_k)$
4.   Find  $\Psi_k = \Psi_k(c, s_1, s_2)$  that minimizes

$$h_k(\theta) = \sum_{\ell} \|\text{diag}\{(\mathbf{G}^{(i_k, j_k, \Psi_k)})^H \mathbf{W}_{k-1}^{(\ell)} \mathbf{G}^{(i_k, j_k, \Psi_k)}\}\|^2$$

5.   Update  $\mathbf{W}_k^{(\ell)} = (\mathbf{G}^{(i_k, j_k, \Psi_k)})^H \mathbf{W}_{k-1}^{(\ell)} \mathbf{G}^{(i_k, j_k, \Psi_k)}$
  6. **End For**
- 

Thus, if a cyclic strategy (2.12) is adopted (the whole gradient is not computed), then the cycle of  $\frac{n(n-1)}{2}$  plane rotations (often called *sweep*) has the complexity  $O(Ln^3)$ .

Algorithm 2.1 requires more care, since we need to have access to the Riemannian gradient (or the matrix  $\Lambda(\mathbf{U})$ ). According to Proposition 4.14,  $O(Ln^2)$  multiplications are needed to compute the matrix  $\Lambda(\mathbf{U})$  in the Riemannian gradient. On the other hand, a plane rotation affects also only a part of the Riemannian gradient (also a cross) hence updating the matrix  $\Lambda(\mathbf{U})$  after each rotation has complexity  $O(Ln)$ . Thus, the complexity of one sweep is again  $O(Ln^3)$ .

Now let us compare with the computational complexity of a first-order method from [3] (e.g., gradient descent). At each iteration, we need at least to compute the Riemannian gradient  $O(Ln^2)$ , and then compute the retraction, which has complexity  $O(n^3)$  for typical choices (QR or polar decomposition). Note that at each step we also need to rotate the matrices, which requires additional  $O(Ln^3)$  multiplications.

*Remark 8.1.* For 3rd order tensors, the complexity of the Jacobi-based methods does not increase, because we again update the cross, which has  $O(Ln)$  elements.

**8.2. Numerical experiments.** In the experiments, we again consider, for simplicity, simultaneous matrix diagonalization (2.1). The general setup is as follows: we generate  $L$  matrices  $\mathbf{A}^{(\ell)} \in \mathbb{C}^{n \times n}$ , and compare several versions of Algorithm 2.1, as well as first-order Riemannian optimization methods implemented in the `manopt` package [10] (using `stiefelcomplexfactory`). We compare the following methods:

1. *Jacobi-G-max*: at each step of Algorithm 2.1, we select the pair  $(i, j)$  that maximizes the absolute value  $\Lambda_{i,j}(\mathbf{U}_{k-1})$  (see Remark 4.3).
2. *Jacobi 0.1*: we select the pairs in a cyclic-by-row order (2.12), but perform the rotations only if (2.13) is satisfied for  $\delta = 0.1\sqrt{2}/n$ .
3. *Jacobi-cyclic*: we use the cyclic-by-row order (2.12), without (2.13).
4. *SD*: steepest descent from [10].
5. *CG*: conjugate gradients from [10].
6. *BFGS*: Riemannian version of BFGS from [10].

In all comparisons,  $\mathbf{U}_0 = \mathbf{I}_n$ . We also plot  $\sum_{\ell} \|\mathbf{A}^{(\ell)}\|^2 - f(\mathbf{U})$  instead of  $f(\mathbf{U})$ .

We first consider a difficult example.  $L = 5$  matrices of size  $10 \times 10$  were generated randomly, such that the real and imaginary part are sampled from the uniform distribution on  $[0; 1]$ . We plot the results in Figure 1.

We do not expect this example to be easy for all of methods: this example is

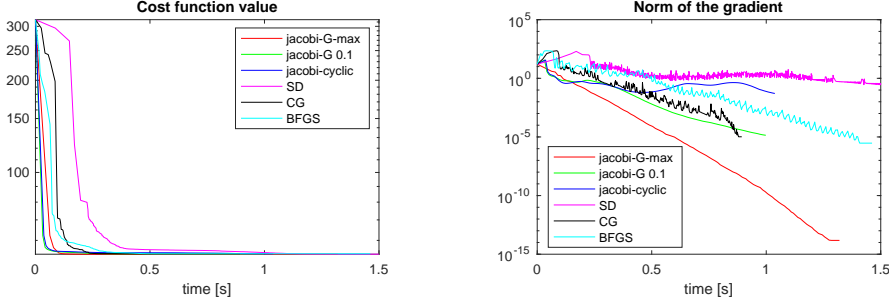


FIG. 1. Cost function value (left) and norm of the gradient (right), Example 1

far from a diagonalizable, and we are not likely to be in a small neighborhood of a local extremum. We see that the Jacobi-type methods converge very fast, and for the versions of [Algorithm 2.1](#) the gradient seems to converge to zero. We also see that the Jacobi-G-max version is the best compared to Jacobi-G with cyclic order and fixed  $\delta$  (we tried different values of  $\delta$ ).

We also consider a nearly diagonalizable case,  $n = L = 20$ . We take  $\mathbf{A}^{(\ell)} = \mathbf{Q}^H \mathbf{D}^{(\ell)} \mathbf{Q} + \mathbf{E}^{(\ell)}$ , where  $\mathbf{Q}$  is a random unitary matrix, elements of  $\mathbf{E}^{(\ell)}$  are i.i.d. realizations of Gaussian random variable with standard deviation  $10^{-6}$ , and  $\mathbf{D}^{(\ell)}$  is a diagonal matrix, whose diagonal elements are equal to 1, except the element  $D_{\ell,\ell}^{(\ell)} = 2$  (note that such matrices, without noise, satisfy [Proposition 7.6](#)).

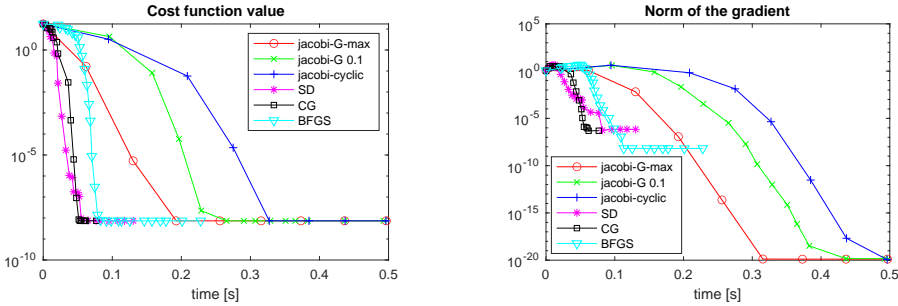


FIG. 2. Cost function value (left) and norm of the gradient (right), Example 2

We plot the results in [Figure 2](#). We see that the convergence of general-purpose Riemannian algorithms is much better in this case. Still, Jacobi algorithms converge in a few sweeps.

Note that in the current implementation (used to produce [Figure 1](#) and [Figure 2](#)), we do not use the  $O(Ln)$  update of  $\mathbf{\Lambda}(\mathbf{U})$  as suggested in [subsection 8.1](#) (i.e., the matrix  $\mathbf{\Lambda}(\mathbf{U})$  is recalculated at each step). This can be observed in [Figure 2](#), where each marker for the Jacobi-type methods corresponds to one sweep. Thus, a further speedup of Jacobi-type methods is possible.

**9. Discussion.** In this paper, we showed that for a class of optimization problems on the unitary group (corresponding to approximate matrix and tensor diagonalization), convergence of Jacobi-type algorithms to stationary points can be proved (together with convergence rates). A gradient-based order of Givens rotations is adopted (which extends the approach of [\[32\]](#) for the real-valued case). By using the

tools based on Łojasiewicz gradient inequality, we can ensure single-point convergence, under regularity conditions on one of the accumulation points; the speed of convergence is linear for the non-degenerate case, and local convergence can be proved. We also provided a characterization of Jacobi rotations for tensors of arbitrary orders.

Still, we believe that stronger results can be obtained. For the matrix case, although the Jacobi-type algorithms are similar in spirit to block-coordinate descent, they enjoy quadratic convergence (of the cost function value) for the classic matrix case [24] and the case of a pair of commuting matrices [13].

Also, in the matrix case, many results are available for cyclic strategies (at least weak convergence is known, see [24]). It would be interesting to see if similar results can be proved for tensor and joint matrix diagonalization cases; in fact, the convergence for the pure cyclic strategy is often observed in practice (see [38] for a comparison in the case of orthogonal group), but there is no convergence proof.

Note that we were not able to prove global single-point convergence, as in [38] (proved for 3rd order tensors or matrices). It seems that in the complex case, not only the order of rotations matters (which makes it similar to the higher-order case [38]). One possible track is to modify of a way to find the Jacobi rotation itself (i.e. adopt proximal-like steps if needed, see also [38]).

Another interesting question is whether we can relax the definition of single-point convergence. Indeed, if the critical point is degenerate (even in the quotient manifold), then a natural question is whether the potentially different accumulation points, belong the same critical manifold. This is, in fact, what is typically proved<sup>20</sup> for the matrix case [20]: if there are multiple eigenvalues, then the convergence of invariant subspaces is guaranteed (which corresponds to the same critical manifold).

### Appendix A. Multilinear algebra proofs.

*Proof of Lemma 4.11.* The “only if” part follows from the fact that if  $\mathcal{B}$  is Hermitian, there exist tensors  $\mathcal{A}^{(1)}, \dots, \mathcal{A}^{(L)}$  of order  $d$  and real numbers  $\alpha_\ell$ , such that

$$(A.1) \quad \mathcal{B} = \sum_{\ell=1}^L \alpha_\ell (\mathcal{A}^{(\ell)})^* \otimes \mathcal{A}^{(\ell)}, \text{ i.e. } \mathcal{B}_{i_1 \dots i_d j_1 \dots j_d} = \sum_{\ell=1}^L \alpha_\ell (\mathcal{A}_{i_1 \dots i_d}^{(\ell)})^* \mathcal{A}_{j_1 \dots j_d}^{(\ell)}.$$

This is nothing but the spectral theorem for Hermitian matrices applied to a matricization of  $\mathcal{B}$ , see also [33, Propositions 3.5 and 3.9]. Then (A.1) implies that

$$g_{\mathcal{B},d}(\mathbf{u}) = \sum_{\ell=1}^L \alpha_\ell |g_{\mathcal{A}^{(\ell)},0}(\mathbf{u})|^2.$$

In order to prove the “if” part, we make the following remarks.

- (a) When restricted to  $\|\mathbf{u}\| = 1$ , the order of the form can be always increased; Indeed, suppose that  $\mathcal{T}$  is  $2(d-1)$ -order Hermitian, then for all  $\|\mathbf{u}\| = 1$

$$g_{\mathcal{T},d-1}(\mathbf{u}) = (\mathcal{T} \otimes \mathbf{I}_n) \bullet_1 \mathbf{u}^* \cdots \bullet_{d-1} \mathbf{u}^* \bullet_d \mathbf{u} \cdots \bullet_{2(d-1)} \mathbf{u} \bullet_{2d-1} \mathbf{u}^* \bullet_{2d} \mathbf{u},$$

where  $\mathbf{I}_n$  is the identity; the expression on the right-hand side is a Hermitian form (4.13), where the  $2d$ -order tensor  $\mathcal{B}$  can be defined by permuting the indices:

$$\mathcal{B}_{i_1 \dots i_d j_1 \dots j_d} = (\mathcal{T} \otimes \mathbf{I}_n)_{i_1 \dots i_{d-1} j_1 \dots j_{d-1} i_d j_d} = \mathcal{T}_{i_1 \dots i_{d-1} j_1 \dots j_{d-1}} (\mathbf{I}_n)_{i_d j_d}.$$

<sup>20</sup>In fact, it seems that [38] is the first paper explicitly showing single-point convergence for the single real-valued matrix case, as the result of [38] also apply to the eigenvalue problems.

(b) Note that for any  $t$ , the function  $|g_{\mathcal{A},t}(\mathbf{u})|^2 = g_{\mathcal{A},t}(\mathbf{u})g_{\mathcal{A},t}^*(\mathbf{u})$  is also a  $2d$ -form:

$$(A.2) \quad |g_{\mathcal{A},t}(\mathbf{u})|^2 = (\mathcal{A} \otimes \mathcal{A}^*)_{\bullet_1} \mathbf{u}^* \cdots \bullet_t \mathbf{u}^* \bullet_{t+1} \mathbf{u} \cdots \bullet_{d+t} \mathbf{u} \bullet_{d+t+1} \mathbf{u}^* \cdots \bullet_{2d} \mathbf{u}^*,$$

which can be written<sup>21</sup> as  $g_{\mathcal{B},d}(\mathbf{u})$  for a tensor  $\mathcal{B}$  obtained by permuting indices:

$$\mathcal{B}_{i_1 \dots i_d j_1 \dots j_d} = (\mathcal{A} \otimes \mathcal{A}^*)_{i_1 \dots i_t j_{t+1} \dots j_d j_1 \dots j_t i_{t+1} \dots i_d} = \mathcal{A}_{i_1 \dots i_t j_{t+1} \dots j_d} \mathcal{A}_{j_1 \dots j_t i_{t+1} \dots i_d}^*.$$

Finally, sums of Hermitian tensors are Hermitian, which completes the proof.  $\square$

*Proof of Theorem 4.16.* Since the cost function has the form (4.11), we have

$$\tilde{h}(c, s_1, s_2) = g_{\mathcal{T},d}(\begin{bmatrix} c \\ s^* \end{bmatrix}) + g_{\mathcal{T},d}(\begin{bmatrix} -s \\ c \end{bmatrix})$$

Let us rewrite the first term using the double contraction:

$$(A.3) \quad \begin{aligned} g_{\mathcal{T},d}(\begin{bmatrix} c \\ s^* \end{bmatrix}) &= \mathcal{T}_{\bullet_1, d+1} \left( \begin{bmatrix} c \\ s^* \end{bmatrix}^* \begin{bmatrix} c & s^* \end{bmatrix} \right) \cdots \bullet_{d, 2d} \left( \begin{bmatrix} c \\ s^* \end{bmatrix}^* \begin{bmatrix} c & s^* \end{bmatrix} \right) \\ &= \frac{1}{2^d} \mathcal{T}_{\bullet_1, d+1} (\mathbf{I}_2 + \mathbf{R}) \bullet_{2, d+2} (\mathbf{I}_2 + \mathbf{R}) \cdots \bullet_{d, 2d} (\mathbf{I}_2 + \mathbf{R}), \end{aligned}$$

where  $\mathbf{R} \stackrel{\text{def}}{=} \begin{bmatrix} 2c^2 - 1 & 2cs^* \\ 2cs & 1 - 2c^2 \end{bmatrix}$ , so that  $\begin{bmatrix} c \\ s^* \end{bmatrix}^* \begin{bmatrix} c & s^* \end{bmatrix} = \begin{bmatrix} c^2 & cs^* \\ cs & |s|^2 \end{bmatrix} = \frac{1}{2}(\mathbf{I}_2 + \mathbf{R})$ .

Similarly, by noting that

$$\begin{bmatrix} -s \\ c \end{bmatrix}^* \begin{bmatrix} -s & c \end{bmatrix} = \begin{bmatrix} |s|^2 & -cs^* \\ -cs & c^2 \end{bmatrix} = \frac{1}{2}(\mathbf{I}_2 - \mathbf{R}),$$

we can rewrite the second term

$$(A.4) \quad g_{\mathcal{T},d}(\begin{bmatrix} -s \\ c \end{bmatrix}) = \frac{1}{2^d} \mathcal{T}_{\bullet_1, d+1} (\mathbf{I}_2 - \mathbf{R}) \bullet_{2, d+2} (\mathbf{I}_2 - \mathbf{R}) \cdots \bullet_{d, 2d} (\mathbf{I}_2 - \mathbf{R}).$$

When summing (A.3) and (A.4), we note that the odd powers of  $\mathbf{R}$  cancel, and the even powers have positive signs; therefore, due to symmetries we get

$$\tilde{h}(c, s_1, s_2) = \frac{1}{2^{d-1}} \sum_{j=0}^m \binom{d}{2j} \mathcal{T}_{\bullet_1, d+1} \mathbf{R} \cdots \bullet_{2j, d+2j} \mathbf{R} \bullet_{2j+1, d+2j+1} \mathbf{I}_2 \cdots \bullet_{d, 2d} \mathbf{I}_2,$$

where the binomial coefficient appears when we sum over all possible locations of  $\mathbf{R}$ .

Next, we remark that  $\mathbf{R}$  can be expressed in the following orthogonal basis

$$(A.5) \quad \mathbf{R} = r_1 \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} + r_2 \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} + r_3 \begin{bmatrix} 0 & i \\ -i & 0 \end{bmatrix},$$

where  $\mathbf{r} = [r_1 \ r_2 \ r_3]^\top = [2c^2 - 1 \ -2cs_1 \ -2cs_2]^\top$  is defined in (4.4). Then by the multilinearity of the contractions, we can rewrite the expression for  $\tilde{h}(c, s_1, s_2)$  as

$$\tilde{h}(c, s_1, s_2) = \sum_{j=0}^m \mathcal{F}^{(j)} \bullet_1 \mathbf{r} \cdots \bullet_{2j} \mathbf{r},$$

<sup>21</sup>An alternative shorter proof of part (b) follows from the fact that a  $2d$ -order form  $g_{\mathcal{B},d}(\mathbf{u})$  is real-valued if and only if it is Hermitian, see [33, Proposition 3.6]

where each  $\mathcal{F}^{(j)}$  is a symmetric complex  $2j$ -order  $3 \times \dots \times 3$  tensor, whose entries are obtained by contractions of  $\mathcal{T}$  with basis matrices in (A.5) or  $\mathbf{I}_2$ .

It is only left to show that all the elements in each of the tensors  $\mathcal{F}^{(j)}$  are real. This is indeed the case, because for a Hermitian tensor  $\mathcal{T}$  contraction with one of the basis matrices keeps it Hermitian:

$$\begin{aligned} (\mathcal{T} \bullet_{1,d+1} \begin{bmatrix} 0 & i \\ -i & 0 \end{bmatrix})_{i_2 \dots i_d, j_2 \dots j_d} &= -i(\mathcal{T}_{2i_2 \dots i_d, 1j_2 \dots j_d} - \mathcal{T}_{1i_2 \dots i_d, 2j_2 \dots j_d}) \\ &= i(\mathcal{T}_{2j_2 \dots j_d, 1i_2 \dots i_d}^* - \mathcal{T}_{1j_2 \dots j_d, 2i_2 \dots i_d}^*) = ((\mathcal{T} \bullet_{1,d+1} \begin{bmatrix} 0 & i \\ -i & 0 \end{bmatrix})_{j_2 \dots j_d, i_2 \dots i_d})^*, \end{aligned}$$

and similarly for contractions with  $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ ,  $\begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$  and  $\mathbf{I}_2$ .

Finally, we note that, since  $\|\mathbf{r}\| = 1$ , all the tensors  $\mathcal{F}^{(j)}$  can be combined in one tensor  $\mathcal{F}$  of order  $2m$ , as in the proof of Lemma 4.11 (see part (a) of the proof).  $\square$

**Acknowledgments.** The authors would like to acknowledge the two anonymous reviewers and the associate editor for their useful remarks that helped to improve the presentation of the results.

#### REFERENCES

- [1] T. E. ABRUDAN, J. ERIKSSON, AND V. KOIVUNEN, *Steepest descent algorithms for optimization under unitary matrix constraint*, IEEE Trans. on Signal Process., 56 (2008), pp. 1134–1147.
- [2] P. A. ABSIL, R. MAHONY, AND B. ANDREWS, *Convergence of the iterates of descent methods for analytic cost functions*, SIAM Journal on Optimization, 16 (2005), pp. 531–547.
- [3] P.-A. ABSIL, R. MAHONY, AND R. SEPULCHRE, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, 2008.
- [4] P. A. ABSIL, R. MAHONY, AND J. TRUMPF, *An extrinsic look at the Riemannian Hessian*, in Geometric Science of Information: First International Conference, GSI 2013, F. Nielsen and F. Barbaresco, eds., Paris, France, 2013, Springer; Berlin Heidelberg, pp. 361–368.
- [5] A. ANANDKUMAR, R. GE, D. HSU, S. M. KAKADE, AND M. TELGARSKY, *Tensor decompositions for learning latent variable models*, Journal of Machine Learning Research, 15 (2014), pp. 2773–2832.
- [6] H. ATTOUCH, J. BOLTE, AND B. F. SVAITER, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss–Seidel methods*, Mathematical Programming, 137 (2013), pp. 91–129.
- [7] A. BANYAGA AND D. E. HURTUBISE, *A proof of the Morse-Bott lemma*, Expositiones Mathematicae, 22 (2004), pp. 365 – 373.
- [8] R. BOTT, *Nondegenerate critical manifolds*, Annals of Mathematics, 60 (1954), pp. 248–261.
- [9] N. BOUMAL, P.-A. ABSIL, AND C. CARTIS, *Global rates of convergence for nonconvex optimization on manifolds*, IMA Journal of Numerical Analysis, 39 (2019), pp. 1–33.
- [10] N. BOUMAL, B. MISHRA, P.-A. ABSIL, AND R. SEPULCHRE, *Manopt, a Matlab toolbox for optimization on manifolds*, Journal of Machine Learning Research, 15 (2014), pp. 1455–1459, <http://www.manopt.org>.
- [11] S. BOYD AND L. VANDENBERGHE, *Convex optimization*, Cambridge University Press, 2004.
- [12] D. BRANDWOOD, *A complex gradient operator and its application in adaptive array theory*, IEE Proceedings H - Microwaves, Optics and Antennas, 130 (1983), pp. 11–16.
- [13] A. BUNSE-GERSTNER, R. BYERS, AND V. MEHRMANN, *Numerical methods for simultaneous diagonalization*, SIAM J. Matr. Anal. and Appl., 14 (1993), pp. 927–949.
- [14] J.-F. CARDOSO AND A. SOULOUMIAC, *Blind beamforming for non-gaussian signals*, IEE Proceedings F-Radar and Signal Processing, 140 (1993), pp. 362–370.
- [15] J.-F. CARDOSO AND A. SOULOUMIAC, *Jacobi angles for simultaneous diagonalization*, SIAM journal on matrix analysis and applications, 17 (1996), pp. 161–164.
- [16] P. COMON, *From source separation to blind equalization, contrast-based approaches*, in Int. Conf. on Image and Signal Processing (ICISP’01), Agadir, Morocco, May 2001, pp. 20–32. preprint: hal-01825729.
- [17] P. COMON, *Contrasts, independent component analysis, and blind deconvolution*, Int. J. Adapt. Control Sig. Proc., 18 (2004), pp. 225–243. preprint: hal-00542916.
- [18] P. COMON AND C. JUTTEN, *Handbook of Blind Source Separation: Independent component analysis and applications*, Academic press, 2010.

- [19] L. DE LATHAUWER, *Signal processing based on multilinear algebra*, Katholieke Universiteit Leuven Leuven, 1997.
- [20] Z. DRMAČ, *A global convergence proof for cyclic Jacobi methods with block rotations*, SIAM J. Matr. Anal. Appl., 31 (2010), pp. 1329–1350.
- [21] 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), pp. 303–353.
- [22] B. EMILE, P. COMON, AND J. LE ROUX, *Estimation of time delays with fewer sensors than sources*, IEEE Transactions on Signal Processing, 46 (1998), pp. 2012–2015.
- [23] P. M. N. FEEHAN, *Optimal Lojasiewicz-Simon inequalities and Morse-Bott Yang-Mills energy functions*, tech. report, 2018. arxiv:1706.09349.
- [24] G. GOLUB AND C. VAN LOAN, *Matrix Computations*, JHU Press, 3rd ed., 1996.
- [25] B. HALL, *Lie groups, Lie algebras, and representations: an elementary introduction*, vol. 222, Springer, 2015.
- [26] V. HARI AND E. B. KOVAČ, *Convergence of the cyclic and quasi-cyclic block Jacobi methods*, Electron. Trans. Numer. Anal, 46 (2017), pp. 107–147.
- [27] V. HARI AND E. B. KOVAČ, *On the convergence of complex Jacobi methods*, Linear and Multilinear Algebra, (2019), pp. 1–26.
- [28] S. HELGASON, *Differential Geometry, Lie Groups, and Symmetric Spaces*, Academic Press, 1978.
- [29] U. HELMKE AND J. B. MOORE, *Optimization and Dynamical Systems*, Springer, 1994.
- [30] A. HJØRUNGNES AND D. GESBERT, *Complex-valued matrix differentiation: Techniques and key results*, IEEE Transactions on Signal Processing, 55 (2007), pp. 2740–2746.
- [31] S. HU AND G. LI, *Convergence rate analysis for the higher order power method in best rank one approximations of tensors*, Numerische Mathematik, 140 (2018), pp. 993–1031.
- [32] M. ISHTEVA, P.-A. ABSIL, AND P. VAN DOOREN, *Jacobi algorithm for the best low multilinear rank approximation of symmetric tensors*, SIAM J. Matrix Anal. Appl., 2 (2013), pp. 651–672.
- [33] B. JIANG, Z. LI, AND S. ZHANG, *Characterizing real-valued multivariate complex polynomials and their symmetric tensor representations*, SIAM J. Matr. Anal. and Appl., 37 (2016), pp. 381–408.
- [34] M. KLEINSTEUBER, U. HELMKE, AND K. HUPER, *Jacobi’s algorithm on compact Lie algebras*, SIAM Journal on Matrix Analysis and Applications, 26 (2004), pp. 42–69.
- [35] S. KRANTZ AND H. PARKS, *A Primer of Real Analytic Functions*, Birkhäuser, Boston, 2002.
- [36] S. G. KRANTZ, *Function theory of several complex variables*, vol. 340, American Mathematical Soc., 2001.
- [37] C. LAGEMAN, *Convergence of gradient-like dynamical systems and optimization algorithms*, doctoralthesis, Universität Würzburg, 2007.
- [38] J. LI, K. USEVICH, AND P. COMON, *Globally convergent Jacobi-type algorithms for simultaneous orthogonal symmetric tensor diagonalization*, SIAM J. Matr. Anal. Appl., 39 (2018), pp. 1–22.
- [39] S. ŁOJASIEWICZ, *Une propri t  topologique des sous ensembles analytiques re ls*, in Colloques internationaux du C.N.R.S., 117. Les  quations aux D riv es Partielles, 1963, pp. 87–89.
- [40] W. F. MASCARENHAS, *On the convergence of the Jacobi method for arbitrary orderings*, SIAM Journal on Matrix Analysis and Applications, 16 (1995), pp. 1197–1209.
- [41] E. MASSART AND P.-A. ABSIL, *Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices*, SIAM J. on Matr. Anal. and Appl., 41 (2020), pp. 171–198.
- [42] P. MCCULLAGH, *Tensor Methods in Statistics*, Monographs on Statistics and Applied Probability, Chapman and Hall, 1987.
- [43] J. MILNOR, *Morse theory*, Princeton University Press, 1963.
- [44] J. NIE AND Z. YANG, *Hermitian tensor decompositions*, (2019), <https://arxiv.org/abs/1912.07175>.
- [45] B. T. POLYAK, *Gradient methods for minimizing functionals*, Zh. Vychisl. Mat. Mat. Fiz., 3 (1963), pp. 643–653.
- [46] T. RAPCSAK, *Geodesic convexity in nonlinear optimization*, Journal of Optimization Theory and Applications, 69 (1991).
- [47] R. SCHNEIDER AND A. USCHMAJEV, *Convergence results for projected line-search methods on varieties of low-rank matrices via lojasiewicz inequality*, SIAM J. Opt., 25 (2015), pp. 622–646.
- [48] A. USCHMAJEV, *A new convergence proof for the higher-order power method and generalizations*, Pac. J. Optim., 11 (2015), pp. 309–321.