4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), April 2003, Nara, Japan

AN ORTHOGONAL MATRIX OPTIMIZATION BY DUAL CAYLEY PARAMETRIZATION TECHNIQUE

Isao Yamada and Takato Ezaki

Dept. of Communications and Integrated Systems,
Tokyo Institute of Technology, Tokyo, 152-8552, Japan.
{isao, eza-o}@comm.ss.titech.ac.jp

ABSTRACT

This paper addresses a mathematically sound technique for the orthogonal matrix optimization problem that has broad applications in recent signal processing problems including the independent component analysis. We propose Dual Cayley parametrization technique that can decompose a slightly restricted version of the original problem into a pair of simple constraint-free optimization problems. This decomposition frees us from using local parametrizations strongly dependent on the location of orthogonal matrices, and hence from numerical approximations of delicate computations, e.g., matrix exponential mappings or SVD.

1. INTRODUCTION

The orthogonal matrix optimization problem is formulated as:

Problem 1.1 *For a given function* $\Theta : \mathbb{R}^{N\times N} \to \mathbb{R}$,

$$\text{Find a matrix } U^* \in \arg \inf_{U\in O(N)} \Theta(U), \tag{1}$$

where $O(N) := \{U \in \mathbb{R}^{N\times N} \mid U^T U = I\}$ *(i.e., the set of all orthogonal matrices) is a manifold called the orthogonal group* [1] *because it constitutes a group.*

An important example having rich applications, in the independent component analysis / blind source separation / unsupervised adaptive filtering (see for example [2–8]), is found in the joint orthogonal matrix diagonalization:

Problem 1.2 *For given* $\mathcal{A} := \{A_k \in \mathbb{R}^{N\times N} \mid k = 1, \dots, K\}$, *minimize*

$$f_{\mathcal{A}}(U) := \sum_{k=1}^{K} \mathrm{off}(U^T A_k U), \quad \forall U \in O(N),$$

where $\mathrm{off}(X) := \sum_{1\le i\ne j\le N} |x_{ij}|^2$, *and* x_{ij} *denotes the* (i,j)*-th entry of* $X \in \mathbb{R}^{N\times N}$.

Problem 1.1 (or more general problem, where U is possibly rectangular real/complex matrix) dates back to the seventies (see [9] and references therein) but it is only in the last decade when the studies, on the algorithmic solutions to Problem 1.1, start extensively [7, 10] (In [7], a standard framework with general local parametrization is stated for minimizing a function on a manifold). One of the central burdens of Problem 1.1 must be in tracking the manifold $O(N)$. Indeed, a great deal of effort has been devoted to resolve this tracking difficulty. For example, the algorithms in [10] perform a series of descent steps taken along the *geodesic* of the manifold. However moving along the *geodesic* requires some delicate numerical approximation [11, §11.3] of the matrix exponential function. For update $U_n \in O(N)$ to $U_{n+1} \in O(N)$ $(n = 0, 1, 2, \cdots)$, the algorithms in [7, 8] assign, to the vicinity of U_n as its local parametrization, the coordinate on the tangent space of $O(N)$ at U_n, through the projection operator onto $O(N)$, and then $U_{n+1} \in O(N)$ is computed in the descent directions of Θ. However numerical approximation is inherently unavoidable because the projection $\pi(X) := \arg\min_{U\in O(N)} \|X - U\|^2, \forall X \in \mathbb{R}^{N\times N}$ ($\|\cdot\|$ denotes the Frobenius norm of a matrix) is given in terms of the SVD (singular-value decomposition) of X, say $X = Y\Sigma Z^T$, by $\pi(X) = YZ^T$.

This paper proposes Dual Cayley parametrization technique that can decompose a slightly restricted version of Problem 1.1 into a pair of constraint-free optimization problems without using any numerical approximation. Because the proposed parametrization realizes inherently a pair of *global* parametrizations corresponding to two disjoint subsets of the orthogonal group, a variety of standard optimization algorithms can be applied without caring the orthogonality constraints. To demonstrate the potential applicability of the proposed technique, we present first for local optimization problems the steepest descent method and the Newton method in the Dual Cayley transform domains. An important side effect of the global nature of the proposed parametrization would be its affinity to global optimization problems. We discuss briefly a possible approach,based on the proposed parametrization, to the global optimization over the orthogonal group.

2. PRELIMINARIES

Throughout this paper, we use the following mathematical notations. Let $\mathbb{R}$ denote the set of all real numbers and $\mathbb{C}$ the set of all complex numbers. $I \in \mathbb{R}^{N\times N}$ denotes the identity matrix. The superscript ^{T} stands for the transposition. $\mathrm{tr}(X)$ denotes the trace of $X \in \mathbb{R}^{N\times N}$. The set of all skew-symmetric matrices is denoted by $Q(N) := \{V \in \mathbb{R}^{N\times N} \mid V = -V^T\}$, which can be identified, in a natural way, with the Euclidean space $\mathbb{R}^L$ ($L := \frac{N(N-1)}{2}$) because $V \in Q(N)$ is fully characterized with all its components in the strictly upper triangular locations.

Definition 2.1 *(Cayley transform* [1, IV.§6]*)*
Let $E := \{U \in O(N) \mid \det(U + I) = 0\}$*. The Cayley transform* $\Phi : O(N) \setminus E \to Q(N)$ *is defined by*

$$\Phi(U) := (I - U)(I + U)^{-1}, \forall U \in O(N) \setminus E. \qquad (2)$$

The mapping Φ is bijective and its inverse $\Phi^{-1} : Q(N) \to O(N) \setminus E$ is symmetrically given by

$$\Phi^{-1}(V) = (I - V)(I + V)^{-1}, \forall V \in Q(N).$$

Because of the eigenvalue distributions of matrices in $O(N) \setminus E$ and $Q(N)$ [i.e., all eigenvalues of every orthogonal matrix have unit magnitude, and all eigenvalues of every skew-symmetric matrix are pure imaginary], the mappings Φ and Φ^{-1} are well-defined (moreover, in principle, their exact computations are possible without requiring any numerical approximation) over $O(N) \setminus E$ and $Q(N)$ respectively. Φ serves as the homeomorphism between $O(N)\setminus E$ and $Q(N)$ [Note: Φ is a matrix version of the bilinear transform (Moebius transform) in the classical complex analysis [12] and the S-Z transform in signal processing [13]]. Although the definition of the Cayley transform is naturally extended to that for the unitary group [1], we restrict, for simplicity, the following discussion to the optimization over the orthogonal group.

3. RESOLUTION OF ORTHOGONAL CONSTRAINTS BY DUAL CAYLEY PARAMETRIZATION

We start from the following proposition that is a key of the proposed technique.

Proposition 3.1 *Let* $SO(N) := \{U \in O(N) \mid \det(U) = 1\}$*. Then we have:*

(a) $SO(N) = \overline{SO(N) \setminus E}$*, where* $\overline{S}$ *denotes the closure of* $S \subset \mathbb{R}^{N\times N}$ *in the sense of the Frobenius norm* $\|\cdot\|$*.*

(b) For any $\mathcal{T} \in O(N)$ *satisfying* $\det(\mathcal{T}) = -1$*,* $\mathcal{T}(SO(N)) := \{\mathcal{T}U \mid U \in SO(N)\} = \overline{\mathcal{T}\,(SO(N) \setminus E)}$*.*

(c) $O(N) = \overline{(O(N) \setminus E)} \cup \overline{\mathcal{T}\,(O(N) \setminus E)}$*.*

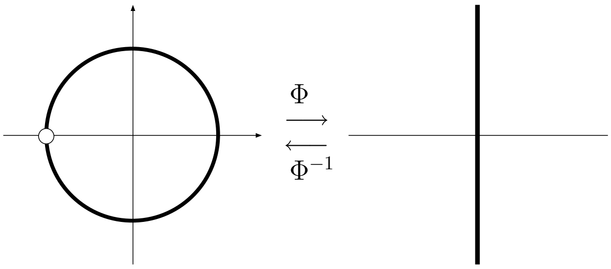

Fig. 1. Cayley transform resolves the orthogonality constraint

Proof: Remind that any matrix $U \in O(N)$ can be expressed as $U = \mathcal{T}_1^T \Lambda \mathcal{T}_1$ with some $\mathcal{T}_1 \in O(N)$ and

$$\Lambda = \begin{bmatrix} I_{n_1} & 0 & 0 & \cdots & 0 \\ 0 & -I_{n_2} & 0 & \cdots & 0 \\ 0 & 0 & \mathcal{R}(\omega_1) & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & \mathcal{R}(\omega_m) \end{bmatrix} \qquad (3)$$

(see for example [1, IV.§5]), where $I_{n_1} \in \mathbb{R}^{n_1\times n_1}$ and $I_{n_2} \in \mathbb{R}^{n_2\times n_2}$ are the identity matrices, $\mathcal{R}(\omega) = \begin{bmatrix} \cos(\omega) & -\sin(\omega) \\ \sin(\omega) & \cos(\omega) \end{bmatrix} \in \mathbb{R}^{2\times 2}$. In particular, the number of -1 in the diagonal entries in Λ is even when $U \in SO(N)$. This implies that every $U \in SO(N)$ can be approximated in any accuracy by $\widetilde{U}(\varepsilon) := \mathcal{T}_1^T \widetilde{\Lambda}(\varepsilon) \mathcal{T}_1 \in SO(N) \setminus E$, where $\widetilde{\Lambda}(\varepsilon)$ is given by replacing each submatrix $-I_2$, in $-I_{n_2}$ [in (3)], by $\mathcal{R}(\pi+\varepsilon)$ satisfying $\det\left(\mathcal{R}(\pi + \varepsilon) + I\right) \neq 0$. This proves (a). (b) is obvious from (a). Application of (a) and (b) to the fact: $St(N, N) = SO(N) \cup \mathcal{T}(SO(N))$ (see [1, IV.§6]), immediately leads to (c). (Q.E.D.)
Note: Both $SO(N)$ and $\mathcal{T}(SO(N))$ are connected subsets of $O(N)$ while $O(N)$ itself is not a connected set. This fact implies that optimization problem over $O(N)$ should be splitted into a pair of problems defined respectively over $SO(N)$ and $\mathcal{T}(SO(N))$. The expression Λ in (3) shows that $O(N) \setminus E = SO(N) \setminus E$.

Proposition 3.1 demonstrates that it is almost always sufficient to deal with the following slightly restricted version of Problem 1.1 (See Remark 3.5).

Problem 3.2 *For a given function* $\Theta : \mathbb{R}^{N\times N} \to \mathbb{R}$*, find a matrix*

$$U^* \in \arg \inf_{U \in (O(N)\setminus E) \cup \mathcal{T}(O(N)\setminus E)} \Theta(U).$$

In the sequel, we focus on Problem 3.2. Thanks to the homeomorphism Φ, between $O(N) \setminus E$ and $Q(N)$ (see Fig. 1), we propose:

Definition 3.3 *(Dual Cayley parametrization)*

$$\left\{ \begin{array}{lll} \Phi^{-1}: & \mathbb{R}^L(\cong Q(N)) \longrightarrow & O(N) \setminus E \\ \mathcal{T}\Phi^{-1}: & \mathbb{R}^L(\cong Q(N)) \longrightarrow & \mathcal{T}(O(N) \setminus E). \end{array} \right.$$

Now, Problem 3.2 is equivalently translated as:

Problem 3.4 *Let*

$$\left. \begin{array}{l} \widehat{\Theta}_1(V) := \Theta\left(\Phi^{-1}(V)\right), \\ \widehat{\Theta}_2(V) := \Theta\left(\mathcal{T}\Phi^{-1}(V)\right) \end{array} \right\} \forall V \in Q(N).$$

Then find matrices

$$U^{*(1)} \in \Phi^{-1}\left(\arg\inf_{V\in Q(N)} \widehat{\Theta}_1(V)\right),$$

and

$$U^{*(2)} \in \mathcal{T}\left(\Phi^{-1}\left(\arg\inf_{V\in Q(N)} \widehat{\Theta}_2(V)\right)\right).$$

Remark 3.5 *The set $SO(N) \cap E$ is only a small portion of $SO(N)$ due to the extra condition $\det(I+U)=0$. However, we should remark here that the transform could be numerically unstable in the vicinity of E. This is observed simply for example by $\mathrm{tr}(V^TV) = \sum_{i=1}^{N} \frac{1-\cos\omega_i}{1+\cos\omega_i} \to \infty$ (as $\det(I+U) = \prod_{i=1}^{N}(1+e^{j\omega_i}) \to 0$), where $e^{j\omega_i}$ denotes the i-th eigenvalue of $U = \Phi^{-1}(V)$. Moreover in the vicinity of E, we deduce in a similar way to* [11, §2.7.2]*:*

$$\tfrac{\|V(\epsilon)-V\|_2}{\|V\|_2} \leq 2\left\|(I+U)^{-1}\right\|_2 \|\epsilon F\|_2 + \mathrm{Order}(\epsilon^2),$$

where $V=\Phi(U)$, $(I+U+\epsilon F)V(\epsilon) = I-U-\epsilon F$, $\forall F \in \mathbb{R}^{N\times N}$, and $\epsilon > 0$ is assumed sufficiently small ($\|\cdot\|_2$ denotes 2-norm). Because $\left\|(I+U)^{-1}\right\|_2 = \max_i \sqrt{\frac{1}{2+2\cos\omega_i}}$ must be large, a small variation $\|\epsilon F\|_2$ can influence significantly the error ratio of V. These facts suggest that the proposed parametrization (Def. 3.3) is hardly applied to an optimization problem around E.

4. OPTIMIZATION ALGORITHMS VIA DUAL CAYLEY PARAMETRIZATION

Since Problem 3.4 is reduced to minimizing $\widehat{\Theta}_t$ $(t=1,2)$ over $\mathbb{R}^L$ $(\cong Q(N))$, we can apply a variety of optimization techniques (see for example [14–17]) developed essentially for the constraint-free (or simply constrained) optimization problems. Fig. 2 illustrates the main idea of the proposed optimization technique over $O(N) \setminus E$, where we assume that the optimal solution in $O(N)$ locates sufficiently far from E. The starting point $U_0^{(1)}$ must be chosen from $O(N) \setminus E$. Then numerical optimization of $\widehat{\Theta}_1$ over $\mathbb{R}^L$ $(\cong Q(N))$ can be proceeded without caring the orthogonality constraint. The final transform is well-defined because of the eigenvalue distribution of $V^{*(1)} \in Q(N)$. [Of course, optimization over $\mathcal{T}(O(N)\setminus E)$ can be realized in a similar way].

4.1. Local optimization

To demonstrate the potential applicability of the proposed parametrization technique to the local optimization over $O(N)$, we present the steepest descent method and the Newton method for Problem 3.4 (hence for Problem 3.2).

Algorithm 1 *(Steepest descent method for Problem 3.4) Choose $U_0^{(t)} \in O(N)\setminus E$ $(t=1,2)$ and let $V_0^{(1)} := \Phi(U_0^{(1)}) \in Q(N)(\cong \mathbb{R}^L)$ and $V_0^{(2)} := \Phi(\mathcal{T}^{-1}U_0^{(2)}) \in Q(N)(\cong \mathbb{R}^L)$. Then generate a pair of sequences $V_n^{(t)}$ $(n=0,1,\ldots)$ by*

$$V_{n+1}^{(t)} := V_n^{(t)} - \gamma_n^{(t)}\nabla\widehat{\Theta}_t(V_n^{(t)}),\ \ \gamma_n^{(t)} \in [0,\infty) \quad (t=1,2).$$

In the algorithm, we need $\nabla\widehat{\Theta}_t(V) \in \mathbb{R}^L$, which fortunately can be computed by applying elementary chain rule to $\nabla\Theta$. A variety of selection of step size $\gamma_n^{(t)}$ is known [14]. For example, we often use a sufficiently small constant $\gamma^{(t)} =: \gamma_n^{(t)}$ or Armijo's rule. For the convergence of the steepest descent method with Armijo's rule, see for example [14, Theorem 8.6.3]. (Note: Specially for Problem 1.2, a steepest descent type algorithm was proposed based on a Cayley parametrization [18], which unfortunately can cover only a small local subspace of $Q(N)$.)

Algorithm 2 *(Newton method for Problem 3.4) Choose $U_0^{(t)} \in O(N)\setminus E$ $(t=1,2)$ and let $V_0^{(1)} := \Phi(U_0^{(1)}) \in Q(N)(\cong \mathbb{R}^L)$ and $V_0^{(2)} := \Phi(\mathcal{T}^{-1}U_0^{(2)}) \in Q(N)(\cong \mathbb{R}^L)$. Then generate a pair of sequences $V_n^{(t)}$ $(n=0,1,\ldots)$ by*

$$V_{n+1}^{(t)} := V_n^{(t)} - H_{\widehat{\Theta}_t}(V_n^{(t)})^{-1}\nabla\widehat{\Theta}_i(V_n^{(t)}) \quad (t=1,2),$$

where $H_{\widehat{\Theta}_t}(V) \in \mathbb{R}^{L\times L}$ is the Hessian matrix of $\widehat{\Theta}_t : \mathbb{R}^L \to \mathbb{R}$.

For each $(v_1, v_2, \ldots, v_L) \in \mathbb{R}^L(\cong Q(N))$, $H_{\widehat{\Theta}_t}(V)$ can be computed by using the elementary chain rule:

$$\frac{\partial^2\widehat{\Theta}_t}{\partial v_i \partial v_j} = \sum_{k,l\in\Omega} \frac{\partial^2\Theta}{\partial u_k^{(t)}\partial u_l^{(t)}} \frac{\partial u_k^{(t)}}{\partial v_i} \frac{\partial u_l^{(t)}}{\partial v_j} + \sum_{l\in\Omega} \frac{\partial\Theta}{\partial u_l^{(t)}} \frac{\partial^2 u_l^{(t)}}{\partial v_i \partial v_j},$$

where $u_k^{(1)}$ and $u_k^{(2)}$ $(k \in \Omega := \{1,2,\ldots,N\}\times\{1,2,\ldots,N\})$ denote respectively the components of $U^{(1)} := \Phi^{-1}(V) \in O(N)\setminus E$ and $U^{(2)} := \mathcal{T}\Phi^{-1}(V) \in O(N)\setminus E$, $\forall V \in Q(N)(\cong \mathbb{R}^L)$. For the standard result on the convergence of the Newton method, see for example [14, Theorem 8.6.5].

4.2. Global optimization

Algorithm 1 (the steepest descent method) and Algorithm 2 (the Newton method) (In Sec. 4.1) can approximate only a locally optimal solution (see for example [14, Theorems

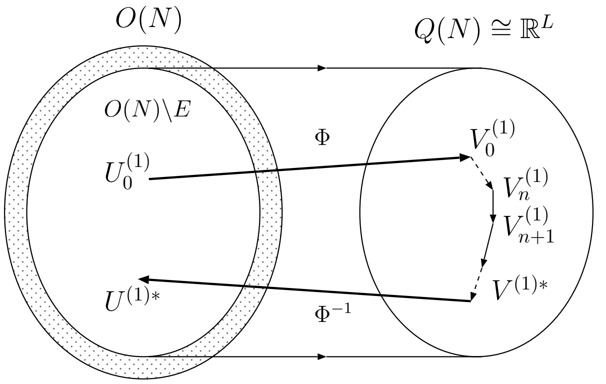

Fig. 2. Basic Idea of the Proposed Optimization over $O(N)\backslash E$

8.6.3 and 8.6.5]). The proposed parametrization technique would also play a central role to deal with more challenging cases where a global ε-optimal solution over $O(N)$ is required. Two major approaches, deterministic methods / stochastic methods (including simulated annealing), can be applied to Problem 3.4 [15, 16]. In this section, we briefly introduce a simple strategy in a deterministic method. Suppose that $\widehat{\Theta}_t : \mathbb{R}^L (\cong Q(N)) \to \mathbb{R}$ in Problem 3.4 is κ-lipschitzian over a simple closed convex set $C \subset \mathbb{R}^L$ [17, 19, 20]. In such a case, we can utilize for example a variety of algorithms developed for *Covering method* [20, Chap.2]: a class of global optimization techniques. The basic idea of covering method is to detect subregions B_i's ($\subset C$) not expected to contain the global minimum and exclude those subregions from further consideration. A primitive construction of such B_i's is found in the *Piyavskii's algorithm* (see for example [17] and [20, p.27]). After obtaining $\left\{\left(V_i, \widehat{\Theta}_t(V_i)\right)\right\}_{i=1}^{K}$, his algorithm constructs hyperspheres:

$$B_i := \left\{ V \in \mathbb{R}^L \mid \widehat{\Theta}_t(V_i) - \kappa \|V - V_i\| \geq \min_{1 \leq i \leq K} \widehat{\Theta}_t(V_i) \right\}$$

($k = 1, \ldots, K$). Since $\widehat{\Theta}_t(V)$ has no smaller value than $\min\limits_{1 \leq i \leq K} \widehat{\Theta}_t(V_i)$ in $\Upsilon_K := \bigcup_{i=1}^{K} B_i$, it is natural to select a next point V_{K+1} from $C \setminus \Upsilon_K$. In order to give a point V_{K+1} from $C \setminus \Upsilon_K$, many techniques have been developed [17, 19–23]. The covering methods usually lead to numerical algorithms that can compute a point $\widetilde{V}$ satisfying $\widehat{\Theta}_t(\widetilde{V}) \leq \inf_{V \in C} \widehat{\Theta}_t(V) + \varepsilon$. Thanks to the proposed parametrization technique, a variety of algorithms developed for the covering methods can be applied naturally to Problem 3.4. Further study of Problem 3.4 in this direction will be reported elsewhere.

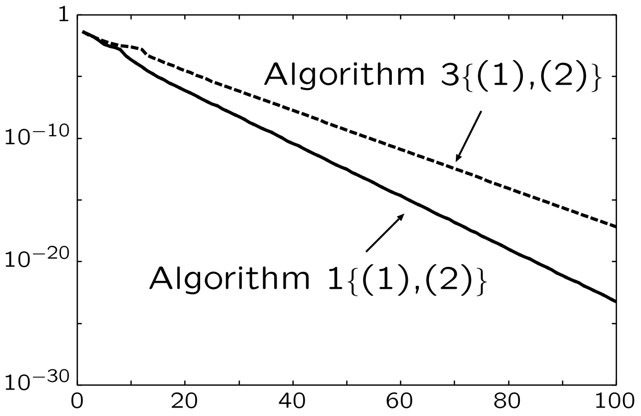

Fig. 3. Graph comparing the performances of the proposed method (Algorithm 1) and an algorithm in [8] (called Algorithm 3 in the present paper) with Armijo's rule for the joint orthogonal matrix diagonalization of 10 randomly generated 9-by-9 matrices with eigenvalues uniformly distributed over $[-0.05, 0.05]$.

5. NUMERICAL EXAMPLES

We present numerical comparisons of the proposed algorithms (Algorithms 1 and 2) with algorithms in [8] (called Algorithms 3 and 4 in the present paper). Algorithm 3 is a steepest descent type method that updates $U_n \in O(N)$ to $U_{n+1} \in O(N)$ in the steepest descent direction of Θ along $O(N)$ in terms of the local parametrization realized by projection onto $O(N)$ (see Section 1). Similarly, Algorithm 4 is a Newton type method that updates $U_n \in O(N)$ to $U_{n+1} \in O(N)$ in the Newton's descent direction of Θ along $O(N)$ in terms of the local parametrization realized by projection onto $O(N)$ (see Section 1). To design the step size γ_n for Algorithms 1 and 3, we used Armijo's rule in Figs. 3 and 5, and constant step sizes $\gamma_n = 5.0$ in Fig. 4 and $\gamma_n = 0.01$ in Fig. 6 respectively.

First, we applied Algorithms 1-4 to Problem 1.2 (i.e., $\Theta := f_{\mathcal{A}}$), where $N = 9$ and $K = 10$. Each target matrix A_k in Problem 1.2 was constructed by multiplying a random orthogonal matrix and its inverse from the left and right hand-sides of a random diagonal matrix whose diagonal entries are distributing uniformly in the range $[-0.05, 0.05]$. For Algorithms 1 and 2, $\mathcal{T} := -I_9$ was used to define $\widehat{\Theta}_2$. We also used as their common starting matrices $V_0^{(t)} \in Q(N)$ ($t = 1, 2$), the matrix $V_0 \in Q(N)$ of which every component in the strictly upper triangle locations is $1/N$.

In Figs. 3 and 4, Algorithm 1(t), depicts the behavior of $\widehat{\Theta}_t(V_n^{(t)})$, $n = 1, \ldots, 100$, ($t = 1, 2$) generated by Algorithm 1. Similarly, Algorithm 3(t) depicts the behavior of $\Theta(U_n^{(t)})$, $n = 1, \ldots, 100$, ($t = 1, 2$) generated by Algorithm 3 with the starting matrices $U_0^{(1)} := \Phi^{-1}(V_0) \in SO(N)$ and $U_0^{(2)} := \mathcal{T}\Phi^{-1}(V_0) \in \mathcal{T}(SO(N))$. The per-

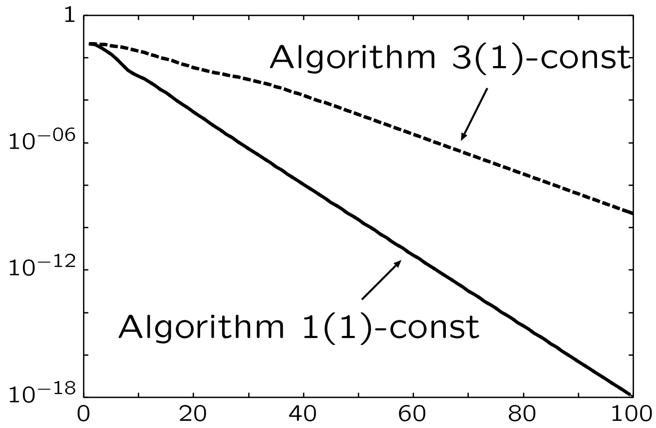

Fig. 4. Graph comparing the performances of the proposed method (Algorithm 1) and an algorithm in [8] (called Algorithm 3 in the present paper) with common step size $\gamma_n = 5.0$ for the same joint orthogonal matrix diagonalization as in Fig. 3.

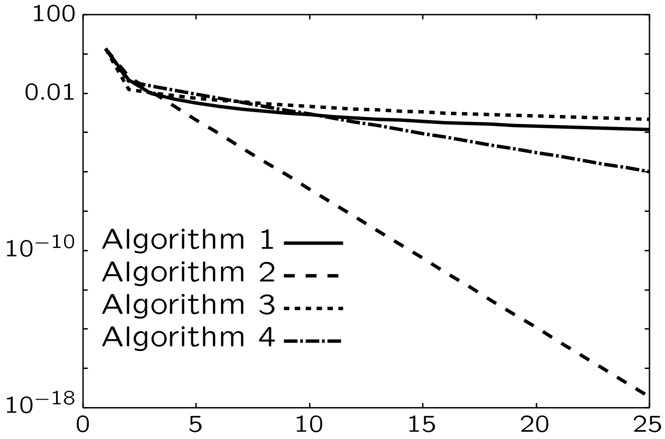

Fig. 5. Graph comparing the performances of the proposed methods (Algorithm 1 with Armijo's rule and Algorithm 2) and algorithms in [8] (called Algorithm 3 with Armijo's rule and Algorithm 4 in the present paper) for minimizing $\Theta(U) := (\mathrm{tr}\{U^T V\})^2$, where $V = \Phi(U) \in Q(N)$ and $N = 10$.

formances of Algorithms i(1) and i(2) ($i = 1, 3$) have no difference because $\Theta(U) = \Theta(\mathcal{T}U)$, $\forall U \in O(N)$ holds in this problem. Figs. 3 and 4 show that Algorithm 1 has better performance than Algorithm 3 for this joint diagonalization problem.

Next, in order to show more clearly the effects made by the proposed parametrization, we applied Algorithms 1-4 to the minimization of $\Theta(U) := (\mathrm{tr}\{U^T V\})^2$, where $V = \Phi(U) \in Q(N)$ and $N = 10$. In this case, by the commutativity of matrices under the trace operator, Θ has the same shape in the orthogonal group $O(N) \setminus E$ and in the Cayley transform domain $Q(N)$, hence we can observe the effects made by the proposed parametrization. In Figs. 5 and 6, we present the behaviors of $\widehat{\Theta}_1(V_n^{(1)})$, $n = 1, \ldots, 25$, generated by applying Algorithms 1 and 2 to $V_0^{(1)} \in Q(N)$ of which every component in the strictly upper triangle locations is $1/N$. We also present there the behaviors of $\Theta(U_n^{(1)})$, $n = 1, \ldots, 25$, generated by applying Algorithms 3 and 4 to $U_0^{(1)} := \Phi^{-1}(V_0^{(1)}) \in SO(N)$. These results demonstrate clearly that the proposed methods (Algorithm 1 and 2) exhibit much better performances than Algorithms 3 and 4.

Finally, we remark that the performances of the proposed algorithms (Algorithms 1 and 2) severely decline in an exceptional case where an optimal solution locates around E. This fact is demonstrated by the vast plateau generated through the Cayley parameterization (See Remark 3.5), and is observed in Fig. 7, where $\widehat{\Theta}(V_n)$ [Algorithm 1] and $\Theta(U_n)$ [Algorithm 3] are depicted for Problem 1.2 with $\begin{bmatrix} -I_2 & O \\ O & I_3 \end{bmatrix} \in E$ as one of its solutions, and both algorithms start from matrix $\begin{bmatrix} \mathcal{R} & O \\ O & I_3 \end{bmatrix}$, where $\mathcal{R} = \begin{bmatrix} \cos(\pi+0.7) & -\sin(\pi+0.7) \\ \sin(\pi+0.7) & \cos(\pi+0.7) \end{bmatrix}$.

ACKNOWLEDGEMENT

The authors thank Prof. K. Sakaniwa and Mr. H. Fujita of Tokyo Institute of Technology, Dr. G. Hori and Dr. T. Tanaka of Brain Science Institute, RIKEN, Prof. N. Murata of Waseda Univ., and Prof. J. Chao of Chuo Univ. for helpful discussions on the preliminary drafts of the present paper.

6. REFERENCES

[1] I. Satake, *Linear Algebra*. NY: Marcel Dekker, Inc., 1975.

[2] J.-F. Cardoso, "Source separation using higher order moments," in *Proceedings of ICASSP*, 1989, pp. 2109–2112.

[3] P. Comon, "Independent component analysis," *Proc. Int. Workshop on Higher-Order Stat.*, pp. 111–120, 1991.

[4] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non gaussian signals," *IEE Proceedings-F*, vol. 140, pp. 362–370, Dec 1993.

[5] P. Comon and P. Chevalier, "Blind source separation: Models, concepts, algorithms and performance," in *Unsupervised adaptive filtering volume I: Blind source separation*, S. Haykin, Ed. John Wiley & Sons, 2000.

[6] N. Murata and S. Ikeda, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, pp. 1–24, October 2001.

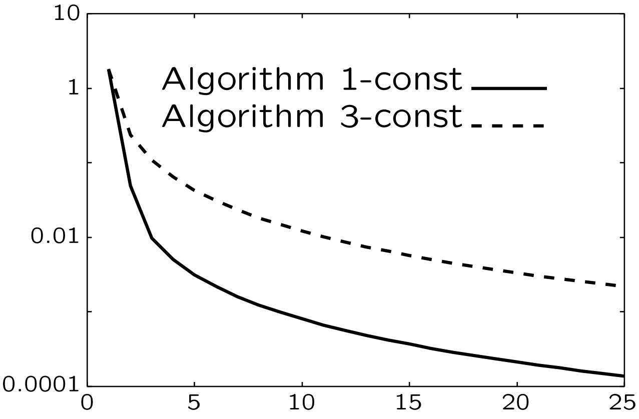

Fig. 6. Graph comparing the performances of the proposed method (Algorithm 1) and an algorithm in [8] (called Algorithms 3 and) with common step size $\gamma_n = 0.01$ for the same problem as in Fig. 5

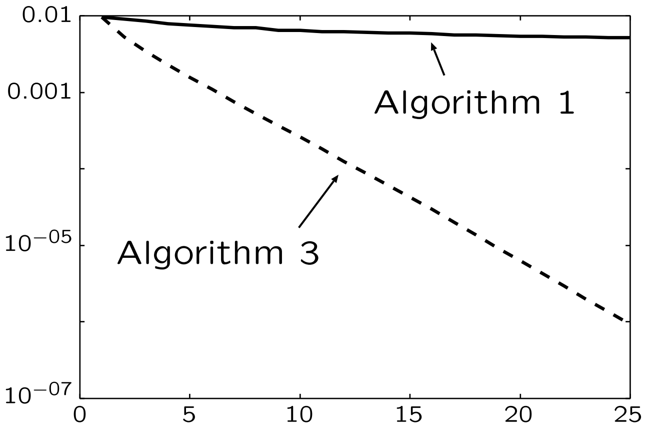

Fig. 7. Graph comparing the performances of the proposed method (Algorithm 1) and an algorithm in [8] (called Algorithms 3 and) with Armijo's rule for the joint diagonalizatioin in the case where an optimal solution locates in E.

[7] J. H. Manton, "Optimization algorithms exploiting unitary constraints," *IEEE Trans. Signal Processing*, vol. 50, no. 3, pp. 635–650, March 2002.

[8] M. Nikpour, J. H. Manton, and G. Hori, "Algorithms on Stiefel manifold for joint diagonalisation," in *Proceedings of ICASSP 2002*, vol. 2, 2002, pp. 1481–1484.

[9] C. Udriste, *Convex Functions and Optimization Methods on Riemannian Manifold*, ser. Mathematics and Its Applications. Kluwer Academic Publishers, 1994, vol. 297.

[10] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, pp. 303–353, 1998.

[11] G. H. Golub and C. F.Van Loan, *Matrix Computations*, 3rd ed. The Johns Hopkins University Press, 1996.

[12] P. Henrici, *Applied and Computational Complex Analysis, Volume 1*. John Wiley and Sons, Inc., 1974.

[13] N. Bose, *Digital Filters*. Krieger Publishing Company, 1993.

[14] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming theory and algorithms*, 2nd ed. John Wiley and Sons, Inc., 1993.

[15] R. Horst and P. M. Pardalos, *Handbook of Global Optimization*, ser. Nonconvex Optimization and its Applications. Dordrecht: Kluwer, Nov. 1994, vol. 2.

[16] P. M. Pardalos and H. E. Romeijn, *Handbook of Global Optimization Vol. 2*, ser. Nonconvex Optimization and its Applications. Dordrecht: Kluwer, June 2002, vol. 62.

[17] R. Horst and H. Tuy, *Global Optimization*. Springer-Verlag, 1987.

[18] M. Klajman and J. A. Chambers, "A novel approximate joint diagonalization algorithm," in *Mathematics in Signal Processing V*, J. G. McWhirter and I. K. Proudler, Eds. Oxford University Press, 2002, pp. 69–85.

[19] Y. Evtushenko, *Numerical Optimization Techniques*. Optimization Software Inc., Publications Division, New York, 1985.

[20] A. Törn and A. Žilinskas, *Global Optimization*, ser. Lecture Notes in Computer Science. Springer-Verlag, 1988, vol. 350.

[21] G. Wood, "The bisection method in higher dimensions," *Mathematical Programming*, vol. 55, pp. 313–337, 1992.

[22] I. Yamada, T. Miyamura, and K. Sakaniwa, "Restricted learning algorithms and improved excluding hyperspheres (in Japanese)," *IEICE Trans. D-II*, no. 9, pp. 1859–1881, 1994.

[23] I. Yamada and K. Sakaniwa, "A global optimization algorithm based on excluding hyperspheres," in *1995 European Conference on Circuit Theory and Design*, 1995.