Riemannian Optimization Method on the Generalized Flag Manifold for Complex and Subspace ICA

Yasunori Nishimori*, Shotaro Akaho ,* and Mark D. Plumbley†

*National Institute of Advanced Industrial Science and Technology (AIST),
AIST Central 2, 1-1-1, Umezono, Tsukuba, Ibaraki 305-8568, Japan
†Dep. of Electronic Engineering, Queen Mary University of London,
Mile End Road, London E1 4NS, UK

Abstract. In this paper we introduce a new class of manifold, the *generalized flag manifold*, which is the manifold of orthogonal *subspaces* and therefore includes the Stiefel and the Grassmann manifolds as special cases. We extend the Riemannian optimization method to include this new manifold by deriving the formulas for the natural gradient and geodesics on the manifold. We show how the complex and subspace ICA problems are solved by optimization of cost functions on the generalized flag manifold. Computer simulations demonstrate our algorithm gives good performance compared with the ordinary gradient descent method.

Key Words: Subspace ICA, Complex ICA, Natural gradient, Geodesics, Generalized flag manifold, Riemannian optimization.

INTRODUCTION

Many neural networks and signal processing tasks including independent component analysis (ICA) can be solved by optimization on a special class of manifolds related to the special orthogonal group $SO(n)$, such as the Stiefel and the Grassmann manifolds. Because a manifold can not be covered by one Euclidean space in general, and curved, optimization on manifolds is harder to solve than optimization on Euclidean space; the ordinary Euclidean optimization methods should be properly modified to be applied to such problems. First, the ordinary Euclidean gradient depends on a parametrization of a manifold, however, there are many ways to parametrize a manifold, thus, we should seek a *geometric* direction the cost function decreases most rapidly. This is made possible by introducing a Riemannian metric to the manifold, and the steepest direction with respect to a metric is called the Riemannian gradient vector (also known as the natural gradient in the neural networks community); it is independent of a parametrization, and its effectiveness in various machine learning problems has been demonstrated [2]. Second, because a manifold is 'curved', as soon as you 'add' an vector (e.g. the natural gradient) to the current point as in the Euclidean case, the updated point doesn't stay on the manifold anymore. To overcome this, we should generalize the concept of a straight line to manifolds; it is also determined by a Riemannian metric, and called a geodesic. The Riemannian optimization method was introduced by putting these ideas together

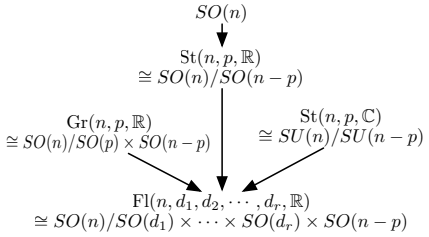


FIGURE 1. Hierarchy of manifolds

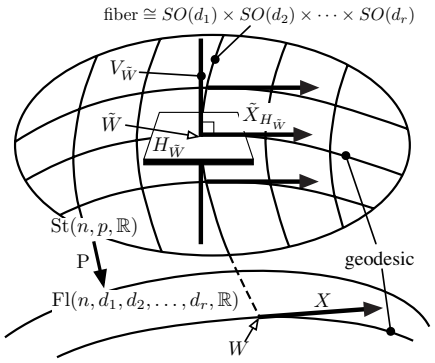


FIGURE 2. Riemannian submersion

[3]. It is formulated as follows: First, an appropriate Riemannian metric g is introduced into a manifold M; next, the ordinary updating direction used for optimization such as the gradient ∇f is switched to the Riemannian counterpart $V_k = \mathrm{grad}_{W_k} f(W_k)$; last, the current point W_k is updated to the next point W_{k+1}in direction to $-V_k$ along the geodesic: $W_{k+1} = \varphi_M(W_k, -\mathrm{grad}_W f(W_k), \eta)$, where $\gamma(t) = \varphi_M(W, V, t)$ denotes the equation of the geodesic on a manifold M starting from $W \in M$(i.e. $\gamma(0) = W$) in direction $V \in T_W M$(i.e. $\dot{\gamma}(0) = V$) with respect to to a Riemannian metric g on M. We call this the Riemannian optimization method. The use of such Riemannian geometrical techniques for optimization on manifolds has been explored by recent authors [1, 3, 4, 5, 9, 10, 11]. However, the target manifolds have been mainly the real Stiefel and Grassmann manifolds. The aim of the present paper is to introduce a new manifold, the generalized flag manifold, which generalizes both the Stiefel and Grassmann manifolds; describe the relationships between this new manifold and the previous manifolds; and extend the Riemannian optimization method to the generalized flag manifold using our previous geodesic formula for the Stiefel manifold [9]. The generalized flag manifold naturally arises when we consider *dependent* component analysis-type problems such as subspace ICA, moreover we show this manifold is a proper manifold to tackle the complex ICA problem as well. Simulations of these problems are carried out to compare the Riemannian optimization method with the ordinary gradient method.

GENERALIZED FLAG MANIFOLD

We summarize the relationships between manifolds (Fig. 1) which have been recently investigated in neural networks, signal processing, numerical analysis, and scientific computing [1, 3, 7, 9]. The most fundamental is the special orthogonal group $SO(n)$, which is the Lie group of the orthogonal matrices $\left\{\tilde{W} \in \mathbb{R}^{n\times n} \middle| \tilde{W}^\top \tilde{W} = I_n, \det(\tilde{W}) = 1\right\}$. The ordinary ICA problem can be solved by minimizing a cost function over $SO(n)$. Other manifolds discussed in this paper are all *descendants* of $SO(n)$. If we consider the manifold which is a set of the orthogonal rectangular matrices

$\{W = (w_1, \ldots, w_p) \in \mathbb{R}^{n\times p} | W^\top W = I_p, n \geq p\}$, we get the (real) Stiefel manifold $\mathrm{St}(n,p,\mathbb{R})$. $SO(n)$ acts transitively on $\mathrm{St}(n,p,\mathbb{R})$ by usual matrix multiplication. The subgroup of $SO(n)$ which fixes a point $W \in \mathrm{St}(n,p,\mathbb{R})$ is called the isotropy subgroup, which is isomorphic to $SO(n-p)$. Manifold theory tells us that $\mathrm{St}(n,p,\mathbb{R})$ can be regarded as the quotient space $SO(n)/SO(n-p)$. In other words, we can interpret $SO(n)$ as a fiber bundle over $\mathrm{St}(n,p,\mathbb{R})$ whose fiber is isomorphic to $SO(n-p)$. A class of manifolds expressed as G/H are called homogeneous spaces, where G is a Lie group and H is a closed subgroup of G. Next comes the (real) Grassmann manifold $\mathrm{Gr}(n,p,\mathbb{R})$, which is defined as the set of p-dimensional subspaces in $\mathbb{R}^n$; $\mathrm{Gr}(n,p,\mathbb{R})$ concerns a subspace in $\mathbb{R}^n$ spanned by $w_1, \ldots, w_p$ instead of the individual frame vectors $(w_1, \ldots, w_p)$ themselves. In other words, any two matrices $W_1, W_2 \in \mathrm{St}(n,p,\mathbb{R})$ related by $W_2 = W_1 R$, where $R \in SO(p)$, correspond to the same point on $\mathrm{Gr}(n,p,\mathbb{R})$: we say we *identify* these two matrices. More formally said, $\mathrm{St}(n,p,\mathbb{R})$ is a fiber bundle over $\mathrm{Gr}(n,p,\mathbb{R})$, whose fiber is isomorphic to $SO(p)$. Therefore, as a homogeneous space, $\mathrm{Gr}(n,p,\mathbb{R}) \simeq SO(n)/SO(p) \times SO(n-p)$.

Let us introduce the generalized flag manifold $\mathrm{Fl}(n,d_1,\ldots,d_r,\mathbb{R})$, which is by definition the set of the direct sum of subspaces $V = V_1 \oplus V_2 \oplus \cdots \oplus V_r \subset \mathbb{R}^n$, where r and each $\dim V_i := d_i$ are fixed ($\sum_{i=1}^r d_i = p$).[1] We represent a point on this manifold by $W \in \mathrm{St}(n,p,\mathbb{R})$, which can be decomposed as $W = (W_1, W_2, \ldots, W_r), W_i = (w_1^i, w_2^i, \ldots, w_{d_i}^i)$, where $w_k^i \in \mathbb{R}^n$, $k = 1, \ldots, d_i$ for some i, form the orthogonal basis of V_i. As in the case of $\mathrm{Gr}(n,p,\mathbb{R})$, we are concerned about each subspace V_i rather than frame vectors w_k^i themselves, hence, as a point on $\mathrm{Fl}(n,d_1,\ldots,d_r,\mathbb{R})$, any two matrices $W_1, W_2 \in \mathrm{St}(n,p,\mathbb{R})$ related by $W_2 = W_1 \mathrm{diag}(R_1, R_2, \ldots, R_r)$, are identified, where $R_i \in SO(d_i)$; namely, $\mathrm{St}(n,p,\mathbb{R})$ is a fiber bundle over $\mathrm{Fl}(n,d_1,\ldots,d_r,\mathbb{R})$, whose fiber is isomorphic to $SO(d_1) \times \cdots SO(d_r)$. As a homogeneous space, $\mathrm{Fl}(n,d_1,\ldots,d_r,\mathbb{R}) \cong SO(n)/SO(d_1) \times \cdots \times SO(d_r) \times SO(n-p)$. $\mathrm{Fl}(n,d_1,\ldots,d_r,\mathbb{R})$ is a generalization of both $\mathrm{St}(n,p,\mathbb{R})$ and $\mathrm{Gr}(n,p,\mathbb{R})$, in the sense that it reduces to the Stiefel manifold if all d_i $(1 \leq i \leq r) = 1$, and it reduces to the Grassmann manifold if $r = 1$.

To derive the update rule for the Riemannian gradient descent geodesic method, we need to obtain the formulas for the natural gradient and geodesics on $\mathrm{Fl}(n,d_1,\ldots,d_r,\mathbb{R})$. By differentiating the constraints on the flag manifold, we see a tangent vector $V = (V_1, \ldots, V_r)$ of $\mathrm{Fl}(n,d_1,\ldots,d_r,\mathbb{R})$ at $W = (W_1, \ldots, W_r)$ is characterized by

$$W^\top V + V^\top W = O, W_i^\top V_i = O, \quad i = 1, \ldots, r. \tag{1}$$

First, let us derive the equation of a geodesic on the flag manifold; it can be obtained based on our geodesic formula for the Stiefel manifold with respect to the normal metric: $g_W^{\mathrm{St}(n,p,\mathbb{R})}(V_1, V_2) = \mathrm{tr} V_1^\top (I - \frac{1}{2} W W^\top) V_2$, where $V_1, V_2 \in T_W \mathrm{St}(n,p,\mathbb{R})$ [9].

$$\varphi_{\mathrm{St}(n,p,\mathbb{R})}(W, -\mathrm{grad}_W^{\mathrm{St}(n,p,\mathbb{R})} f, t) = \exp(-t(\nabla f(W) W^\top - W \nabla f(W)^\top)) W \tag{2}$$

[1] This definition is slightly different from the usual definition of the flag manifold [7], yet both are isomorphic to each other. For more details, see [10].

Here we recall the following theorem: Let $p : \tilde{M} \to M$ be a Riemannian submersion (see Fig. 2), that is, for any $\tilde{m} \in \tilde{M}$, $(dp)_{\tilde{m}}$ is an isometry between $H_{\tilde{m}}$ and $T_{p(\tilde{m})}M$, where $H_{\tilde{m}}$ is the horizontal space in $T_{\tilde{m}}\tilde{M}$. Let $\tilde{c}(t)$ be a geodesic of $(\tilde{M}, \tilde{g})$. If the vector $\dot{\tilde{c}}(0)$ is horizontal, then $\dot{\tilde{c}}(t)$ is horizontal for any t, and the curve $p(\tilde{c}(t))$ is a geodesic of (M, g) of the same length as $\tilde{c}(t)$ [6, 9]. Because the projection $\pi : \mathrm{St}(n, p, \mathbb{R}) \to \mathrm{Fl}(n, d_1, \ldots, d_r, \mathbb{R})$ is a Riemannian submersion, and any tangent vector $V \in T_W \mathrm{Fl}(n, d_1, \ldots, d_r, \mathbb{R})$ belongs to H_W in $T_W \mathrm{St}(n, p, \mathbb{R})$, this theorem ensures that the normal metric $g_W^{\mathrm{Fl}(n,d_1,\ldots,d_r,\mathbb{R})}$ coincides with $g_W^{\mathrm{St}(n,p,\mathbb{R})}$ on $T_W \mathrm{Fl}(n, d_1, \ldots, d_r, \mathbb{R})$, and that

$$\varphi_{\mathrm{Fl}(n,d_1,\ldots,d_r,\mathbb{R})}(W, -\mathrm{grad}_W^{\mathrm{Fl}(n,d_1,\ldots,d_r,\mathbb{R})} f, t) = \varphi_{\mathrm{St}(n,p,\mathbb{R})}(W, -\mathrm{grad}_W^{\mathrm{Fl}(n,d_1,\ldots,d_r,\mathbb{R})} f, t). \quad (3)$$

Next, using the following notations: $G = I - \frac{1}{2} W W^\top$, $X = \nabla_W f = \left(\frac{\partial f}{\partial w_{ij}} \right) = (X_1, \ldots, X_r)$, $Y = G^{-1} \nabla_W f, Y_i = G^{-1} X_i$, we can get the natural gradient V of a function f on $\mathrm{Fl}(n, d_1, \ldots, d_r, \mathbb{R})$ at W with respect to $g^{\mathrm{Fl}(n,d_1,\ldots,d_r,\mathbb{R})}$ by the orthogonal projection of Y onto $T_W \mathrm{Fl}(n, d_1, \ldots, d_r, \mathbb{R})$ relative to $g_W^{\mathrm{Fl}(n,d_1,\ldots,d_r,\mathbb{R})}$. In other words, V is obtained by minimizing $\mathrm{tr}\left\{(V-Y)^\top G (V-Y)\right\}$ under the tangency constraints of V. This can be solved by the Lagrangian multiplier method and we get:

$$V_i = X_i - (W_i W_i^\top X_i + \sum_{j \neq i} W_j X_j^\top W_i). \quad (4)$$

SUBSPACE ICA

Subspace ICA (a.k.a. independent subspace analysis) was proposed by Hyvärinen and Hoyer [8] by relaxing the assumption of normal ICA, namely each source signal is statistically independent. The subspace ICA task is to decompose a gray-scale image $I(x, y)$ into linear combination of basis images $a_i(x, y)$: $I(x, y) = \sum_{i=1}^n s_i a_i(x, y)$, where s_i is a coefficient. Let the inverse filter of this model be $s_i = \langle w_i, I \rangle = \sum_{x,y} w_i(x, y) I(x, y)$. The goal is to estimate s_i (or equivalently $w_i(x, y)$) from a set of given images. In the subspace ICA model, we assume $s = (s_1, \ldots, s_n)^\top$ is decomposed into disjoint subspaces $S_1, \ldots, S_r, (\dim S_i = d_i)$, where signals within each subspace are allowed to be dependent on each other, and signals belonging to different subspaces are statistically independent. As a cost function to solve this task, we take the negative log-likelihood:

$$f(\{w_i\}) = -\sum_{k=1}^K \log L(I_k; \{w_i\}) = -\sum_{k=1}^K \sum_{j=1}^r \log p \left(\sum_{i \in S_j} \langle w_i, I_k \rangle^2 \right) \quad (5)$$

where k denotes the index of sample images and p denotes the exponential distribution $p(x) = \alpha \exp(-\alpha x)$. Since the subspace ICA algorithm uses pre-whitening, solving the subspace ICA task reduces to minimizing f over the special orthogonal group $SO(n)$, as normal ICA. However, because of the statistical dependence of signals within each S_i, the objective function f is invariant under rotation within each subspace:

$W \mapsto W\mathrm{diag}(R_1, \ldots, R_r)$, where $R_i \in SO(d_i)$. Therefore, the subspace ICA task should be regarded as optimization on the generalized flag manifold $\mathrm{Fl}(n, d_1, \ldots, d_r, \mathbb{R})$ instead of simply $SO(n)$.

To demonstrate the Riemannian optimization method is effective, we applied it to the following subspace ICA task: We prepared 10000 image patches of 16×16 pixels at random locations extracted from monochrome photographs of natural images. (The dataset and subspace ICA code is distributed by Hyvärinen http://www.cis.hut.fi/projects/ica/data/images). As a preprocessing step, the mean gray-scale value of each image patch was subtracted; then the dimension of the image was reduced from 256 to 160 by PCA ($n = 160$), and the data were whitened. We performed subspace ICA on this dataset; the 160-dimensional vector space were decomposed into 40×4-dimensional subspaces (i.e. $r = 40$, $d_i = 4$) by minimizing f over $\mathrm{Fl}(160, 4, \ldots, 4, \mathbb{R})$. We compared the Riemannian optimization method with the ordinary gradient descent method used in [8] for this minimization problem. The former is: $W_{k+1} = \varphi_{\mathrm{Fl}(160,4,\ldots,4,\mathbb{R})}(W_k, -\mathrm{grad}^{\mathrm{Fl}(160,4,\ldots,4,\mathbb{R})}_{W_k} f(W_k), \eta_k) := \gamma_1(\eta_k)$, while the latter is: $W_{s+1} = \mathrm{proj}(W_s - \mu_s \frac{\partial f}{\partial W_s}) := \gamma_2(\mu_s)$, where proj means the projection onto $SO(160)$ via SVD. The learning constant η_k, μ_s was chosen at each iteration based on the Armijo rule such that

$$f(W_k) - f(\gamma_1(\eta_k)) \geq \frac{1}{2}\eta_k g^{\mathrm{Fl}}_{W_k}(\Delta_k, \Delta_k), f(W_k) - f(\gamma_1(2\eta_k)) \leq \eta_k g^{\mathrm{Fl}}_{W_k}(\Delta_k, \Delta_k) \quad (6)$$

$$f(W_s) - f(\gamma_2(\mu_s)) \geq \frac{1}{2}\mu_s \langle \delta_s, \delta_s \rangle, f(W_s) - f(\gamma_2(2\mu_s)) \leq \mu_s \langle \delta_s, \delta_s \rangle \quad (7)$$

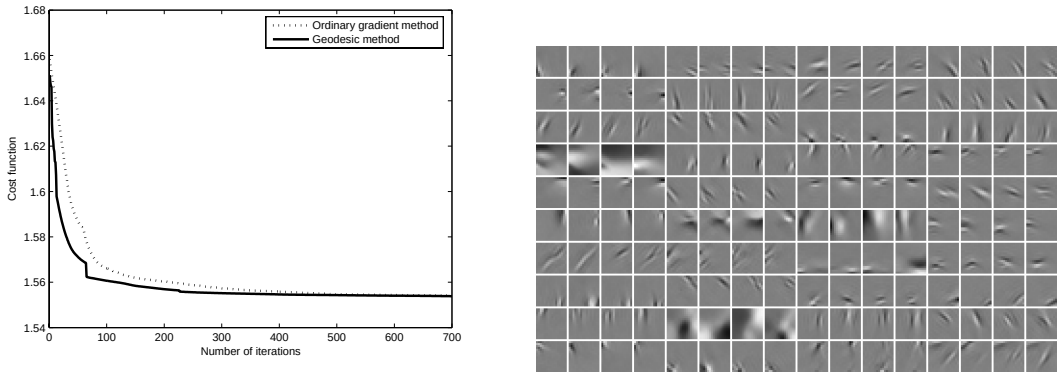
are satisfied, where Fl denotes $\mathrm{Fl}(160, 4, \ldots, 4, \mathbb{R})$, $\Delta = -\mathrm{grad}^{\mathrm{Fl}(160,4,\ldots,4,\mathbb{R})}_{W_k} f(W_k)$, δ_s denotes the orthogonal projection of $-\frac{\partial f}{\partial W_s}$ onto $T_{W_s}SO(160)$ with respect to the Euclidean metric $\langle \quad , \quad \rangle$. The behavior of these algorithms is shown in Fig. 3(a). In the early stages of learning, the cost associated with the geodesic method decreased much faster than the cost with the ordinary gradient method. The recovered inverse filters by the geodesic method $w_i(x, y)$ are shown in Fig. 3(b). We obtained complex cell-like filters, which were grouped into 4-dimensional subspaces. We found no significant difference between the points of convergence of the two methods, and neither method appeared to get 'stuck' in a local minimum.

COMPLEX ICA

Let us consider an optimization problem on the complex Stiefel manifold.

$$F : \mathrm{St}(n, p, \mathbb{C}) \to \mathbb{R}, \quad (8)$$

where $\mathrm{St}(n, p, \mathbb{C}) = \{W = (w_1, \ldots, w_p) = W_{\Re} + iW_{\Im} \in \mathbb{C}^{n \times p} | W^H W = I_p\}$ (H denotes the Hermitian transpose operator). We assume F is a smooth function of the norm of column vectors $||w_i||(i = 1, \ldots, p)$, which is satisfied by many signal processing tasks including complex ICA.



(a) (b)

FIGURE 3. Results, showing (a) learning curves; (b) recovered inverse filters

Because the cost function F is real-valued, $\mathrm{St}(n,p,\mathbb{C})$ should be regarded as a *real manifold* rather than a complex manifold. The real manifold underlying $\mathrm{St}(n,p,\mathbb{C})$ is a submanifold M in $\mathbb{R}^{2n\times p}$ defined by the constraints:

$$M := \{\overline{W} = \begin{pmatrix} W_{\Re} \\ W_{\Im} \end{pmatrix} \in \mathbb{R}^{2n\times p} | W_{\Re}^{\top} W_{\Re} + W_{\Im}^{\top} W_{\Im} = I_p, W_{\Re}^{\top} W_{\Im} - W_{\Im}^{\top} W_{\Re} = O_p\}. \quad (9)$$

The cost function F over $\mathrm{St}(n,p,\mathbb{C})$ corresponds to the function $F'(\overline{W}) := F(W)$ over M. However, it is difficult to deal with the constraints (9) as is; we embed M into $\mathbb{R}^{2n\times 2p}$ by the following map:

$$\tau : \begin{pmatrix} W_{\Re} \\ W_{\Im} \end{pmatrix} = \begin{pmatrix} w_1^{\Re} \cdots w_p^{\Re} \\ w_1^{\Im} \cdots w_p^{\Im} \end{pmatrix} \mapsto \tilde{W} = \begin{pmatrix} w_1^{\Re} & -w_1^{\Im} & w_2^{\Re} & -w_2^{\Im} & \cdots & w_p^{\Re} & -w_p^{\Im} \\ w_1^{\Im} & w_1^{\Re} & w_2^{\Im} & w_2^{\Re} & \cdots & w_p^{\Im} & w_p^{\Re} \end{pmatrix}. \quad (10)$$

We consider the embedded manifold $N = \tau(M)$ in $\mathbb{R}^{2n\times 2p}$ and the function $f : N \to \mathbb{R}$ associated with the embedding τ. $\tilde{W} \mapsto f(\tilde{W}) := F'(\overline{W})$. If $\overline{W} \in M$, then $\tilde{W} \in \mathrm{St}(2n,2p,\mathbb{R})$ holds. It turns out that $N = \mathrm{St}(2n,2p,\mathbb{R}) \cap T$, where $T = \tau(\mathbb{R}^{2n\times p})$ forms a subspace in $\mathbb{R}^{2n\times 2p}$. As such, minimizing F over $\mathrm{St}(n,p,\mathbb{C})$ is transformed to minimizing f over N.

Furthermore, the assumption of F gives N an additional structure. We see the transformation on $\mathrm{St}(n,p,\mathbb{C})$:

$$W = (w_1, \ldots, w_p) \mapsto \left(e^{i\theta_1} w_1, \ldots, e^{i\theta_p} w_p\right) \quad (11)$$

corresponds to the transformation on N :

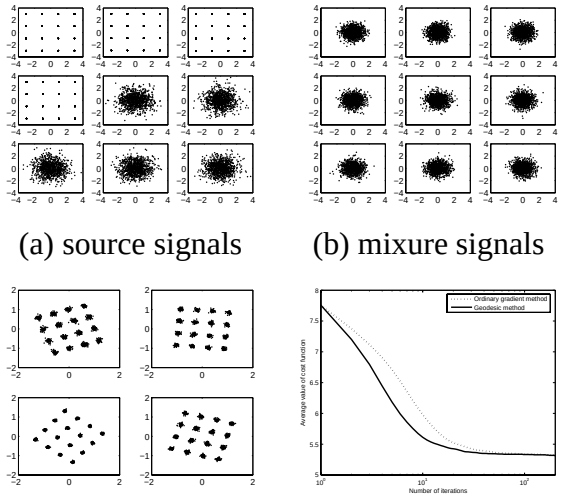
$$\tilde{W} \mapsto \tilde{W}\mathrm{diag}(R(\theta_1), R(\theta_2), \cdots, R(\theta_p)), \text{where } R(\theta_i) = \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix}, \quad (12)$$

and F is invariant under the transformation (11) from the assumption. So is f with respect to the transformation (12). Therefore, f can be interpreted as a function over a submanifold of the flag manifold: $N' = \mathrm{Fl}(2n, 2, \ldots, 2, \mathbb{R}) \cap T$.

In fact, the following two facts allow us to consider just $\mathrm{Fl}(2n,2,\ldots,2,\mathbb{R})$ instead of its submanifold N'. First, N' is a totally geodesic submanifold of $\mathrm{Fl}(2n,2,\ldots,2,\mathbb{R})$, that is, a geodesic on $\mathrm{Fl}(2n,2,\ldots,2,\mathbb{R})$ emanating from $\tilde{W} \in N'$ in direction $\tilde{V} \in T_{\tilde{W}}N'$ is always contained in N'. Second, the natural gradient of f on N' at $\tilde{W}$ coincides with the natural gradient of f on $\mathrm{Fl}(2n,2,\ldots,2,\mathbb{R})$ at $\tilde{W}$, that is, we can obtain $\mathrm{grad}^{N'}_{\tilde{W}} f$ by substituting $\begin{pmatrix} \frac{\partial f}{\partial w_i^{\Re}} & -\frac{\partial f}{\partial w_i^{\Im}} \\ \frac{\partial f}{\partial w_i^{\Im}} & \frac{\partial f}{\partial w_i^{\Re}} \end{pmatrix}$ for X_i in the formula for the natural gradient of $\mathrm{Fl}(2n,2,\ldots,2,\mathbb{R}), (d_i = 2, r = p)$. Note that $(X_1,\ldots,X_p)$ is the gradient of f in T relative to the Euclidean metric. To summarize, minimizing F over $\mathrm{St}(n,p,\mathbb{C})$ can be solved by minimizing the function f over the submanifold N' of $\mathrm{Fl}(2n,2,\ldots,2,\mathbb{R})$; for minimizing f on N', we have only to apply the Riemannian optimization method for $\mathrm{Fl}(2n,2,\ldots,2,\mathbb{R})$ to f.

To explore the behavior of the Riemannian gradient descent geodesic method on the complex Stiefel manifold as described above, we performed a numerical experiment for complex ICA. Let us assume we are given 9 source signals $x = (x_1,\ldots,x_9)^\top$ (Fig. 4(b)) which are complex-valued instantaneous linear mixture of four independent QAM16 signals $s = (s_1,\ldots,s_4)^\top$ and five complex-valued Gaussian noise signals $u = (s_5,\ldots,s_9)^\top$ (Fig. 4(a)) such that $x = A \begin{pmatrix} s \\ u \end{pmatrix}$, where A is a randomly generated nonsingular 9×9 matrix. We assume we know in advance the number of the noise signals and the task of complex ICA is to recover only non-noise signals $y = (y_1,\ldots,y_4)^\top$ so that $y = W^\top x$. As a preprocessing stage, we first center the data and then whiten it by SVD. Thus, $n \times p$ demixing matrix W can be regarded as a point on the complex Stiefel manifold $\mathrm{St}(n,p,\mathbb{C})$, namely $W^H W = I_p$. As an object function, we use a kurtosis-like higher-order statistics: $F(W) = \sum_{i=1}^{4} E\left[|y_i(t)|^4\right]$ [4], then by minimizing $F(W)$ over $\mathrm{St}(n,p,\mathbb{C})$ we can solve the task.

We compared two algorithms for optimizing $F(W)$ over $\mathrm{St}(n,p,\mathbb{C})$. One is the Riemannian optimization method: $\tilde{W}_{k+1} = \varphi_{\mathrm{Fl}(2n,2,\ldots,2,\mathbb{R})}(\tilde{W}_k, -\mathrm{grad}_{\tilde{W}_k} f(\tilde{W}_k), \eta_k) := \gamma_1(\eta_k)$, and another is the ordinary (Euclidean) gradient method followed by projection: $W_{s+1} = \mathrm{proj}(W_s - \mu_s \frac{\partial f}{\partial \overline{W}_s}) := \gamma_2(\mu_s)$, where $\frac{\partial f}{\partial \overline{W}_s}$ denotes $\frac{\partial f}{\partial W_{\Re}} + i\frac{\partial f}{\partial W_{\Im}}$, and proj means the projection onto $\mathrm{St}(n,p,\mathbb{C})$ via complex SVD. Both $\mathrm{grad}_{\tilde{W}_k} f(\tilde{W}_k)$ and $\frac{\partial f}{\partial \overline{W}_s}$ are computed by substituting $\frac{\partial ||y_i||^4}{\partial w_i^{\Re}} = 2||y_i||^2(y_i^* x + y_i x^*)$ and $\frac{\partial ||y_i||^4}{\partial w_i^{\Im}} = 2i||y_i||^2(y_i^* x - y_i x^*)$. Recall that we map $\mathrm{St}(n,p,\mathbb{C})$ to $\mathrm{Fl}(2n,2,\ldots,2,\mathbb{R})$ and the Riemannian optimization method for f updates the matrices on $\mathrm{Fl}(2n,2,\ldots,2,\mathbb{R})$ via the correspondence between W and $\tilde{W}$ (10). After $\tilde{W}$ converges to $\tilde{W}_\infty$, $\tilde{W}_\infty$ is pulled back to $\mathrm{St}(n,p,\mathbb{C})$ to give a demixing matrix W_∞. We followed the Armijo rule to set the learning constant at each iteration as the subspace ICA experiment. The separation result is shown in Fig. 4(c). The QAM 16 constellation was well-recognized after recovery. Both algorithms were tested for 100 trials. On each trial, a random nonsingular matrix was used to generate the data; a random unitary matrix was chosen as a initial demixing matrix; we iterated for 200 steps. The plots of Fig. 4(d) show the average behavior of these two algorithms over 100 trials. We observed that the Riemannian optimization method outperformed the ordinary gradient method followed by projection, particularly in the early stages of



(a) source signals (b) mixure signals

(c) recovered signals (d) cost function

FIGURE 4. Complex ICA experiment

learning much the same way as the subspace ICA experiment.

ACKNOWLEDGEMENTS

This work is partly supported by JSPS Grant-in-Aid for Exploratory Research 16650050, and MEXT Grant-in-Aid for Scientific Research on Priority Areas 17022033.

REFERENCES

1. P-A. Absil, R. Mahony, and R. Sepulchre, Riemannian geometry of Grassmann manifolds with a view on algorithmic computation, *Acta Applicandae Mathematicae*, **80**(2), pp.199-220, 2004.
2. S. Amari, Natural gradient works efficiently in Learning, *Neural Computation*, **10**, pp.251-276, 1998.
3. A. Edelman, T.A. Arias, and S.T. Smith, The Geometry of algorithms with orthogonality constraints, *SIAM Journal on Matrix Analysis and Applications*, **20**(2), pp.303-353, 1998.
4. S. Fiori, Complex-Weighted One-Unit 'Rigid-Bodies' Learning Rule for Independent Component Analysis, *Neural Processing Letters*, **15**(3), 2002.
5. S. Fiori, Quasi-Geodesic Neural Learning Algorithms over the Orthogonal Group: A Tutorial, *Journal of Machine Learning Research*, **6**, pp.743-781, 2005.
6. S. Gallot, D. Hulin, and J. Lafontaine, *Riemannian Geometry*, Springer, 1990.
7. U. Helmke, J. B. Moore, *Optimization and dynamical systems*, Springer-Verlag, 1994.
8. A. Hyvärinen and P.O. Hoyer, Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, **12**(7), pp.1705-1720, 2000.
9. Y. Nishimori and S. Akaho, Learning Algorithms Utilizing Quasi-Geodesic Flows on the Stiefel Manifold, *Neurocomputing*, **67** pp.106-135, 2005.
10. Y. Nishimori, S. Akaho, and M D. Plumbley Riemannian Optimization Method on the Flag Manifold for Independent Subspace Analysis, *Proceedings of 6th International Conference ICA2006*, pp.295-302, 2006.

11. M. D. Plumbley, Algorithms for non-negative independent component analysis. *IEEE Transactions on Neural Networks*, **14**(3), pp.534-543, 2003.