Riemannian-Gradient-Based Learning on the Complex Matrix-Hypersphere

Simone Fiori

***Abstract*—The present manuscript tackles the problem of learning over the complex-valued matrix-hypersphere $\mathbb{S}^{\alpha}_{n,p}(\mathbb{C})$. The developed learning theory is formulated in terms of Riemannian-gradient-based optimization of a regular criterion function and is implemented by a geodesic-stepping method. The stepping method is equipped with a geodesic-search sub-algorithm to compute the optimal learning stepsize at any step. Numerical results show the effectiveness of the developed learning method and of its implementation.**

***Index Terms*—Complex matrix-hypersphere; Complex-valued neural networks; Riemannian-gradient-based learning; Geodesic-stepping; Geodesic-search; MIMO broadcast channels.**

I. Introduction

Complex-valued neural networks are based on artificial neurons with complex-valued weights and complex-valued activation functions. A number of original solutions in pattern recognition and classification, artificial neural information processing, image processing and in the theory of artificial neurons and neural networks that are based on the use of complex-valued neurons have been proposed. The use of complex-valued weights and complex-valued activation functions is not simply a theoretical generalization of the real-valued case but it makes it possible to extend the functionality of a single neuron and of a network to obtain more stable learning algorithms and to solve applied problems that do not accommodate well in the framework of real-valued neural networks [1].

Complex-valued neural networks find applications in adaptive signal processing for highly functional sensing and imaging, control in unknown and changing environment, brain-like information processing and robotics inspired by human neural systems [22]. In the field of signal processing, complex-valued neural networks are widely applied (examples of signal processing applications of complex-valued neural networks are to land-surface classification, to removal of phase singular points to create digital elevation maps and to pitch-asynchronous overlap-add waveform-concatenation speech synthesis by optimizing phase spectrum in frequency domain).

The present analysis of the problem of learning by optimization of a regular criterion function over the complex-valued matrix-hypersphere $\mathbb{S}^{\alpha}_{n,p}(\mathbb{C})$ is motivated by the technical problem posed in [23], which concerns the computation of an optimal precoding matrix via maximization of a weighted sum rate in MIMO broadcast channels. For a recent review of weighted sum rate in MIMO broadcast channels see, e.g., [21], [24], [25]. The main focus of the present contribution is to develop a suitable learning algorithm on the basis of a formulation of the problem drawn from the specific literature.

The present paper fits within the research line of differential geometrical methods for machine learning and neural networks design [2]. In particular, it lies on the intersection between the author's research line about learning by optimization on Riemannian (as well as pseudo-Riemannian) manifolds with application, e.g., to blind signal deconvolution [9], [13], blind source separation, latent variable analysis and independent component analysis [11], [17], unsupervised machine learning by optimization on differentiable manifolds and Lie groups [4], [8], [12], [15], [18], [19], and the author's research line about complex-valued neural networks [7], [10], [14].

As the real line $\mathbb{R}$ is a subfield of the complex plane $\mathbb{C}$, the developed optimization algorithm translates to the real-valued parameter space $\mathbb{S}^{\alpha}_{n,p}(\mathbb{R})$ in a straightforward way, which proves useful in certain applications. An application of learning by optimization on the manifold $\mathbb{S}^{1}_{n,p}(\mathbb{R})$ is to dimension reduction for image retrieval [30], which was designed to optimize class separation with respect to metrics derived from cross-correlation of spectral histograms. A problem formally similar to the one discussed in [23] tailored to the real line was discussed in the paper [29] under the name of "maximum relative uncertainty theory" for linear-neural-networks' unsupervised learning (with special reference to principal component/subspace analysis). Moreover, the problem of optimization over the manifold $\mathbb{S}^{\alpha}_{n,1}(\mathbb{R})$ arises in applications such as blind channel deconvolution [9], [13] and one-unit principal/independent component analysis [16].

The sought-for learning rule over the manifold $\mathbb{S}^{\alpha}_{n,p}(\mathbb{C})$ is formulated in terms of Riemannian-gradient based optimization of a regular criterion function, as the space of interest $\mathbb{S}^{\alpha}_{n,p}(\mathbb{C})$ is a smooth manifold that may be endowed with a Riemannian geometry. The discussed Riemannian-gradient based optimization theory is implemented by a geodesic-stepping method. Geodesic stepping is based on the calculation of geodesic arcs in closed form and provides a geometrically sound way of moving from a point along a given direction on a Riemannian manifold proportionally to an assigned stepsize. A method to compute a numerically-optimal learning stepsize schedule is discussed as well, which resembles the line-search method on Euclidean spaces, termed *geodesic-search* method. An advantage of the devised geodesic-search method is that, in the space $\mathbb{S}^{\alpha}_{n,p}(\mathbb{C})$, the geodesic curve is periodic of finite period, hence the learning stepsize belongs to a closed interval. Moreover, unlike other methods previously adopted [13], it

The author is with the Dipartimento di Ingegneria dell'Informazione, Facoltà di Ingegneria, Università Politecnica delle Marche, Via Brecce Bianche, Ancona I-60131, Italy (E-mail: s.fiori@univpm.it)

does not rely on any local approximation of the criterion function and hence does not limit to short steps.

Numerical results about the solution of the formal problem related to the optimal precoding for MIMO-broadcast-channels show the effectiveness of the learning method and of its numerical implementation. In particular, the obtained numerical results show that the developed optimization algorithm converges steadily and in a few iterations.

II. Derivation of a Riemannian-gradient optimization algorithm on the hypersphere $\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$

The parameter space of interest is the matrix-hypersphere $\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$, defined as:

$$\mathbb{S}_{n,p}^{\alpha}(\mathbb{C}) \stackrel{\text{def}}{=} \{x \in \mathbb{C}^{n\times p} | \text{tr}(x^H x) = \alpha\}, \tag{1}$$

where $n \geq p$ and the symbol $\mathbb{C}^{n\times p}$ denotes the set of $n \times p$ matrices with complex-valued entries. Superscript ^{H} denotes Hermitian transpose, the operator $\text{tr}(\cdot)$ denotes matrix trace and $\alpha > 0$.

A. Geometric characterization of the space $\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$

The present section makes use of notions from differential geometry. A reference on differential geometry is [27].

The tangent space to the manifold $\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$ at a point $x \in \mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$ is denoted by $T_x\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$. Given any smooth curve $\gamma(t)$ such that $\gamma : [-a, a] \to \mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$, with $a > 0$ and $\gamma(0) = x$, the tangent space $T_x\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$ is spanned by the vectors $\dot{\gamma}(0)$, where an over-dot denotes derivative with respect to parameter t. Hand-by-hand derivation of the condition $\text{tr}(\gamma^H(t)\gamma(t)) = \alpha$ gives, for any $t \in [-a, a]$:

$$\text{tr}(\dot{\gamma}^H(t)\gamma(t) + \gamma^H(t)\dot{\gamma}(t)) = 2\Re\text{tr}(\dot{\gamma}^H(t)\gamma(t)) = 0,$$

where symbol $\Re$ denotes real part. Setting $v \stackrel{\text{def}}{=} \dot{\gamma}(0) \in T_x\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$, one finds that the tangent space of the smooth manifold $\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$ at a point $x \in \mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$ is described by:

$$T_x\mathbb{S}_{n,p}^{\alpha}(\mathbb{C}) = \{v \in \mathbb{C}^{n\times p} | \Re\text{tr}(v^H x) = 0\}. \tag{2}$$

By embedding the manifold $\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$ into the ambient space $\mathbb{C}^{n\times p}$ equipped with the inner product $\langle z, w\rangle \stackrel{\text{def}}{=} \Re\text{tr}(z^H w)$, $z, w \in \mathbb{C}^{n\times p}$, the normal space $N_x\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$ to the manifold $\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$ at a point $x \in \mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$ may be defined as the collection of vectors that are orthogonal to the tangent space $T_x\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$, namely as:

$$N_x\mathbb{S}_{n,p}^{\alpha}(\mathbb{C}) \stackrel{\text{def}}{=} \{\nu \in \mathbb{C}^{n\times p} | \Re\text{tr}(\nu^H v) = 0,\ \forall v \in T_x\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})\}. \tag{3}$$

It represents the orthogonal complement of the tangent space with respect to the ambient space. The normal space $N_x\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$ admits the following characterization:

$$N_x\mathbb{S}_{n,p}^{\alpha}(\mathbb{C}) = \{\lambda x | \lambda \in \mathbb{R}\}, \tag{4}$$

in fact, $\Re\text{tr}(\nu^H v) = \Re\text{tr}((\lambda x)^H v) = \Re\text{tr}(\lambda x^H v) = \lambda\Re\text{tr}(x^H v) = 0$, for all $v \in T_x\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$ and $\lambda \in \mathbb{R}$.

Endowing the smooth manifold $\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$ with a inner product turns it into a Riemannian manifold. The following inner product at every point $x \in \mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$ is chosen:

$$\langle u, v\rangle_x \stackrel{\text{def}}{=} \Re\text{tr}(u^H v),\ u, v \in T_x\mathbb{S}_{n,p}^{\alpha}(\mathbb{C}). \tag{5}$$

The above inner product defines the norm $\|v\|_x \stackrel{\text{def}}{=} \sqrt{\text{tr}(v^H v)}$, $v \in T_x\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$, at every point $x \in \mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$.

The Riemannian gradient of a regular function $f : \mathbb{S}_{n,p}^{\alpha}(\mathbb{C}) \to \mathbb{R}$ is denoted as $\nabla_x f$ and satisfies the following conditions:

- *Tangency*: It holds that $\nabla_x f \in T_x\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$.
- *Compatibility with the metric*: It holds that $\langle v, \nabla_x f\rangle_x = \Re\text{tr}(v^H \partial_x f)$, for all $v \in T_x\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$.

The gradient $\partial_x f$ represents the best linear approximation of a regular function $f(x)$ in the sense that:

$$\lim_{\langle y,y\rangle^{\frac{1}{2}} \to 0} \frac{f(x+y) - f(x) - \langle \partial_x f, y\rangle}{\langle y, y\rangle^{\frac{1}{2}}} = 0. \tag{6}$$

From the metric compatibility condition, it follows that $\Re\text{tr}(v^H \nabla_x f) = \Re\text{tr}(v^H \partial_x f)$, $\forall v \in T_x\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$, which clearly implies that $\nabla_x f - \partial_x f \in N_x\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$, namely, that $\nabla_x f = \partial_x f + \lambda_x x$ for some $\lambda_x \in \mathbb{R}$. In addition, the tangency condition implies that $0 = \Re\text{tr}(x^H \nabla_x f) = \Re\text{tr}(x^H \partial_x f) + \lambda_x \Re\text{tr}(x^H x) = \Re\text{tr}(x^H \partial_x f) + \alpha\lambda_x$. Applying both conditions yields thus the Riemannian gradient on the manifold $\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$:

$$\nabla_x f = \partial_x f - \frac{x}{\alpha}\Re\text{tr}(x^H \partial_x f). \tag{7}$$

A smooth curve $\gamma : [0, 1] \to \mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$ is referred to as 'geodesic arc with normal parametrization' if it solves the following variational problem:

$$\delta \int_0^1 \langle \dot{\gamma}(t), \dot{\gamma}(t)\rangle_{\gamma(t)} dt = 0. \tag{8}$$

In the above expression, symbol δ denotes the variation of the integral. The variation $\delta\gamma \in T_\gamma\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$ is arbitrary, except at the boundaries of the curve, $\gamma(0)$ and $\gamma(1)$, where the variation vanishes to zero. The variation of the integral in (8) may be written explicitly as:

$$\int_0^1 \delta\text{tr}(\dot{\gamma}^H \dot{\gamma}) dt = 2\int_0^1 \Re\text{tr}\left(\dot{\gamma}^H \tfrac{d\delta\gamma}{dt}\right) dt = -2\int_0^1 \Re\text{tr}(\ddot{\gamma}^H \delta\gamma) dt,$$

upon integration by parts. As the last integral must vanish to zero for any admissible variation $\delta\gamma$, the geodesic arc is characterized by the condition $\ddot{\gamma} \in N_x\mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$, namely $\ddot{\gamma} = \lambda_\gamma \gamma$ for $\lambda_\gamma \in \mathbb{R}$. As $\gamma(t) \in \mathbb{S}_{n,p}^{\alpha}(\mathbb{C})$, it must hold $\Re\text{tr}(\gamma^H(t)\gamma(t)) = \alpha$ for any $t \in [0, 1]$. Deriving twice with respect to the parameter t gives the condition $\Re\text{tr}(\ddot{\gamma}^H \gamma + \dot{\gamma}^H \dot{\gamma}) = 0$. Replacing the term $\ddot{\gamma}$ with $\lambda_\gamma\gamma$ in the last equation gives $\Re\text{tr}(\lambda_\gamma \gamma^H \gamma + \dot{\gamma}^H \dot{\gamma}) = 0$ from which $\lambda_\gamma = -\alpha^{-1}\text{tr}(\dot{\gamma}^H \dot{\gamma})$. The equation of the geodesic curve reads, therefore:

$$\ddot{\gamma} + \alpha^{-1}\text{tr}(\dot{\gamma}^H \dot{\gamma})\gamma = 0. \tag{9}$$

The solution $G^{\alpha}_{x,v} : \mathbb{R} \to \mathbb{S}^{\alpha}_{n,p}(\mathbb{C})$ of the variational problem (8) with initial conditions $\gamma(0) = x$ and $\dot{\gamma}(0) = v$ reads:

$$\begin{aligned} G^{\alpha}_{x,v}(t) &= x \cos\left(t\sqrt{\frac{\mathrm{tr}(v^H v)}{\alpha}}\right) \\ &+ v\sqrt{\frac{\alpha}{\mathrm{tr}(v^H v)}} \sin\left(t\sqrt{\frac{\mathrm{tr}(v^H v)}{\alpha}}\right), \end{aligned} \quad (10)$$

for $v \neq 0$, while $G^{\alpha}_{x,0}(t) = x$, as it may be verified by substitution. Note that the inner product $\langle \dot{G}^{\alpha}_{x,v}(t), \dot{G}^{\alpha}_{x,v}(t) \rangle_{G^{\alpha}_{x,v}(t)} = \mathrm{tr}(v^H v)$ keeps constant over any geodesic arc and for any $t \neq 0$, $G_{x,v}(t) = G_{x,tv}(1)$; moreover, $\lim_{t\to 0} G_{x,tv}(1) = G_{x,v}(0)$.

The distance between two geodesically-connectible points on a Riemannian manifold defines as follows:

$$d(G_{x,v}(0), G_{x,v}(1)) \stackrel{\mathrm{def}}{=} \int_0^1 \|\dot{G}_{x,v}(t)\|_{G_{x,v}(t)} dt, \quad (11)$$

namely, as the length of the geodesic arc that connects them.

B. Geodesic-stepping optimization method

Given a regular function $f : \mathbb{S}^{\alpha}_{n,p}(\mathbb{C}) \to \mathbb{R}$, a gradient-steepest ascent algorithm to compute its maximum (or a local maximum) compatible with the geometrical structure of the parameter space is the geodesic-stepping method (see, for instance, [20], [26]). (Gradient-steepest-descent search to minimize the criterion function f may be achieved by applying the gradient-steepest-ascent search to maximize the criterion function $-f$, whenever appropriate.)

Geodesic-stepping methods are regarded as the counterparts of Euler stepping methods on curved spaces. Euler-stepping-based optimization consists in moving in the direction of the gradient of a criterion function along a straight line. Geodesic stepping extends Euler stepping by replacing the notion of straight line with the notion of geodesic arc. Geodesic steepest-gradient-ascent stepping may be expressed as:

$$x_{k+1} = G^{\alpha}_{x_k, \nabla_{x_k} f}(h_k) \text{ for } k \geq 0, \quad (12)$$

$$h_k = \arg\max_{t>0} \{f(G^{\alpha}_{x_k, \nabla_{x_k} f}(t))\}, \quad (13)$$

where $x_k \in \mathbb{S}^{\alpha}_{n,p}(\mathbb{C})$ denotes a sequence of discrete steps on the manifold of parameters with step-counter $k \in \mathbb{N}$. The term $h_k > 0$ denotes a sequence of optimization stepsizes. Likewise in Euler stepping methods, in the context of geodesic stepping methods, the length of a step is proportional to the learning stepsize. In fact, from the properties 1) and 2) of section II-A and from the definition of geodesic distance (11), it follows:

$$d(x_k, x_{k+1}) = h_k \|\nabla_{x_k} f\|_{x_k}. \quad (14)$$

By setting $\omega_k \stackrel{\mathrm{def}}{=} \alpha^{-\frac{1}{2}} \|\nabla_{x_k} f\|_x$ the geodesic-stepping algorithm (12) may be implemented on the manifold $\mathbb{S}^{\alpha}_{n,p}(\mathbb{C})$ as:

$$x_{k+1} = x_k \cos(h_k \omega_k) + \nabla_{x_k} f \sin(h_k \omega_k)/\omega_k \text{ for } k \geq 0, \quad (15)$$

as long as $\omega_k \neq 0$, otherwise the algorithm stops. In practice, the condition $\omega_k = 0$ corresponds to a critical point of the criterion function f and may be used to define a stopping criterion for the optimization algorithm. Denoting with $\varepsilon > 0$ a desired precision, the iteration may be halted as soon as $\|\nabla_{x_k} f\| < \varepsilon$. The geodesic stepping method on the manifold $\mathbb{S}^{\alpha}_{n,p}(\mathbb{C})$ is illustrated in the Figure 1.

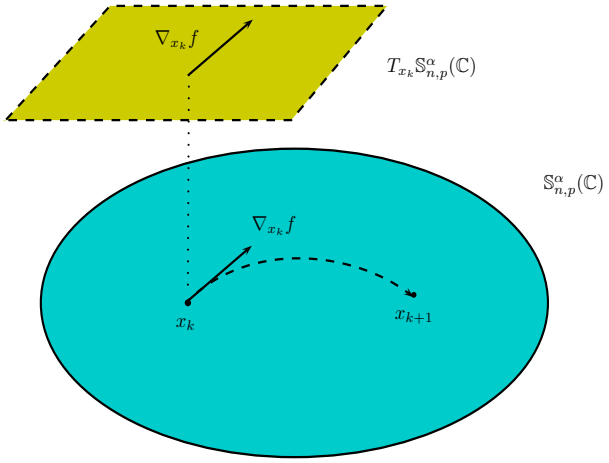

Fig. 1. Rendering of geodesic-stepping-based optimization of a criterion function f on the manifold $\mathbb{S}^{\alpha}_{n,p}(\mathbb{C})$.

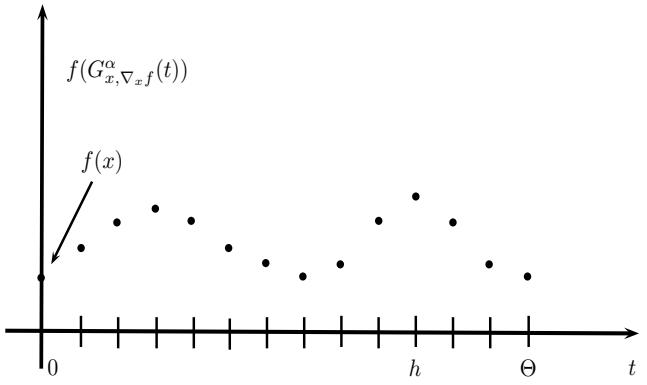

Fig. 2. Geodesic-search method: The learning/optimization stepsize is selected at any step as the value of the parameter t that ensures maximal increase of the criterion function over a geodesic arc, namely $f(G^{\alpha}_{x,\nabla_x f}(t))$, with respect to the value $f(G^{\alpha}_{x,\nabla_x f}(0)) = f(x)$, up to the finesse of the sampling.

The choice of the learning-stepsize schedule h_k is facilitated by the observation that the geodesic function (10) with $\|v\|_x \neq 0$, is periodic in the argument t of period $\Theta \stackrel{\mathrm{def}}{=} \frac{2\pi\sqrt{\alpha}}{\|v\|_x}$. Namely, given a smooth function $f : \mathbb{S}^{\alpha}_{n,p}(\mathbb{C}) \to \mathbb{R}$ to optimize, it holds that $f(G^{\alpha}_{x,\nabla_x f}(t+\Theta)) = f(G^{\alpha}_{x,\nabla_x f}(t))$. The search of an optimal learning stepsize may thus be restricted to the closed interval $[0, \Theta]$. Numerically, for a fixed matrix $x \in \mathbb{S}^{\alpha}_{n,p}(\mathbb{C})$, the optimal stepsize (13) may be approximated by sampling the function $f(G^{\alpha}_{x,\nabla_x f}(t))$ at points $t = \frac{s\Theta}{S}$, where constant S denotes the number of sampling locations and $s = 1, 2, \ldots, S$. The stepsize is selected as the value $\frac{s\Theta}{S}$ that guarantees the maximum increase of the function f with respect to the preceding learning step, namely, with respect to value $f(x)$. If no increase can be achieved, the learning process stops. The geodesic-search method is illustrated in the Figure 2.

C. Properties of the gradient-steepest descent learning method

The learning method (12)-(13) falls within the class of geodesic-step Riemannian-gradient-steepest-ascent methods studied by Luemberger and and by Gabay, among others [20].

The convergence of the geodesic-based Riemannian-gradient-steepest-descent optimization method (12), endowed with the stepsize selection rule (13), assumed to start from the point $x_0 \in \mathbb{S}^{\alpha}_{n,p}(\mathbb{C})$ which is supposed to differ from a critical point of the function f, is ensured by the following results:

1) Assume that the function f is continuously differentiable, that its critical values are distinct and that the connected subset of the level-set $\{x \in \mathbb{S}^{\alpha}_{n,p}(\mathbb{C}) | f(x) < f(x_0)\}$ containing x_0 is compact. Then the sequence x_k constructed by the method (12) and (13) converges to a critical point of the function f.
2) Moreover, under appropriate conditions, the gradient-based optimization method converges linearly.

The property 1) may be proven by applying Theorem 4.3 of paper [20], while property 2) follows from the companion Theorem 4.4.

A learning-by-optimization problem related to the one discussed within the present manuscript was discussed in the contribution [5] on the basis of the manifold $\mathbb{S}^{D}_{n,p}(\mathbb{R}) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^{n \times p} | x^T x = D\}$, with D being a semi-positive-definite diagonal matrix. Paper [6] showed that a maximization/minimization algorithm on $\mathbb{S}^{D}_{n,p}(\mathbb{R})$ will encounter the problem that the manifold is stable for one of the algorithm and unstable for the other, or both are neutral stable. Such a difficulty is due to the fact that, although the learning rule discussed in [5] was derived in a geometrically-sound way, its numerical implementation was not based on geodesic stepping. Whenever a dynamics happen in the ambient space ($\mathbb{R}^{n \times p}$, in this case) it is necessary to verify whether the manifold of interest ($\mathbb{S}^{D}_{n,p}(\mathbb{R})$, in the present case) is an attractor. Whenever a geometrically-sound numerical integration method is made use of, as in the present case, the numerical dynamics takes place directly inside the manifold of interest, hence there is no need to check whether it is an attractor.

III. Application to optimal precoding in MIMO broadcast channels

The present section describes an application that implies optimization over the hypersphere $\mathbb{S}^{\alpha}_{n,p}(\mathbb{C})$.

In a multiple-input multiple-output (MIMO) communication channel, transmitters and receivers may cooperate [3]. The transmitters can cooperate if the messages are jointly encoded into the components of the input vector, instead of being carried on by each entry separately. Likewise, receivers can cooperate if the whole output vector instead of each individual entry of the output is used to decode the messages. Cooperation of transmitters implies that the elements in the input of the channel are correlated. If both transmitters and receivers are allowed to cooperate, it represents a single-user MIMO Gaussian channel, arising in multiple antenna wireless systems. If only the receivers are allowed to cooperate and the transmitters are constrained to encode their signals independently, then the MIMO system represents a Gaussian multiple-access channel, arising in code-division multiple access (CDMA). If only the transmitters are allowed to cooperate and the receivers are constrained to decode their signals independently, it represents a Gaussian broadcast channel (GBC), arising in the downlink of a wireless system where the base station is equipped with an antenna array. Finally, if neither the transmitters and the receivers are allowed to cooperate, then the MIMO system represents an interference Gaussian channel, arising, for example, in peer-to-peer wireless communication networks.

When a base-station of a MIMO broadcast channel does not have enough antennas for full multiplexing, a precoding matrix is sought for that maps the data streams to the antenna elements of the user that does not apply full multiplexing. All other variables like power allocation and covariance matrices of fully multiplexing users are already completely determined, therefore the weighted sum rate solely depends on the precoder matrix. As the particular choice of the precoding matrix defines the subspace in which the transmitted signals lie, it has also a considerable impact on the achievable rates of the other users. Such coupling effect leads to the fact that the weighted sum rate utility can be formulated as the maximization of a quotient of two Hermitian form determinants where the one in the numerator is raised to an exponent larger than one [23].

A. Learning by optimization of a MIMO-broadcast-related criterion over the space $\mathbb{S}^{\alpha}_{n,p}(\mathbb{C})$

The learning problem discussed in the present section, concerning the maximization of a weighted sum rate in the context of MIMO broadcast channels, may be cast as the maximization of the criterion function $F : \mathbb{S}^{\alpha}_{n,p}(\mathbb{C}) \to \mathbb{R}$ defined as [23]:

$$F(x) \stackrel{\text{def}}{=} \frac{\det^{\beta}(x^H B x)}{\det(x^H A x)}, \tag{16}$$

where $x \in \mathbb{S}^{\alpha}_{n,p}(\mathbb{C})$ denotes a precoding matrix, matrix $A \in \mathbb{C}^{n \times n}$ is Hermitian positive-definite, matrix $B \in \mathbb{C}^{n \times n}$ is Hermitian positive-semidefinite and such that $\operatorname{rank}(B) \geq p$ and the exponent $\beta > 1$. An example of criterion function (16) on the circle $\mathbb{S}^{1}_{2,1}(\mathbb{R})$ is illustrated in the Figure 3.

In order to put into effect the learning scheme described in the section II, it is necessary to compute the gradient $\partial_x F$ of the learning criterion (16). For $K \in \mathbb{C}^{n \times n}$ Hermitian and $x^H K x \in \mathbb{C}^{p \times p}$ nonsingular, it holds that:

$$\partial_x \det(x^H K x) = 2 \det(x^H K x) K x (x^H K x)^{-1}. \tag{17}$$

In fact, according to the definition (6), the gradient $\partial_x \det(x^H K x)$ must satisfy:

$$\det((x+y)^H K (x+y)) - \det(x^H K x) = \Re\mathrm{tr}(y^H \partial_x \det(x^H K x)) + o(\|y\|),$$

where symbol $o(\cdot)$ denotes higher-order infinitesimal. Calcu-

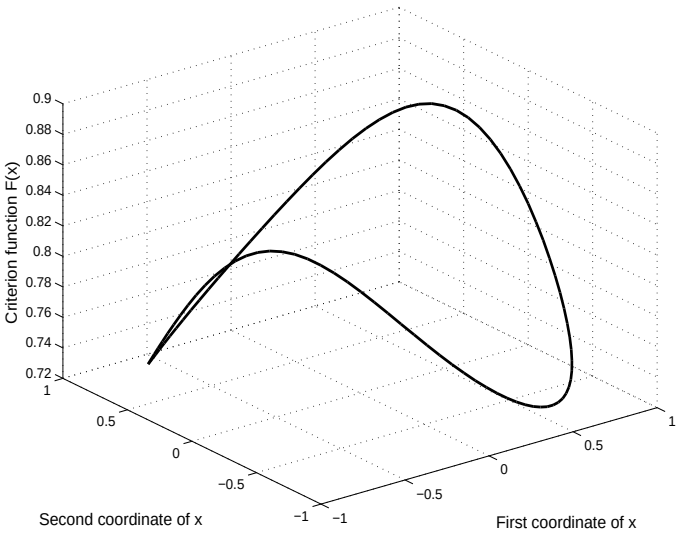

Fig. 3. Example of criterion function (16) on the circle $\mathbb{S}^1_{2,1}(\mathbb{R})$. It appears as a continuous loop and, as it is a regular function defined an a compact domain, it is bounded from above.

lations show that:

$$\begin{aligned}&\det((x+y)^H K(x+y)) = \\ &\det(x^H Kx + y^H Kx + x^H Ky + o(y)) = \\ &\det(x^H Kx(e_p + (x^H Kx)^{-1} y^H Kx + \\ &\quad (x^H Kx)^{-1} x^H Ky + o(y))) = \\ &\det(x^H Kx)\det(e_p + (x^H Kx)^{-1} y^H Kx + \\ &\quad (x^H Kx)^{-1} x^H Ky + o(y)),\end{aligned}$$

where symbol e_p denotes a $p \times p$ identity matrix. For an arbitrary small $y \in \mathbb{C}^{p\times p}$, the following identity holds true:

$$\det(e_p + y) = 1 + \mathrm{tr}(y) + o(\|y\|). \tag{18}$$

Hence:

$$\begin{aligned}&\det((x+y)^H K(x+y)) - \det(x^H Kx) = \\ &\det(x^H Kx)\mathrm{tr}((x^H Kx)^{-1} y^H Kx + \\ &\quad (x^H Kx)^{-1} x^H Ky) + o(\|y\|) = \\ &2\det(x^H Kx)\Re\mathrm{tr}(y^H Kx(x^H Kx)^{-1}) + o(\|y\|).\end{aligned}$$

The last term must equate $\Re\mathrm{tr}(y^H \partial_x \det(x^H Kx))$, up to an high-order infinitesimal, hence (17) follows.

The result (17) holds true at every point $x \in \mathbb{S}^\alpha_{n,p}(\mathbb{C})$ such that $\det(x^H Kx) \neq 0$. It may be extended to the whole matrix space $\mathbb{S}^\alpha_{n,p}(\mathbb{C})$ thanks to the notion of *adjugate* matrix [28]. The adjugate of a matrix $z \in \mathbb{C}^{p\times p}$, denoted as $\mathcal{A}(z)$, satisfies the identities $\mathcal{A}(z)z = z\mathcal{A}(z) = \det(z)e_p$. If $x^H Kx$ is nonsingular, it holds that:

$$(x^H Kx)^{-1}\det(x^H Kx) = \mathcal{A}(x^H Kx). \tag{19}$$

The right-hand side of the above equation is defined for every $x \in \mathbb{S}^\alpha_{n,p}(\mathbb{C})$, hence the gradient (17) may be prolonged to the whole space $\mathbb{S}^\alpha_{n,p}(\mathbb{C})$ and reads:

$$\partial_x \det(x^H Kx) = 2Kx\mathcal{A}(x^H Kx). \tag{20}$$

- Set $k = 0$
- Set x_0 to an initial guess in $\mathbb{S}^\alpha_{n,p}(\mathbb{C})$
- Set ε to desired precision
- Repeat:
  - Set $a_k = \det(x_k^H Ax_k)$ and $b_k = \det(x_k^H Bx_k)$
  - Set $g_k = 2\frac{b_k^{\beta-1}}{a_k^2}[\beta a_k Bx_k \mathcal{A}(x_k^H Bx_k) - b_k Ax_k \mathcal{A}(x_k^H Ax_k)]$
  - Set $v_k = g_k - \frac{x_k}{\alpha}\Re\mathrm{tr}(x_k^H g_k)$
  - If $\|v_k\| < \varepsilon$ then Stop
  - Set $\omega_k = \alpha^{-\frac{1}{2}}\|v_k\|$
  - Set $\Theta_k = 2\pi/\omega_k$ and determine the stepsize h_k according to the method explained in section II-B
  - Set $x_{k+1} = x_k\cos(\omega_k h_k) + v_k \sin(\omega_k h_k)/\omega_k$
  - Set $k = k + 1$

Fig. 4. Pseudocode to implement the proposed procedure to optimize the criterion function (16).

Such result allows computing the gradient of the criterion function (16) that takes on the form:

$$\begin{aligned}\partial_x F(x) = 2\tfrac{\det^{\beta-1}(x^H Bx)}{\det^2(x^H Ax)} \times \\ [\beta\det(x^H Ax)Bx\mathcal{A}(x^H Bx) - \\ \det(x^H Bx)Ax\mathcal{A}(x^H Ax)].\end{aligned} \tag{21}$$

The above expression together with the general expression (7) of the Riemannian gradient on the hypersphere $\mathbb{S}^\alpha_{n,p}(\mathbb{C})$ gives the Riemannian gradient of the criterion function (16).

The proposed procedure to optimize the criterion function (16) may be summarized by the pseudocode listed in the Figure 4, where the quantity g_k denotes the gradient of the learning criterion function (16) while the quantity v_k denotes its Riemannian gradient.

B. Numerical results

To start the iteration (15), an initial guess x_0 may be picked up randomly in $\mathbb{S}^\alpha_{n,p}(\mathbb{C})$. The learning progress may be monitored by computing iteratively the value of the learning criterion function $F(x_k)$. As the range of values of the criterion function may vary considerably, the following performance index may be considered instead:

$$10\log_{10}\left[\frac{F(x_k)}{F(x_0)}\right] \text{ for } k \geq 0. \tag{22}$$

The Figure 5 shows the result of a single run on the manifold $\mathbb{S}^5_{8,5}(\mathbb{C})$. In the test-problem, the matrix A is generated by the rule $A = UPU^H$ with U being the orthogonal projection of a 8×8 random matrix $C + iD$ with C and D having random entries drawn from a normal distribution and P being a 8×8 diagonal matrix whose in-diagonal elements are drawn from a uniform distribution with support $[0,\ \frac{8}{10}]$. Likewise, the matrix B is generated by the rule $B = VRV^H$ with V being randomly generated likewise U and R being a diagonal matrix with at least 5 non-zero in-diagonal entries which are again drawn from a uniform distribution with support $[0,\ \frac{8}{10}]$. The rank of the matrix B, namely, the number of non-zero in-diagonal entries, is randomly selected in the integer-set

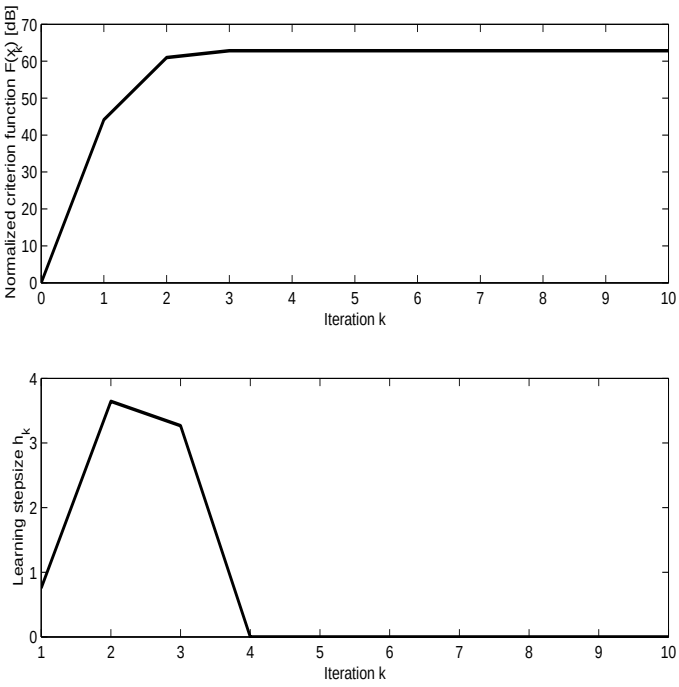

Fig. 5. Result of a run of the algorithm (15) on the manifold $\mathbb{S}^5_{8,5}(\mathbb{C})$.

$\{5, 6, 7, 8\}$. The number of sampling locations to approximate the optimal stepsize was set to $S = 50$. The obtained numerical result shows that the optimization algorithm (15) converges steadily and in a few iterations.

IV. Conclusion

The aim of the present manuscript is to discuss the problem of learning by optimization over the complex-valued matrix-hypersphere $\mathbb{S}^{\alpha}_{n,p}(\mathbb{C})$. Similar optimization problems have been dealt with by other authors for functions defined on spheres of matrices with real entries. Although the investigated optimization problems can be treated with existing methods for real matrices, there are advantages and merit in treating them directly as complex matrices. For example, one may recognize that the formulation is more elegant by avoiding going back and forth from complex to real and the method should become more transparent and accessible to researchers who are not familiar with optimization on curved spaces.

The learning theory to tackle such problem has been formulated in terms of Riemannian-gradient based optimization of a regular learning criterion function. Such learning theory has been implemented through a geodesic-stepping method, which allows moving on a smooth manifold along the direction of the Riemannian gradient of an assigned criterion function. In order to choose a numerically-optimal learning stepsize schedule for the geodesic-stepping-based learning algorithm, a geodesic-search sub-algorithm has been made use of.

The algorithm has been applied to a formal problem related to multiple-input multiple-output broadcast channels. Numerical results illustrate the effectiveness of the proposed learning method and that the algorithm converges steadily and in a few iterations.

References

[1] I. Aizenberg, *Book Reviews: Complex-Valued Neural Networks: Theories and Applications A. Hirose, Ed. (NJ.: World Scientific Publishing Co. Pte. Ltd., 2004)*, IEEE Transactions on Neural Networks, Vol. 17, No. 2, p. 534, March 2006

[2] S.-i. Amari and S. Fiori, *Editorial: Special issue on "Geometrical methods in neural networks and learning"* Neurocomputing, Vol. 67, pp. 1 – 7, August 2005

[3] G. Caire and S. Shamai, *On the achievable throughput of a multi-antenna Gaussian broadcast channel*, IEEE Transactions on Information Theory, Vol. 49, No. 7, pp. 1691 – 1705, July 2003

[4] E. Celledoni and S. Fiori, *Neural learning by geometric integration of reduced 'rigid-body' equations*, Journal of Computational and Applied Mathematics (JCAM), Vol. 172, No. 2, pp. 247 – 269, December 2004

[5] T.-P. Chen, S.-I. Amari and Q. Lin, *A unified algorithm for principal and minor component extraction*, Neural Networks, Vol. 11, No. 3, pp. 365 – 369, 1998

[6] T.-P. Chen and S.-I. Amari, *Unified stabilization approach to principal and minor components extraction algorithms*, Neural Networks, Vol. 14, No. 10, pp. 1377 – 1387, 2001

[7] S. Fiori, *Blind separation of circularly distributed source signals by the neural extended APEX algorithm*, Neurocomputing, Vol. 34, No. 1-4, pp. 239 – 252, August 2000

[8] S. Fiori, *A theory for learning based on rigid bodies dynamics*, IEEE Transactions on Neural Networks, Vol. 13, No. 3, pp. 521 – 531, May 2002

[9] S. Fiori, *A fast fixed-point neural blind deconvolution algorithm*, IEEE Transactions on Neural Networks, Vol. 15, No. 2, pp. 455 – 459, March 2004

[10] S. Fiori, *Non-linear complex-valued extensions of hebbian learning: An essay*, Neural Computation, Vol. 17, No. 4, pp. 779 – 838, 2005

[11] S. Fiori, *Quasi-geodesic neural learning algorithms over the orthogonal group: A tutorial*, Journal of Machine Learning Research, Vol. 6, pp. 743 – 781, May 2005

[12] S. Fiori, *Formulation and integration of learning differential equations on the Stiefel manifold*, IEEE Transactions on Neural Networks, Vol. 16, No. 6, pp. 1697 – 1701, November 2005

[13] S. Fiori, *Geodesic-based and projection-based neural blind deconvolution algorithms*, Signal Processing, Vol. 88, No. 3, pp. 521 – 538, March 2008

[14] S. Fiori, *A study on neural learning on manifold foliations: The case of the Lie group SU(3)*, Neural Computation, Vol. 20, No. 4, pp. 1091 – 1117, April 2008

[15] S. Fiori, *Lie-group-type neural system learning by manifold retractions*, Neural Networks, Vol. 21, No. 10, pp. 1524 – 1529, December 2008

[16] S. Fiori, *On vector averaging over the unit hypersphere*, Digital Signal Processing (Elsevier), Vol. 19, No. 4, pp. 715 – 725, July 2009

[17] S. Fiori and P. Baldassarri, *Approximate joint matrix diagonalization by Riemannian-gradient-based optimization over the unitary group (with application to neural multichannel blind deconvolution)*, in "Neural Computation and Particle Accelerators: Research, Technology and Applications" (ed.s: E. Chabot and H. D'Arras, Series of Neuroscience Research Progress), NOVA Publisher, 2009

[18] S. Fiori, *Learning by natural gradient on noncompact matrix-type pseudo-Riemannian manifolds*, IEEE Transactions on Neural Networks, Vol. 21, No. 5, pp. 841 – 852, May 2010

[19] S. Fiori, *Extended Hamiltonian learning on Riemannian manifolds: Theoretical aspects*, IEEE Transactions on Neural Networks, Vol. 22, No. 5, pp. 687 – 700, May 2011

[20] D. Gabay, *Minimizing a differentiable function over a differentiable manifold*, Journal of Optimization Theory and Applications, Vol. 37, No. 2, pp. 177 – 219, 1982

[21] C. Guthy, W. Utschick, R. Hunger and M. Joham, *Efficient weighted sum rate maximization with linear precoding*, IEEE Transactions on Signal Processing, Vol. 58, No. 4, pp. 2284 – 2297, April 2010

[22] A. Hirose, *Complex-Valued Neural Networks*, Studies in Computational Intelligence, Vol. 32, Springer-Verlag, Berlin Heidelberg, 2006

[23] R. Hunger, P. de Kerret and M. Joham, *An algorithm for maximizing a quotient of two Hermitian form determinants with different exponents*, Proceeding of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 2010, Dallas (TX, USA), March 2010), pp. 3346 – 3349, 2010

[24] M. KOBAYASHI AND G. CAIRE, *An iterative water-filling algorithm for maximum weighted sum-rate of Gaussian MIMO-BC*, IEEE Journal on Selected Areas in Communications, Vol. 24, No. 8, pp. 1640 – 1646, 2006

[25] J. LIU, Y.T. HOU AND H.D. SHERALI, *On the maximum weighted sum-rate of MIMO Gaussian broadcast channels*, Proceedings of the IEEE International Conference on Communications (ICC, Beijing (China), May 2008), pp. 3664 – 3668, 2008

[26] D.G. LUENBERGER, *The gradient projection methods along geodesics*, Management Science, Vol. 18, pp. 620 – 631, 1972

[27] M. SPIVAK, *A Comprehensive Introduction to Differential Geometry*, Volume 1, 2nd Edition, Berkeley, CA: Publish or Perish Press, 1979

[28] G.W. STEWART, *On the adjugate matrix*, Linear Algebra and Its Applications, Vol. 283, No.s 1-3, pp. 151 – 164, November 1998

[29] L. XU, *Theories for unsupervised learning: PCA and its nonlinear extension*, in Proceedings of the International Joint Conference on Neural Networks (IJCNN 1994, Orlando (FL, USA), June 1994), Vol. II, pp. 1252 – 1257, 1994

[30] Y. ZHU, W. MIO AND X. LIU, *Optimal dimension reduction for image retrieval with correlation metrics*, Proceedings of the International Conference on Neural Networks (IJCNN 2009, Atlanta (GA, USA), June 2009), pp. 3565 – 3570, 2009