Relative Uncertainty Learning Theory: An Essay

Simone Fiori

Affiliation:
Facoltà di Ingegneria, Università di Perugia
Polo Didattico e Scientifico del Ternano
Loc. Pentima bassa, 21, I-05100 Terni (Italy)

E-mail: fiori@unipg.it

Pages: 33, **Figures**: 1, **References**: 56

Manuscript accepted for publication on:
International Journal of Neural Systems

September 22, 2004

Relative Uncertainty Learning Theory: An Essay

Simone Fiori

Abstract

The aim of this manuscript is to present a detailed analysis of the algebraic and geometric properties of relative uncertainty theory (RUT) applied to neural networks learning. Through the algebraic analysis of the original learning criterion, it is shown that RUT gives rise to principal-subspace-analysis-type learning equations. Through an algebraic-geometric analysis, the behavior of such matrix-type learning equations is illustrated, with particular emphasis to the existence of certain invariant manifolds.

Keywords: Relative uncertainty; Learning theory; Unsupervised neural learning; Principal subspace analysis; Stiefel manifold.

1 Introduction

In the neural network field, many learning problems can be re-formulated as optimizations ones, i.e. often a learning-strategy design work consists in choosing a suitable loss or objective function, which measures the adequateness of network parameters values, and in establishing a set of proper constraints that take into account the inherent physical restrictions to the possible optimal solutions. In fact, some learning problems possess a particular structure such that it is *a priori* known that the associated optimal parameters values must fulfill mutual constraints, as it is the case for limited-resource-based learning theories, for instance.

The aim of this paper is to study the behavior of a particular set of first-order matrix-type differential equations, which arise by the gradient-based optimization of a class of criterion functions originating from the relative uncertainty theory, as proposed by Xu in [50].

The relative uncertainty theory is closely tied to a widely investigated problem, namely *principal subspace analysis*. Subspace analysis methods play an important role in high-dimensional data handling, such as in computer vision research, for instance. In visual modeling and recognition, the principal subspaces are extracted and utilized for description, detection, and classification. They are also used in parametric descriptions of shapes, target detection, visual learning, face recognition and linear discriminant analysis [38, 54]. Subspace analysis often significantly simplifies tasks such as regression, classification, and density estimation by computing low-dimensional uncorrelated subspaces [12].

Learning with constraints may be achieved through different ways. The most widely known way relies on the augmentation of the original cost/objective function through suitable functions that take into account the necessary constraints,

which are formally expressed as equality or inequality restrictions on the free variables, as recalled with details in the next section. However, by analyzing the behavior of the relative uncertainty learning equations, we observe that there exist learning systems that do not require to add constraints to the original learning criterion function, since they inherently maintain the constraints fulfilled at any time. As already pointed out in [16], there are essentially two classes of learning rules that do not break the invariants: The class of Jacobian-based rules, which contains all the gradient-based learning rules in which the gradient is defined as the plain Euclidean gradient (or Jacobian) of the criterion function and still they fulfill the orthogonality constraints, and the class of the learning rules in which the adaptation essentially arises as a matrix-flow of some Riemannian-gradient differential learning equation. The class relative uncertainty learning equation turns out to be an excellent example of learning rule belonging to the former category.

The paper is organized as follows. In Section 2, a general discussion is presented about the most commonly known methods for learning with constraints. In Section 3, some specific definitions are recalled from the scientific literature and an insight on the relative uncertainty theory is given. Section 4 is devoted to the formal presentation of two lemmas that will help us when we shall prove the principal results about the relative uncertainty optimization theory in Section 5. The Section 6 is devoted to the study of the geometry of relative-uncertainty system and to the illustration of several properties of it, like stability, convergence and steady-state behavior. In Section 7, some complementary comments and a discussion on possible future investigations are presented. Section 8 concludes the paper.

2 Generalities on learning via constrained optimization

Let a neural system be described by a set of tunable parameters arranged in a vector w. A widely known procedure allowing to find the optimal network parameter vector $w^\star$ consists in:

1. Defining a proper loss function and a set of mathematical constraints on the values of w.

2. Constructing an augmented optimization criterion that takes into account the learning goal and the necessary constraints.

3. Selecting a suitable searching method that allows to find the optimal solution.

Usually a loss measure takes on the form of a function $C(w)$ of the design vector w and the constraints are expressed through equality restrictions $\alpha_i(w) = 0$ and inequality restrictions $\beta_i(w) \geq 0$. The function $C(\cdot)$ should be convex at least around the optimal solution to ensure the convergence of the searching algorithm, but usually it is not convex over the whole search-space and this circumstance makes it difficult to prove the global convergence of the learning procedure.

A way to construct an augmented optimization criterion on the basis of the above information is to define the following function:

$$\Gamma(w, \lambda; \kappa_1, \kappa_2, \kappa_3) := C(w) + \Lambda(w, \lambda) + \Pi(w; \kappa_1) + B(w; \kappa_2) + R(w; \kappa_3) \ ,$$

where Λ, Π, B and R are scalar functions of the vector of network parameters and the κ_i are vectors of internal parameters of the above criterion functions, which are often used to weight the different terms in the augmented criterion Γ. One or more of the functions Λ, Π, B and R may lack, and this may be simply formalized by setting their relative weight equal to zero.

If only the function Λ is present, the criterion $\Gamma(w, \lambda)$ is termed the *ordinary Lagrange function* and the parameters in the vector λ are termed *Lagrange multipliers*. The function Λ can take into account both equality and inequality constraints; in addition to the optimal value $w^\star$ an optimal multipliers' vector $\lambda^\star$ has to be learned from the available network training data. In order to analyze the existence of global minima of Γ depending on the nature of function C and of functions α_i and β_i, a useful tool is provided by the standard Kuhn-Tucker theory [20, 41].

The functions $\Pi(w; \kappa_1)$ and $B(w; \kappa_2)$ stand, respectively, for exterior penalty function (or simply *penalty*) and interior penalty function (simply *barrier*). Penalty functions [11] can deal both with equality and inequality constraints. They take on high values when the restrictions are violated and low (typically zero) values when the restrictions are fulfilled. The parameters in the vector κ_1 should be carefully chosen to ensure a correct behavior of the augmented function Γ. Barrier functions [20, 45] take on small values when the constraints are respected, while they take on large values near the separation surface between the allowed and non-allowed solutions. Indeed, the barrier methods may deal only with inequality restrictions which are fulfilled by the so-termed *feasible points*.

When penalty and ordinary Lagrange method are used together, i.e. penalty and Lagrange functions are linearly mixed, the resulting approach is referred to as the *augmented Lagrange method* [4, 24, 44].

As a useful improvement of the above approaches, another term may be added to the augmented optimization function in order to enhance the numerical conditioning of the problem. It is the *regularization term* $R(w; \kappa_3)$ which comes from the fruitful regularization theory by Tikhonov [36, 48, 49].

About the cost function $C(w)$, often it is computed on the basis of an error term that represents the distance among the actual network's output and the corresponding learning target. In these cases, the function $C(w)$ is expressed by means of functions of the error term as the L_p–norm as well as Huber's or Talwar's non-linear functions [5, 22, 26, 40]. Each of these may imply different outcomes of the learning task.

As an example of learning with constraints, in unsupervised learning theory for multilayer-perceptron-like networks formed by the interconnection of basic neurons, learning the optimal set of connection patterns may be interpreted as selecting the best directions in the space that the weight-vectors belong to. If a learning error criterion is defined over the weight-space so that it measures how much interesting the different directions are, and if the criterion is designed to be unbounded, its optimization must be performed under some constraint, and the weight-space reduces to the subset of directions satisfying such mutual restrictions.

Some further highly-involved examples of learning with restrictions may be found e.g. in text classification via the information geometry of multinomial simplex [31] and in multichannel blind deconvolution via exploitation of the geometrical structure of the FIR-filter-banks manifolds [55, 56].

After constructing a suitable optimization criterion, a proper method that allows to look for its minimum should be chosen. Several procedures have been developed and continuously improved through years. It is worth citing here the ordinary gradient-based searching method, the most popular first-order procedure, but also the Newton method and its improvements, as the Levemberg-Marquard one [41], the Broyden-Fletcher-Goldfarb-Shanno technique [20] and the conjugate gradient method [20]. As it is well-known, the use of one of them instead of another generally may greatly affect the speed of learning and the accuracy of the solution $w^\star$ found.

It might be interesting to observe that, in the previous discussion, only vector-type optimization problems and methods have been considered. However, in some technical areas as in the artificial neural network context [3, 23], matrix optimization problems arise, as well. It is known from the scientific literature that a class of learning systems for artificial neural networks may be formulated in terms of matrix-type differential equations of network's learnable parameters. Not infrequently, such differential equations are defined over parameter spaces endowed with a specific geometry (such as the general linear group, the compact Stiefel manifold, the Grassman manifold or the orthogonal group [2, 16, 18]). Also, from an abstract viewpoint, the differential equations describing the internal dynamics of neural systems may be studied through the instruments of complex dynamical systems analysis.

Although a matrix problem may be always re-formulated as a vector problem, from a theoretical point of view this procedure may cause the loss of important information on the learning task to solve, which might result in a less powerful learning method. Conversely, although the search for an optimal network connection pattern can be performed by using constrained optimization, the exploitation of the intrinsic geometry of network's parameters space may lead to efficient and versatile algorithms (see, e.g., [33] for an example of mixed gradient/MCMC algorithm for optimization over the Grassman manifold). Also, a detailed study on the adequateness of the Lagrange-multiplier-type learning has been recently presented in [13].

From a more general perspective, differential geometry provides the mathematical tool for studying e.g. physical or engineering problems where it is worthwhile taking into account the intrinsic geometry of the variables space. Perhaps, the best known example is given by the general theory of relativity, where the effect of gravitation is completely described by the effect produced on the geometry of the four-dimensional space [15].

In this work we consider the special problem of finding the minimum (or maximum) of a cost function of a network connection-matrix W subject to the constraint of orthonormality of W. In other words, we analyze the solution of an *orthonormal problem*. Orthonormal problems arise in several contexts, as is the case in eigenvalue and generalized eigenvalue problems, joint diagonalization and optimal linear compression [10, 25, 51], numerical simulation of the physics of bulk materials [14], linear programming and sequential quadratic programming [6, 14], minimal linear system realization from noise-injection measured data and invariant subspace computation [14, 34], steering of antennas arrays [1], analysis of neural activity for natural three-dimensional movement [53], electrical networks fault detection [32], adaptive image coding [37], signal denoising via sparse coding [39], dynamic texture recognition [46], unsupervised learning for blind signal processing [7, 16, 18, 29, 30, 35, 43, 52] and images modeling/representation [33, 47].

3 General definitions and problem formulation

In this section, some specifications are given concerning the notation used in the whole manuscript and the statistics of the signals involved in the subsequent analysis. Also, formal definitions are given about the already-mentioned concepts of principal subspace and principal subspace analysis.

In this work the following notation and conventions are used:

- The symbol $Im(\cdot)$ denotes the matrix operator *range*, which returns the range-space (i.e. the space spanned by the column vectors) of the matrix contained within, $det(\cdot)$ denotes the determinant of the matrix contained

within, the operator $rk[\cdot]$ returns the value of the rank of the matrix contained within. The symbol I_p denotes a $p \times p$ identity matrix while symbol $I_{p,q}$ denotes a $p \times q$ pseudo-identity matrix. The notation $\mathcal{M}(p,q)$ is used to denote the space of the $p \times q$ matrices endowed with real and bounded entries and such that $p \geq q$. The superscript ^{T} denotes the usual matrix transposition.

- For the sake of notation conciseness, we find it convenient to define the matrix set $\mathbf{\Omega}_r(p,q)$ of the *r–orthonormal $p \times q$ matrices*, i.e. that subset of $\mathcal{M}(p,q)$ defined as:

$$\mathbf{\Omega}_r(p,q) := \{H \in \mathcal{M}(p,q) : H^T H = r^2 I_q\}, \tag{1}$$

  By definition, r must be non-null. When we shall not be interested about the value of r, we shall use the "don't care" notation, namely, we shall indicate that subset with $\mathbf{\Omega}(p,m)$ and we shall simply refer to its elements as *pseudo-orthonormal matrices*. Conversely, in the case in which r is constant and equal to 1, the resulting set-structure is known in the literature as *compact Stiefel manifold*, whose members are termed orthonormal matrices.

- The operator $E[\cdot]$ returns the statistical average (expectation) of its argument. As a further convention, through the paper we consider only wide-sense stationary random processes endowed with a probability density function. Stationarity is required because the artificial neural systems take time to learn the statistical features of the input signals, therefore these should not change during the learning process (of course, variability of the statistical features of the signals is admissible as long as it is slow compared to the learning adaptation process). As a further convention, in this work we are interested in real-valued random processes, with positive-definite covariance matrices.

Next, the formal definition of the concept of principal subspace analysis is recalled from the literature and the basic idea behind maximum relative uncertainty theory for achieving principal subspace analysis is also sketched from [50].

Formally, a generic principal subspace of a real-valued random process is defined in the following way:

Definition 1 (Principal subspace of order q of a random process.) *Let x be a zero-mean random process in $\mathcal{R}^p$ endowed with a finite covariance matrix $\Sigma_x := E[xx^T]$ with p distinct positive eigenvalues and let $q \leq p$ be a positive integer number arbitrarily fixed. Let us also denote with f_i an eigenvector of the matrix Σ_x and with λ_i the corresponding eigenvalue. The eigenpairs (f_i, λ_i) are supposed to be ordered so that $\lambda_1 > \lambda_2 > \ldots > \lambda_p$.*

It is defined principal subspace of order q of the process x *the space* $\mathcal{P}_q(x)$ *spanned by the first* q *eigenvectors of the covariance matrix of the process* x*. In symbols:* $\mathcal{P}_q(x) := span\{f_1, f_2, \ldots, f_q\}$.

It is fundamental to note that the above definition involves only a second-order statistical feature of the analyzed random process: Its covariance matrix. Also, if, as a basis for $\mathcal{P}_q(x)$, we choose the vectors $f_1, f_2, \ldots, f_q$, then the projections of the random process on the basis results in the principal components of the random process.

Given the preceding fundamental definition, we can define the principal subspace analysis (PSA) of order q of a random process in the following way:

Definition 2 (PSA of order q of a random process.) *Let* x *be a zero-mean random process in* $\mathcal{R}^p$ *with a finite covariance matrix with* p *distinct positive eigenvalues and let* $q \leq p$ *be a positive integer number arbitrarily fixed. Let* $\mathcal{P}_q(x)$ *be a principal subspace of order* q *of the random process* x*.*

It is defined PSA *of order* q *of the random process* x *a generic matrix* B *in* $\mathcal{M}(p, q)$ *such that* $Im(B) = \mathcal{P}_q(x)$*. If, moreover, matrix* B *belongs to the space* $\mathbf{\Omega}_1(p, q)$ *it is said an* orthonormal PSA *and is denoted with the symbol* PSA$^{\perp}$*.*

If B is a PSA (or, as a special case, a PSA$^{\perp}$) of the random process x, the new random process $B^T x$ is the *PSA transformed process*, whose main characteristic is that it contains the main fraction of the power of the process x compatibly with the reduction in dimensionality chosen. For more detail about this aspect of the theory, see the abundant related field's literature [23, 27, 42, 50].

One of the diverse theoretical foundations of PSA is the maximum relative uncertainty theory (RUT) [50], which is based on the following considerations.

A random process x takes on values relative to its determinations with an uncertainty whose degree depends on its probability density function (p.d.f.). For a Gaussian random process x with covariance Σ_x, for instance, as a proper measure of such an uncertainty or, equivalently, of the information carried on by x, the following index may be used [50]:

$$J_X^{\varphi}(x) := \varphi[det(\Sigma_x)] \ . \tag{2}$$

It is worth noting that, in the above expression, if the hypotheses of **Definition 1** hold, the quantity $det(\Sigma_x)$ is always positive. In fact, by eigenvalue decomposition we known that there exist $\Phi \in \mathbf{\Omega}_1(p, p)$ and a diagonal matrix Λ of the eigenvalues of the covariance matrix Σ_x, such that $\Sigma_x = \Phi\Lambda\Phi^T$. Now, it is easily seen that $det(\Sigma_x) = det(\Phi\Lambda\Phi^T) = det(\Lambda) > 0$, being $det(\Lambda)$ the product of all positive eigenvalues. Henceforth, in the expression (2), the function $\varphi(u) > 0$ may be assumed differentiable and monotonically increasing for $u > 0$.

The uncertainty associated with x reflects the complexity of its p.d.f., or the richness of information contained in the probability density function. Notice

that if $\varphi(z) := \frac{1}{2}ln(z)+h$, with h being a properly chosen constant, the measure $J_X^{\varphi}(x)$ coincides with the well-known *Shannon differential entropy* [28] of the process x. That measure also describes the dispersion of the values of the process: The greater the measure $J_X^{\varphi}(\cdot)$'s value, the larger the range of the dispersion, the greater the uncertainty associated to the process.

It is well-known [28] that for a zero-mean Gaussian process, the second order statistics entirely describes its property, but this is no longer true for a non-Gaussian process. Despite this, by extension, even when x is not Gaussian, it is still possible to use the same Xu's measure (2) in order to describe its uncertainty: As a consequence of such an approach there is a loss of information, but for principal subspace analysis of the process that, as emerges from the **Definition 2**, is a second order task, still the optimization criterion is valid.

Now, let x_1 and x_2 be distinct random processes with different p.d.f.s. The function:

$$\rho(x_1, x_2) := \frac{J_X^{\varphi}(x_1)}{J_X^{\psi}(x_2)} \ , \tag{3}$$

measures the uncertainty associated with x_1 relatively to x_2, if $\psi(u) > 0$ is differentiable and monotonically increasing for $u > 0$.

Since our target is to determine a neural linear transformation of the process $x \in \mathcal{R}^p$ into a new process $y := W^T x \in \mathcal{R}^q$ with less components than x, such that y retains the greatest possible quantity of uncertainty, we can use a system that learns the matrix W such that in a fixed number of variables it is concentrated the major fraction of the uncertainty contained in the x's entries, compatibly with the fixed reduction in dimensionality. In other words, we can attempt to *maximize the uncertainty of the random process 'y'*.

Such an approach would clearly lead to an ill-posed problem. Indeed, if some consistent constraints about the values reachable by the entries of W are not imposed, a method as that just proposed would lead to the divergence of matrix W's entries. Therefore, a restriction criterion is necessary. Following Xu, it is possible to utilize the uncertainty about the values of the process η, produced by the same neural transformation, derived from an input reference process $\xi \in \mathcal{R}^p$, for instance drawn from a Gaussian distribution $N(0, I_p)$. A graphical representation of such dual network system is given in the Figure 1. Since such a reference process already contains the maximum quantity of uncertainty, when the uncertainty of the random process $\eta := W^T \xi \in \mathcal{R}^q$ grows it is only because the entries of W grow in value. By consequence, the quantity $J_X^{\psi}(\eta)$ measures that part of uncertainty introduced in η by the variations of the entries in W, while the function $\rho(y, \eta)$ defined by equation (3) measures the actual variation of the uncertainty relative to the process y due to a true concentration of information. Thus, following Xu, the objective function to be maximized

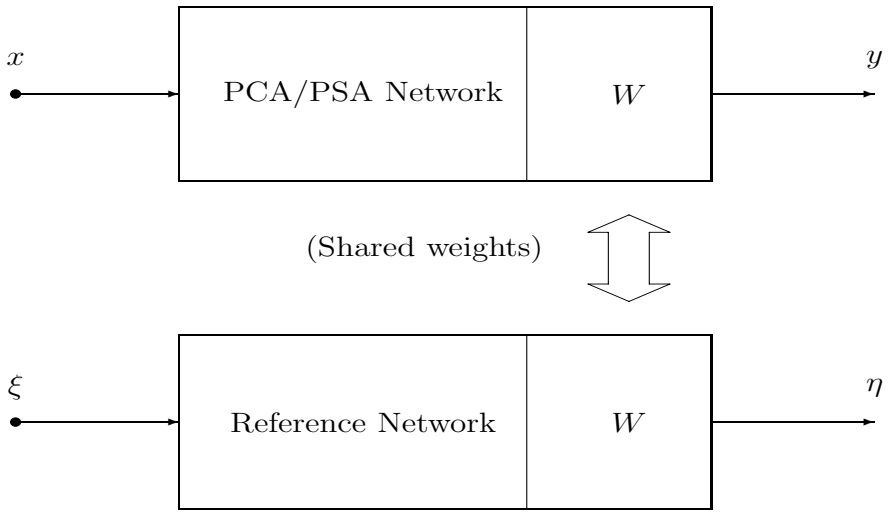

Figure 1: A representation of the dual network system that the MUT theory relies on.

with respect to the neural network connection matrix W is:

$$\rho(y,\eta) = \frac{\varphi\left[det\left(\int_{\mathcal{R}^q} \bar{y}\bar{y}^T p_y(\bar{y};W)d\bar{y}\right)\right]}{\psi\left[det\left(\int_{\mathcal{R}^q} \bar{\eta}\bar{\eta}^T p_\eta(\bar{\eta};W)d\bar{\eta}\right)\right]} \ , \tag{4}$$

where $p_y(\cdot;W)$ and $p_\eta(\cdot;W)$ are the p.d.f.s of the random processes y and η, respectively, which depend on network's configuration. Note that with the above conventions, the signals' covariance matrices are $\Sigma_\eta = W^T W$ and $\Sigma_y = W^T \Sigma_x W$. The above description explains why such learning principle is referred to as maximum relative uncertainty.

Since our purpose is to determine a solution to a criterion-maximization problem, we can adopt the GSA (gradient steepest ascent) method, obtaining the adaptation system:

$$\frac{dW}{dt} = \gamma \frac{\partial \rho}{\partial W} \ , \tag{5}$$

with γ being a positive and constant adapting rate.

From a practical viewpoint, the above differential equation should be implemented on a computer, so a discretization method in the time-domain that allows converting it into a discrete-time algorithms should be carefully developed, in order to retain (up to reasonable precision) the geometric properties that characterize the developed learning rule (that is, that allows preserving the constraints). This is a research field by itself and falls outside the scope of the present manuscript. Interested Readers might find some recent results on this topic in the contributions [7, 17].

4 Two useful lemmas

Under some conditions about the distribution of the input signal x, the learning system (5) is of PSA-type. Before we are proving the previous statement, we need to formally state two preliminary results.

The first result concerns an application defined on abstract matrix spaces, and with real values, based on the determinant operator.

Lemma 3 (Maximum–determinant lemma). *Let A be a fixed matrix in $\mathcal{M}(p,p)$ symmetric and endowed with p distinct and positive eigenvalues. The function:*

$$d_A : \mathbf{\Omega}_1(p,q) \mapsto \mathcal{R} \ , \ d_A(\Phi) := det(\Phi^T A \Phi) \ , \tag{6}$$

enjoys the following properties:

1. *It is invariant with respect to the orthonormal right-multiplications of the argument.*
2. *It is positive for every choice of the argument, namely $d_A(\Phi) > 0$ for all $\Phi \in \mathbf{\Omega}_1(p,q)$.*
3. *It reaches its maximum value in $\Phi = \tilde{\Phi}$ (up to arbitrary rotations) where the columns of $\tilde{\Phi}$ coincide with the eigenvectors, normalized to one, corresponding to the q largest eigenvalues of A. If such eigenvalues are denoted with λ_i (for $i = 1, 2, \ldots, q$), the maximum value of the function can be written as $d_A(\tilde{\Phi}) = \prod_{i=1}^{q} \lambda_i$.*

Proof

Let us denote with f_i the i^{th} eigenvector of A, normalized to one, corresponding to the eigenvalue λ_i. Let the eigenvalues be ordered in a descending manner, i.e. so that $\lambda_1 > \lambda_2 > \ldots > \lambda_p$. By these conventions the matrix $\tilde{\Phi}$ defined in the claim 3) has the vector f_i as the i^{th} column, namely:

$$\tilde{\Phi} := [f_1 \ f_2 \ f_3 \ \ldots \ f_{q-1} \ f_q] \ . \tag{7}$$

We also define the following matrices, which are instrumental in the proof:

$$Q := [f_{i_1} \ f_{i_2} \ f_{i_3} \ \ldots \ f_{i_{q-1}} \ f_{i_q}] \ , \tag{8}$$

$$F := [f_1 \ f_2 \ f_3 \ \ldots \ f_{p-1} \ f_p] \ , \tag{9}$$

where the set $\{i_k\}$ denotes an arbitrary choice of distinct integer indices within $\{1, \ldots, p\}$. In particular, the matrix Q contains a subset of q distinct eigenvectors of the matrix A, while matrix F contains all the eigenvectors.

The first claim can be proven by showing that anyhow an orthonormal matrix K is chosen in $\mathbf{\Omega}_1(q,q)$ it results that $d_A(\Phi K) = d_A(\Phi)$. In fact due to the properties of determinants and of the orthonormal square matrices, the equality

chain $d_A(\Phi K) = det^2(K) \cdot d_A(\Phi) = d_A(\Phi)$ hold true, therefore $d_A(\cdot)$ is invariant with respect to orthonormal right-transformations.

The second claim may be proven by recalling that the determinant of a positive definite matrix is always positive [21]. Therefore, it suffices to prove that the matrix $\Phi^T A \Phi$ is positive definite. To this aim, let us note that the matrix A has, by hypothesis, all positive eigenvalues, and this implies it is positive definite, namely, for all vectors $\bar{\xi} \in \mathcal{R}^p$ is holds $\bar{\xi}^T A \bar{\xi} > 0$. Let us now consider an arbitrary vector $\bar{\eta} \in \mathcal{R}^q$ and let us form the product $\bar{\eta}^T \Phi^T A \Phi \bar{\eta}$. By the variable change $\xi := \Phi \bar{\eta}$ we readily obtain that:

$$(\Phi \bar{\eta})^T A (\Phi \bar{\eta}) = \bar{\xi}^T A \bar{\xi} > 0 \ .$$

As the above inequality is true for any arbitrary vector $\bar{\eta} \in \mathcal{R}^q$ and for all the matrices $\Phi \in \mathbf{\Omega}_1(p, q)$, we may conclude that the product matrix $\Phi^T A \Phi$ is positive definite and hence that $d_A(\Phi) > 0$ always.

The third claim may be proven by the following argument. Since we are searching for the maxima of the function $d_A(\Phi)$ under the constraint that $\Phi \in \mathbf{\Omega}_1(p, q)$, we can utilize the Lagrange's method, i.e. we can search for the free maxima of the Lagrangian function:

$$\mu(\Phi) := det(\Phi^T A \Phi) - \frac{1}{2} tr((\Phi^T \Phi - I_q) H) \ , \tag{10}$$

where H is a symmetric matrix in $\mathcal{M}(q, q)$. The free extremes of this function coincide to the points Φ where the gradient $\frac{\partial \mu(\Phi)}{\partial \Phi}$ zeros, and within this set of solutions we need to select those matrices that maximize $d_A(\Phi)$. The vanishing-gradient equation, particularized to the present case, has the following expression:

$$d_A(\Phi) A \Phi (\Phi^T A \Phi)^{-1} = \Phi H \ . \tag{11}$$

By pre-multiplying both members of the above equation by Φ transposed, remembering that $\Phi^T \Phi = I_q$, it is immediately found that $H = d_A(\Phi) I_q$. Therefore, equation (11) becomes:

$$A \Phi = \Phi (\Phi^T A \Phi) \ . \tag{12}$$

This is an *auto-subspace equation* for A, that is, each of the q columns of Φ must be a linear combination of the same q eigenvectors of A, which is a rotation. This assertion may be proven by considering the value $\Phi = QK$, where K denotes again an arbitrary rotation in $\mathbf{\Omega}_1(q, q)$. In order to verify that this is a solution of equation (12), it is sufficient to replace it within the equation, which gives:

$$AQK = QKK^T Q^T AQK = QQ^T AQK \ . \tag{13}$$

Now, as Q contains a set of q eigenvectors of matrix A, the matrix $\Delta := Q^T AQ$ is diagonal and contains the corresponding eigenvalues of A. Therefore, the

leftmost and the rightmost members of the above equivalence chain read:

$$AQK = Q\Delta K \;\text{ and } QQ^TAQK = Q(Q^TQ)\Delta K = Q\Delta K \;,$$

which shows the equation chain (13) is an identity. In order to prove that the only solutions of equation (12) are given by the combinations QK, we must consider the most general solution $\Phi = FRK$, where $R \in \mathbf{\Omega}_1(p,q)$ is arbitrary, that takes into account a linear combination of all eigenvectors and is such that $\Phi \in \mathbf{\Omega}_1(p,q)$. As we already know that the last rotation K cancels out, we may safely assume $K = I_q$ without loss of generality. By substituting the quantity FR to Φ in equation (12) and by defining $\tilde{\Delta} := F^TAF$, we get:

$$AFR = FR(R^TF^TAFR) \;\Rightarrow\; F\tilde{\Delta}R = F(RR^T)\tilde{\Delta}R \;.$$

The above equation may be satisfied only if $R = \Pi S I_{p,q}$, where $\Pi \in \mathcal{M}(p,p)$ is an arbitrary permutation matrix and $S \in \mathcal{M}(p,p)$ is a diagonal matrix with entries in $\{-1,+1\}$. In fact, by plugging this solution into the above equation yields:

$$F\tilde{\Delta}\Pi S I_{p,q} = F\Pi S I_{p,q} I_{q,p} S\Pi^T\tilde{\Delta}\Pi S I_{p,q} \;. \tag{14}$$

Thanks to the properties of permutation, diagonal and ± 1-valued diagonal matrices, the following equalities hold true: $\Pi^T\tilde{\Delta}\Pi = \tilde{\Delta}$, $S\tilde{\Delta}S = \tilde{\Delta}$, $I_{p,q}I_{q,p}\tilde{\Delta}I_{p,q} = \tilde{\Delta}I_{p,q}$ and $\Pi S\tilde{\Delta} = \tilde{\Delta}\Pi S$, therefore the expression (14) is an identity. In this way, $p-q$ eigenvectors are cancelled out from the product FR. The remaining eigenvectors may be picked arbitrarily and their sign may be arbitrarily switched.

It is now necessary to prove that any solution of the form $\Phi = Q$ is a saddle point for the function $d_A(\cdot)$ except for $\Phi = \tilde{\Phi}$, but for arbitrary rotation. (Note that permutation and sign switch are encompassed by rotation).

To this aim, let us consider arbitrary curves $\Phi = \Phi(t)$ in $\mathbf{\Omega}_1$ passing by the points of interest in $t = 0$, namely $\Phi(0) = Q$: It is then possible to investigate on the behavior of the function $d_A(\Phi(t))$ in a neighbor of Q by computing the second derivative of $d_A(\Phi(t))$ with respect to the parameter t and by evaluating its sign. If the sign varies depending on the particular curve, then the point of interest is a saddle point. If the sign of the second derivative is negative irrespective of the considered curve, then the point of interest is a point of maximum.

For the first- and second-order derivatives of the function $d_A(\Phi(t))$ with respect to t, we have:

$$\begin{aligned}
\frac{d}{dt}det(\Phi^TA\Phi) &= det(\Phi^TA\Phi)tr\left[(\Phi^TA\Phi)^{-1}\frac{d}{dt}(\Phi^TA\Phi)\right] \;, \qquad (15)\\
\frac{d^2}{dt^2}det(\Phi^TA\Phi) &= \frac{d}{dt}[det(\Phi^TA\Phi)]tr\left[(\Phi^TA\Phi)^{-1}\frac{d}{dt}(\Phi^TA\Phi)\right] +
\end{aligned}$$

$$
\begin{aligned}
& det(\Phi^T A\Phi) tr\left[\frac{d}{dt}\left((\Phi^T A\Phi)^{-1}\frac{d}{dt}(\Phi^T A\Phi)\right)\right] \ , \\
= \ & det(\Phi^T A\Phi)\left\{tr^2\left[(\Phi^T A\Phi)^{-1}\frac{d(\Phi^T A\Phi)}{dt}\right] + \right. \\
& tr\left[\frac{d(\Phi^T A\Phi)^{-1}}{dt}\frac{d(\Phi^T A\Phi)}{dt}\right] + \\
& \left. tr\left[(\Phi^T A\Phi)^{-1}\frac{d^2(\Phi^T A\Phi)}{dt^2}\right]\right\} \ . \qquad (16)
\end{aligned}
$$

Some useful definitions and formulas for the following computations are:

$$
\begin{aligned}
& \frac{dX^{-1}}{dt} = -X^{-1}\frac{dX}{dt}X^{-1} \ , \text{ for an invertible matrix flow } X(t) \in \mathcal{M}(q,q) \ , \\
& V := \frac{d\Phi}{dt} \ , \ C := \frac{d^2\Phi}{dt^2} \ , \\
& \frac{d}{dt}(\Phi^T A\Phi) = V^T A\Phi + \Phi^T AV \ , \\
& \frac{d^2}{dt^2}(\Phi^T A\Phi) = C^T A\Phi + 2V^T AV + \Phi^T AC \ .
\end{aligned}
$$

Also, it is necessary to take into account the restrictions that the matrix variable Φ and its first- and second-order derivatives are subjected to, namely:

$$
\begin{aligned}
& \Phi^T\Phi = I_q \ , \text{ which, derived hand-by-hand, gives:} \\
& V^T\Phi + \Phi^T V = 0 \ , \text{ which, derived hand-by-hand, gives:} \\
& C^T\Phi + \Phi^T C = -2V^T V \ .
\end{aligned}
$$

Let us now evaluate the terms of the expression (16) for $t = 0$, namely when $\Phi = Q$. It is just the case to note that the first-order derivative of $d_A(\Phi(t))$ with respect to t vanishes, according to the fact that the points $\Phi = Q$ are extremes of the function $d_A(\Phi)$. The involved terms assume the expressions:

$$
\begin{aligned}
& tr^2\left[(\Phi^T A\Phi)^{-1}\frac{d(\Phi^T A\Phi)}{dt}\right]\Big|_{t=0} = \\
& \quad tr^2\left[\Delta^{-1}(V^T AQ + Q^T AV)\right] = tr^2\left[\Delta^{-1}(V^T Q\Delta + \Delta Q^T V)\right] = \\
& \quad tr^2\left[V^T Q + Q^T V\right] \ , \qquad (17) \\
& - tr\left[(\Phi^T A\Phi)^{-1}\frac{d(\Phi^T A\Phi)}{dt}(\Phi^T A\Phi)^{-1}\frac{d(\Phi^T A\Phi)}{dt}\right]\Big|_{t=0} = \\
& \quad -tr[(\Delta^{-1}(V^T AQ + Q^T AV))^2] = \\
& \quad -2tr[V^T QV^T Q + \Delta^{-1}(V^T Q\Delta Q^T V)] \ , \qquad (18) \\
& tr\left[(\Phi^T A\Phi)^{-1}\frac{d^2(\Phi^T A\Phi)}{dt^2}\right]\Big|_{t=0} = \\
& \quad tr[\Delta^{-1}(C^T AQ + 2V^T AV + Q^T AC)] = \\
& \quad tr[\Delta^{-1}(C^T Q\Delta + 2V^T AV + \Delta Q^T C)] = \\
& \quad tr[C^T Q + Q^T C + 2\Delta^{-1}V^T AV] = 2tr[\Delta^{-1}V^T AV - V^T V] \ . \qquad (19)
\end{aligned}
$$

It is worth noting that the only variable in the preceding expressions is the matrix $V \in \mathcal{M}(p,q)$, that geometrically represents the tangent to the curve $\Phi(t)$ in the point $t = 0$. The velocity matrix V is clearly allowed to point toward any tangent direction to the manifold $\mathbf{\Omega}_1$ in Q with any magnitude.

Now, the term (17) is identically null. The terms (18) and (19) simplify considerably if the following linear-algebra identities are employed:

$$A = Q\Delta Q^T + Q_c \Delta_c Q_c^T \ , \ I_p = QQ^T + Q_c Q_c^T \ .$$

In the above identities, the matrix Q_c contains the eigenvectors of the matrix A that are not in Q and $\Delta_c := Q_c^T A Q_c$. By noting that $V^T V = V^T I_p V^T$ and by plugging the above identities into equation (19) and by slightly rearranging terms in the equation (18), we get:

$$\begin{aligned} -tr[V^T Q V^T Q + \Delta^{-1}(V^T Q \Delta Q^T V)] &= tr[V^T Q Q^T V] - \\ & \quad tr[\Delta^{-1}(V^T Q \Delta Q^T V)] \ , \\ tr[\Delta^{-1} V^T A V - V^T V] &= -tr[V^T Q Q^T V] - tr[V^T Q_c Q_c^T V] + \\ & \quad tr[\Delta^{-1} V^T Q \Delta Q^T V] + \\ & \quad tr[\Delta^{-1} V^T Q_c \Delta_c Q_c^T V] \ . \end{aligned}$$

By gathering the above partial results, we come up with the final expression for the second-order derivative of the function $d_A(\cdot)$:

$$\left.\frac{d^2}{dt^2} det(\Phi^T A \Phi)\right|_{t=0} = 2det(\Delta) tr[\Delta^{-1} V^T Q_c \Delta_c Q_c^T V - V^T Q_c Q_c^T V] \ . \qquad (20)$$

The behavior of the above expression as the matrix V varies is better understood in components representation. By denoting as v_i the i^{th} column of V, it is easily seen that:

$$\left.\frac{d^2}{dt^2} det(\Phi^T A \Phi)\right|_{t=0} = 2det(\Delta) \sum_{k=1}^{q} \sum_{r=q+1}^{p} (v_k^T f_{i_r})^2 \left(\frac{\lambda_{i_r}}{\lambda_{i_k}} - 1 \right) \ , \qquad (21)$$

where again the symbols i_k denote an arbitrary choice of distinct integer indices within $\{1, \ldots, p\}$.

When $Q \neq \tilde{\Phi}$, the sign of the terms $\frac{\lambda_{i_r}}{\lambda_{i_k}} - 1$ may be either positive or negative, so that the sign of the derivative $\left.\frac{d^2}{dt^2} det(\Phi^T A \Phi)\right|_{t=0}$ may vary with the local orientation of the curve $\Phi(t)$ at $t = 0$. The only way in which such terms are surely non-positive, and so is the derivative $\left.\frac{d^2}{dt^2} det(\Phi^T A \Phi)\right|_{t=0}$, is that the set $\{\lambda_{i_k}\}_k$ contains the largest eigenvalues of the matrix A, that is when $Q = \tilde{\Phi}$.

Finally, by definition of the function $d_A(\cdot)$, it follows that:

$$d_A(\tilde{\Phi}) = \lambda_1 \ \lambda_2 \ \cdots \ \lambda_q,$$

and this completes the proof. □

Using statements 1) and 3) of the above Lemma, one gathers that $d_A(\Phi)$ has its maxima in $\Phi = \tilde{\Phi}K$ for any $K \in \mathbf{\Omega}_1$.

The second result of this section concerns a particular matrix equation that will arise in the main result presented in the next section.

Lemma 4 *Let A be a matrix in $\mathcal{M}(p,p)$, symmetric and endowed with p distinct and positive eigenvalues. The following matrix equation:*

$$A\Phi = \Phi(\Phi^T\Phi)^{-1}(\Phi^T A\Phi) \ , \tag{22}$$

with rank-q unknown $\Phi \in \mathcal{M}(p,q)$ enjoys the following properties:

1. *Any matrix $Q \in \mathbf{\Omega}_1(p,q)$ whose columns are q eigenvectors of A is a particular solution to the equation (22).*
2. *If $\tilde{\Phi}$ is a solution of the equation (22), then anyhow an invertible matrix $M \in \mathcal{M}(q,q)$ is chosen, the matrix $\tilde{\Phi}M$ is also a solution to the equation (22).*
3. *There do not exist other solutions to the equation (22) but for those given by 1) and 2).*

Proof

In order to prove the first claim, let us first replace matrix Φ with Q into equation (22), which then rewrites as:

$$AQ = Q(Q^TQ)^{-1}(Q^TAQ) \ ,$$

By definition of Q, a diagonal square matrix Δ exists such that $AQ = Q\Delta$. Using this property in the first equation yields:

$$AQ = Q(Q^TQ)\Delta = Q\Delta \ ,$$

which is an identity, whereby the proof of claim 1) follows.

In order to prove the second claim, it is sufficient to replace the expression $\Phi = \tilde{\Phi}M$ into the equation (22). In this way, we obtain:

$$\begin{aligned}
& A\tilde{\Phi}M = \tilde{\Phi}M(M^T\tilde{\Phi}^T\tilde{\Phi}M)^{-1}(M^T\tilde{\Phi}^TA\tilde{\Phi}M) \\
\Leftrightarrow \quad & A\tilde{\Phi}M = \tilde{\Phi}(MM^{-1})(\tilde{\Phi}^T\tilde{\Phi})^{-1}(M^{-T}M^T)(\tilde{\Phi}^TA\tilde{\Phi})M \\
\Leftrightarrow \quad & 0 = (A\tilde{\Phi} - \tilde{\Phi}(\tilde{\Phi}^T\tilde{\Phi})^{-1}\tilde{\Phi}^TA\tilde{\Phi})M \\
\Leftarrow \quad & 0 = A\tilde{\Phi} - \tilde{\Phi}(\tilde{\Phi}^T\tilde{\Phi})^{-1}\tilde{\Phi}^TA\tilde{\Phi} \ .
\end{aligned}$$

By hypothesis, $\tilde{\Phi}$ is a solution of the equation (22), thus the last equality holds true and through the above equivalence-chain the first equality is implied, hence the thesis.

The last claim is proven through the following arguments. Let us denote by F the $p \times p$ matrix whose columns are the p eigenvectors of the matrix A. The most general solution of the equation (22) is then $\Phi = FN$, where $N \in \mathcal{M}(p,q)$. By plugging this solution into equation (22) and by defining $\tilde{\Delta} := F^T A F$, we get:

$$F\tilde{\Delta}N = FN(N^T N)^{-1} N^T \tilde{\Delta} N \ . \tag{23}$$

This equation in N has only the solution $N = \Pi S I_{p,q}$, where $\Pi \in \mathcal{M}(p,p)$ is an arbitrary permutation matrix and $S \in \mathcal{M}(p,p)$ is a diagonal matrix with entries in $\{-1,+1\}$. (Of course, every solution of this kind post-multiplied by an arbitrary invertible $q \times q$ matrix is also a solution: As it was already granted by the proof of claim 2), such possibility may be safely ignored.) It is now easy to prove that $(N^T N)^{-1} = I_q$ and the remaining terms of the above equation are identical to those of equation (22), therefore the same arguments may be used to prove that equation (23) becomes an identity. This circumstance, proves the last claim. □

The above result characterizes completely the solutions of the matrix equations (22).

All the elements necessary to prove the adequateness of the Xu's original criterion (4) in the GSA learning system with respect to the target of extracting the principal subspace of a fixed order from a multivariate random process are now available.

5 Principal result about RUT

We can now study in details the RUT criterion function and present a discussion of its features through the following result, that stems as a formal review of arguments developed in [50].

Theorem 5 *Let A be a matrix in $\mathcal{M}(p,p)$, symmetric, positive definite and endowed with p distinct eigenvalues. Let the eigen-matrices Q and $\tilde{\Phi}$ defined as in **Lemma 3** and the value $\lambda_\star$ defined as:*

$$\lambda_\star := det(\tilde{\Phi}^T A \tilde{\Phi}) \ . \tag{24}$$

In addition, let the following function be defined:

$$J_X(\Phi) := \frac{\varphi[det(\Phi^T A \Phi)]}{\psi[det(\Phi^T \Phi)]} \ , \tag{25}$$

where $\Phi \in \mathcal{M}(p,q)$, with $q \leq p$, is a variable matrix, and φ and ψ are real functions of real variable such that:

(a) The function $\varphi(u) > 0$ is differentiable and monotonically increasing for $u > 0$ and $\psi(u) > 0$ is differentiable for $u > 0$;

(b) The function $\frac{\varphi(\lambda_\star u)}{\psi(u)}$ *assumes the maximum value in an unique point* $u = \bar{u}$ *with* $\bar{u}(\lambda_\star)$ *positive and bounded.*

Then, the function (25) assumes the maximum value in correspondence of the matrix-point $\Phi = \tilde{\Phi}M$*, where* M *is whatever invertible matrix in* $\mathcal{M}(q,q)$ *such that* $det^2(M) = \bar{u}(\lambda_\star)$*.*

Proof

The first issue is that an extremum of the function (25) must be a full-rank matrix. In fact, assuming a solution Φ such that $rk[\Phi] < q$ results in $det(\Phi^T A\Phi) = det(\Phi^T\Phi) = 0$. Therefore, such solution may not be a point of maximum, in fact, from the condition (b) it follows that another matrix Φ_1 could be found such that $J_X(\Phi_1) > J_X(\Phi)$, because condition (b) requires the maximum argument to be positive. Consequently, we may restrict to searching among rank-q extremes of the function (25).

The extremes of the function $J_X(\cdot)$ are among the solutions of the equilibrium equation $\frac{\partial J_X}{\partial \Phi} = 0$, that is:

$$A\Phi = a\Phi(\Phi^T\Phi)^{-1}(\Phi^T A\Phi), \text{ where:} \tag{26}$$

$$a := \frac{\psi'\varphi det(\Phi^T\Phi)}{\psi\varphi' det(\Phi^T A\Phi)} \ , \tag{27}$$

and where we made use of the following notation, for the sake of conciseness:

$$\varphi = \varphi[det(\Phi^T A\Phi)], \quad \psi = \psi[det(\Phi^T\Phi)] \ ,$$
$$\varphi' = \varphi'[det(\Phi^T A\Phi)], \quad \psi' = \psi'[det(\Phi^T\Phi)] \ .$$

It is worth observing that the equation (26) is consistent, in fact, from the above considerations about the rank of the admissible solutions, the product matrix $\Phi^T\Phi$ is invertible.

Pre-multiplying both sides of equation (26) by Φ^T gives the scalar condition $a = 1$, so that the equation (26) simplifies into:

$$A\Phi = \Phi(\Phi^T\Phi)^{-1}(\Phi^T A\Phi) \ . \tag{28}$$

It is the auto-subspace equation of **Lemma 4**, from which it is known that all the solutions of equation (28) are of the form $\Phi = QM$, where M is whatever invertible matrix in $\mathcal{M}(q,q)$. (It is worth noting that it results $rk[\Phi] = q$, as required.) Replacing such a solution in the definition (28) yields:

$$J_X(QM) = \frac{\varphi[det(M^T Q^T AQM)]}{\psi[det(M^T Q^T QM)]} = \frac{\varphi[\mu \cdot det(Q^T AQ)]}{\psi(\mu)} \ . \tag{29}$$

where, for the sake of notation conciseness, the auxiliary variable $\mu := det^2(M)$ has been introduced.

Due to peculiar dependency of the function $J_X(\cdot)$ upon the variable μ and the quantity $det(Q^T AQ)$, the function $J_X(\cdot)$ may be easily optimized with respect

to the latter quantity. In fact, by definition the function $\varphi(\cdot)$ is monotonically increasing so $J_X(\cdot)$ is monotonically increasing with $det(Q^T AQ)$. From **Lemma 3** we know that this function assumes in fact the maximum value $\lambda_\star$.

In virtue of this result, the equation (29) and the condition $a = 1$ can be rewritten in the unknown μ as:

$$\tilde{J}_X(\mu) = \frac{\varphi(\lambda_\star \mu)}{\psi(\mu)} , \tag{30}$$

$$\tilde{a}(\mu) - 1 = \frac{\psi'(\mu)\varphi(\lambda_\star \mu)}{\lambda_\star \psi(\mu)\varphi'(\lambda_\star \mu)} - 1 = 0 . \tag{31}$$

The optimal point Φ must satisfy both of them. Thanks to the condition (b), an unique point $\bar{u} \neq 0$ exists such that the function $\tilde{J}_X(\mu)$ maximizes, namely such that:

$$\lambda_\star \varphi'(\lambda_\star \bar{u})\psi(\bar{u}) - \varphi(\lambda_\star \bar{u})\psi'(\bar{u}) = 0 .$$

By the equation (31), this point coincides to the point where $\tilde{a} = 1$, that is $\mu = \bar{u}$. In other words, M can be arbitrarily chosen as long as condition $det^2(M) = \bar{u}(\lambda_\star)$ holds true. □

The Theorem just proven shows that the maximum of the Xu's function is reached when the matrix Φ contains as its columns *whatever invertible linear combination* of the principal eigenvectors of the (covariance) matrix A, provided the squared determinant of the matrix M equals the constant $\bar{u}$, which depends on the shape of the functions $\varphi(\cdot)$ and $\psi(\cdot)$ and on the product of the q largest eigenvalues $\lambda_\star$.

The main condition (b) of **Theorem 5** may be easily fulfilled. For instance, a useful choice that (as it will be shown later) implies some interesting consequences on the structure of the system, is $\varphi(u) := ln(u)$. If we further assume, for instance, $\psi(u) := u$, then the ratio $\varphi(\lambda_\star u)/\psi(u)$ assumes its maximum value in $\bar{u} = e/\lambda_\star$, where '$e$' is the Neper's number. Moreover, with the above assumptions, the function $a(\cdot)$ simplifies into $a(\Phi) = ln[det(\Phi^T A\Phi)]$.

6 Orthonormal dynamics

The present section is devoted to the study of the dynamics of the GSA system with the Xu's index. As an important analysis we are going to show that gradient steepest ascent heavily affects the characteristics of the solutions that can be learnt by the neural system, in the sense that, depending on the initial state chosen for the system, only a restricted subset of the set of the solutions described in the **Theorem 5** can be effectively achieved. In particular, we shall be able to demonstrate that, starting with certain particular initial conditions, the Xu's learning method can determine only PSA$^\perp$ of the random process

under analysis. That is to say, there exists an invariant submanifold for the learning algorithm.

6.1 Analysis of the GSA system with the RUT index

We are now interested in the dynamical properties of Xu's principal subspace estimation system for the PSA of a random process with covariance matrix Σ_x, namely:

$$\left\{ \begin{array}{lcl} \frac{dW}{dt} & = & \gamma \frac{\partial J_X(W)}{\partial W} \ , \\ \frac{\partial J_X(W)}{\partial W} & = & 2\frac{\varphi' \psi det(W^T \Sigma_x W) \Sigma_x W (W^T \Sigma_x W)^{-1} - \varphi \psi' det(W^T W) W (W^T W)^{-1}}{\psi^2} \ , \end{array} \right. \tag{32}$$

with $J_X(\cdot)$ as defined in the equation (25) with $A = \Sigma_x$. In the present section we consider, for simplicity, $\eta = \frac{1}{2}$. The properties of the above system are elucidated in the following result.

Theorem 6 (Orthonormal evolution.) *If in the system (32) an initial condition $W(0) \in \mathbf{\Omega}(p,q)$ is assumed, then for all $t > 0$ it results $W(t) \in \mathbf{\Omega}(p,q)$.*

Proof.

From the hypothesis, the matrix $W(0)$ is such that $W^T(0)W(0) = w_0^2 I_q$ for some $w_0 \in \mathcal{R} \setminus \{0\}$.

By handling the plain expression of the gradient at the second member of the equation (32), and defining the function:

$$\zeta(W) := \frac{\psi[det(W^T W)]}{\varphi'[det(W^T \Sigma_x W)] det(W^T \Sigma_x W)} \ , \tag{33}$$

the learning system (32) assumes the following expression:

$$\frac{dW}{dt} = \frac{1}{\zeta(W)} [\Sigma_x W (W^T \Sigma_x W)^{-1} - a(W) W (W^T W)^{-1}] \ , \tag{34}$$

with $a(W)$ being defined as in formula (27).

Pre-multiplying the preceding equation by W^T yields:

$$\begin{array}{rcl} \zeta(W) W^T \frac{dW}{dt} & = & (W^T \Sigma_x W)(W^T \Sigma_x W)^{-1} - a(W)(W^T W)(W^T W)^{-1} = \\ & = & [1 - a(W)] I_q \ . \end{array}$$

From $\zeta W^T \dot{W} = (1-a) I_q$ it also follows that $\zeta \dot{W}^T W = (1-a) I_q$, and from both equations it further follows that:

$$W^T \frac{dW}{dt} + \frac{dW^T}{dt} W = \frac{d(W^T W)}{dt} = \frac{2(1-a)}{\zeta} I_q \ ,$$

therefore the following strong property holds for all $t \geq 0$:

$$W^T(t) W(t) = w^2(t) I_q \ , \tag{35}$$

where $w(t)$ is a real function of the time and, in general, of w_0. □

It is interesting to note that when the initial conditions specified in the preceding Theorem are fulfilled, the function $\zeta(W)$ defined in (33) may assume a simplified form. In fact, with the above notation, from the equation (33) it follows that at each temporal instant the identity $det(W^T W) = w^{2q}$ holds true. Consequently, in the special case that $\varphi(x) := ln(x)$, it results $\zeta(W) = \psi(w^{2q})$. Thus the evolution of the variable $w(t)$ defined in the equation (33) is governed by the *scalar* differential equation:

$$\frac{dw^2}{dt} = \frac{2(1-a)}{\psi(w^{2q})} , \tag{36}$$

with the initial condition $w(0) = w_0$.

Unfortunately, such an equation does not result, in general, resolvable in closed form, since the variable a does not only depend on the scalar w but on the whole matrix W.

As a consequence of the Theorem just proven, if we assume for the continuous-time GSA learning system (32) a pseudo-orthonormal initial condition, it determines only pseudo-orthonormal solutions. Naturally, this fact restricts the set of the solutions admitted by the **Theorem 5** to the only subset $\mathbf{\Omega}(p,q)$ of $\mathcal{M}(p,q)$. More formally, we can state the following:

Corollary 7 *Let x be a random process in $\mathcal{R}^p$, having zero mean and being endowed with a finite covariance, and let W be a matrix variable in $\mathcal{M}(p,q)$, with $q \leq p$ fixed. Let, furthermore, be defined the criterion $J_X(\cdot)$ as in the equation (25), and the whole hypotheses and definitions of the **Theorem 5** be fulfilled. If as initial state $W(0)$ for the system (32) a configuration in $\mathbf{\Omega}(p,q)$ is assumed, and supposing the system be asymptotically stable, then the state $W(t)$ asymptotically converges to a PSA$^{\perp}$ of order q of the random process x.*

Proof.

By the hypotheses, the dynamical system is asymptotically stable, therefore the state-matrix W asymptotically converges to an equilibrium configuration $W_\star$. From the **Theorem 5** and the **Theorem 6**, the matrix $W_\star$ must contain orthonormal linear combinations of q generators of the principal subspace $\mathcal{P}_q(x)$ of the process as its columns, therefore, from the **Definition 2**, $W_\star$ is a PSA$^{\perp}$ of order q of x. □

Another particular result is that it is possible to formally determine the asymptotic solutions of the differential equation (36).

Corollary 8 *Reassuming the whole definitions and supposing fulfilled the whole conditions specified in the **Theorem 5** for the system (32), if an initial state*

$W(0) \in \mathbf{\Omega}_r(p,q)$ is assumed, then the admissible equilibrium points for it are all the matrices W of the form $W = QM$, with $M \in \mathbf{\Omega}_m(q,q)$, only if either $m = +\sqrt[2q]{\bar{u}}$ or $m = -\sqrt[2q]{\bar{u}}$.

Proof.

From the **Theorem 5**, we know that $W = QM$, with $det^2(M) = \bar{u}$. But from the **Theorem 6**, however, we also know that $W^T W = w^2 I_q$, therefore the $M^T Q^T Q M = w^2 I_q$ must hold. By definition, yet, it holds that $Q^T Q = I_q$, thus $M^T M = w^2 I_q := m^2 I_q$, whereby the $m^{2q} = det^2(M)$ follows. Equating yields $m^{2q} = \bar{u}$, that is a necessary condition. □

6.2 Stability considerations and convergence results

The results about the GSA system shown in the preceding subsection are sufficient to conclude that the dynamical system (32), with orthonormal initial conditions, may of course be either stable or unstable, but its stability properties are determined only by the stability properties of the scalar system (36). In particular:

- The system (32) is "rotationally asymptotically stable" if and only if the functions φ and ψ are such that it exist a real bounded value w_∞ such that for the solution $w(t)$ of the equation (36) the following condition:

  $$\lim_{t \to +\infty} w(t) = w_\infty,$$

  holds true.

- The system (32) is "rotationally simply stable" if and only if the functions φ and ψ are such that a real bounded positive value w_B and a temporal-instant t_B exist such that condition:

  $$\forall t \geq t_B : |w(t)| \leq w_B,$$

  is satisfied.

Rotational stability implies that the norm of the state-matrix W is bounded.

The purpose of this section is to advance this knowledge by stating and proving an asymptotic stability Theorem, that ensures the learning algorithm is convergent to the expected solutions.

Theorem 9 (Asymptotic convergence). *Let us consider the dynamical system:*

$$\frac{dW}{dt} = \frac{1}{2} \frac{\partial}{\partial W} \frac{\varphi(det(W^T A W)}{\psi(det(W^T W))} \ , \tag{37}$$

where $A \in \mathcal{M}(p,p)$ is a symmetric matrix endowed with positive distinct eigenvalues, $W(0) \in \mathbf{\Omega}_{w_0}(p,q)$, $\varphi(u) > 0$ and $\psi(u) > 0$ are smooth functions such

that $\varphi(u)$ is monotonically increasing for $u > 0$, and for every $\lambda > 0$ the function $\varphi(\lambda u)/\psi(u)$ has a only bounded maximum for $u > 0$ bounded. Then the dynamical system (37) is asymptotically convergent to the equilibrium points of ***Corollary 7*** *and* ***Corollary 8****.*

Proof. In order to carry out the proof, we need the plain expression of the gradient in (37), that is:

$$\begin{aligned}\frac{\partial}{\partial W}\frac{\varphi(det(W^TAW)}{\psi(det(W^TW))} &= 2\frac{\varphi' det(W^TAW)AW(W^TAW)^{-1}}{\psi} \\ &- 2\frac{\varphi\psi' det(W^TW)W(W^TW)^{-1}}{\psi^2} \ . \end{aligned} \tag{38}$$

The dynamical system (37) is a special case of the system (32), obtained when $\eta = \frac{1}{2}$, for which the **Theorem 6** holds. Therefore, for every $t > 0$ it holds $W(t) \in \mathbf{\Omega}_{w(t)}$. The most general representation for every such matrix is $W(t) = w(t)FR(t)$, where F is defined as in **Lemma 3**, and $R(t)$ is a one-parameter family of curves in $\mathbf{\Omega}_1(p, q)$.

Thanks to the above decomposition, the quantities involved in the gradient (38) simplify as follows:

$$\begin{aligned} &W^TAW = w^2R^TF^TAFR = w^2R^T\tilde{\Delta}R \ , \text{ where } \tilde{\Delta} := F^TAF \ , \\ &det(W^TAW) = w^{2q}det(R^T\tilde{\Delta}R) \ , \\ &det(W^TW) = w^{2q} \ , \\ &AW = wAFR = wF\tilde{\Delta}R \ . \end{aligned}$$

By defining the scalar quantity $\lambda := det(R^T\tilde{\Delta}R)$ it is clearly seen that the dynamics of variables w and λ are responsible for the convergence of the learning rule (37). In fact, thanks to the above decomposition and identities, we have:

$$\begin{aligned} &\frac{1}{2}\frac{\partial}{\partial W}\frac{\varphi}{\psi} = w^{2q-1}F\frac{\varphi'(\lambda w^{2q})\psi(w^{2q})\lambda\tilde{\Delta}R(R^T\tilde{\Delta}R)^{-1} - \varphi(\lambda w^{2q})\psi'(w^{2q})R}{\psi^2(w^{2q})} \ , \\ &\frac{dW}{dt} = F\left(\frac{dw}{dt}R + w\frac{dR}{dt}\right) \ . \end{aligned}$$

By equating the right-hand sides of the latter two equations and by observing that the matrix F may be cancelled out because it is square and full-rank, we obtain the expression:

$$\begin{aligned}\frac{dw}{dt}R + w\frac{dR}{dt} &= w^{2q-1}\frac{\varphi'(\lambda w^{2q})\lambda\tilde{\Delta}R(R^T\tilde{\Delta}R)^{-1}}{\psi(w^{2q})} \\ &- w^{2q-1}\frac{\varphi(\lambda w^{2q})\psi'(w^{2q})R}{\psi^2(w^{2q})} \ . \end{aligned} \tag{39}$$

In order to search for an equation that captures the dynamics of the variable w, let us pre-multiply both members of the equation (39) by R^T. This operation

yields:

$$\frac{dw}{dt}I_q + wR^T\frac{dR}{dt} = w^{2q-1}\frac{\varphi'(\lambda w^{2q})\psi(w^{2q})\lambda - \varphi(\lambda w^{2q})\psi'(w^{2q})}{\psi^2(w^{2q})}I_q \ .$$

Hand-by-hand transposition of the above equation gives:

$$\frac{dw}{dt}I_q + w\frac{dR^T}{dt}R = w^{2q-1}\frac{\varphi'(\lambda w^{2q})\psi(w^{2q})\lambda - \varphi(\lambda w^{2q})\psi'(w^{2q})}{\psi^2(w^{2q})}I_q \ .$$

Then, by summing hand-by-hand the two latter equations, we get:

$$\frac{dw^2}{dt} = 2w^{2q}\frac{\varphi'(\lambda w^{2q})\psi(w^{2q})\lambda - \varphi(\lambda w^{2q})\psi'(w^{2q})}{\psi^2(w^{2q})} \ , \tag{40}$$

thanks to the identities:

$$\frac{dR^T}{dt}R + R^T\frac{dR}{dt} = 0 \ , \ \frac{dw^2}{dt} = 2w\frac{dw}{dt} \ .$$

Now, it is easy to verify that for Xu's criterion $J_X = \frac{\varphi(\lambda w^{2q})}{\psi(w^{2q})}$ it holds:

$$\frac{\partial J_X}{\partial w^2} = qw^{-2}w^{2q}\frac{\varphi'(\lambda w^{2q})\psi(w^{2q})\lambda - \varphi(\lambda w^{2q})\psi'(w^{2q})}{\psi^2(w^{2q})} \ , \tag{41}$$

therefore, by comparing equations (40) and (41), we obtain the following result:

$$\frac{dw^2}{dt} = \left(\frac{2w^2}{q}\right)\frac{\partial}{\partial w^2}\left[\frac{\varphi(\lambda w^{2q})}{\psi(w^{2q})}\right] \ . \tag{42}$$

In order to derive an equation for the dynamics of the variable $\lambda = det(R^T\tilde{\Delta}R)$, let us start by the general formulas:

$$\frac{d}{dt}det(R^T\tilde{\Delta}R) = tr\left[\left(\frac{\partial}{\partial R}det(R^T\tilde{\Delta}R)\right)^T\frac{dR}{dt}\right] \ ,$$
$$\frac{\partial}{\partial R}det(R^T\tilde{\Delta}R) = 2det(R^T\tilde{\Delta}R)\tilde{\Delta}R(R^T\tilde{\Delta}R)^{-1} \ .$$

By straightforward computations and by exploiting the previous identities and the properties of the trace operator, we get:

$$\frac{d\lambda}{dt} = \lambda\, tr\left[(R^T\tilde{\Delta}R)^{-1}\frac{d(R^T\tilde{\Delta}R)}{dt}\right] \ . \tag{43}$$

It is now worth pre-multiplying both sides of the equation (39) by $R' := \frac{dR}{dt}$, which gives:

$$w'R'^TR + wR'^TR' = w^{2q-1}\frac{\varphi'\psi\lambda(R'^T\tilde{\Delta}R)(R^T\tilde{\Delta}R)^{-1} - \varphi\psi'R'^TR}{\psi^2} \ .$$

The above expression, transposed hand-by-hand, reads:

$$w'R^TR' + wR'^TR' = w^{2q-1}\frac{\varphi'\psi\lambda(R^T\tilde{\Delta}R)^{-1}(R^T\tilde{\Delta}R') - \varphi\psi'R^TR'}{\psi^2} \ .$$

Summing up the latter two equations hand-by-hand gives:

$$R'^T R' = \frac{w^{2(q-1)}\varphi'\lambda}{2\psi}[(R'^T\tilde{\Delta}R)(R^T\tilde{\Delta}R)^{-1} + (R^T\tilde{\Delta}R)^{-1}(R^T\tilde{\Delta}R')] \ . \qquad (44)$$

Finally, by applying the trace operator hand-by-hand we obtain:

$$\begin{aligned} tr(R'^T R') &= \frac{w^{2(q-1)}\varphi'\lambda}{2\psi} tr\left[(R^T\tilde{\Delta}R)^{-1}\frac{d}{dt}(R^T\tilde{\Delta}R)\right] \\ &= \frac{w^{2(q-1)}\varphi'\lambda}{2\psi}\cdot\frac{1}{\lambda}\frac{d\lambda}{dt} \ , \end{aligned}$$

because of the identity (43). As a consequence, the dynamics of the variable λ is governed by the differential equation:

$$\frac{d\lambda}{dt} = \frac{2\psi}{w^{2(q-1)}\varphi'} tr(R'^T R') \ . \qquad (45)$$

By definition, the functions φ and ψ' assume positive values, while the quantity $tr(R'^T R') \geq 0$, with equality holding only for $R' = 0$.

In summary, to the purpose of convergence analysis, the dynamics of the learning system (37) is described by the differential system:

$$\left\{ \begin{array}{l} \frac{d\lambda}{dt} = \frac{2\psi(w^{2q})tr(R'^T R')}{w^{2(q-1)}\varphi'(\lambda w^{2q})} \ , \\ \frac{dw^2}{dt} = \frac{2w^2}{q}\frac{\partial}{\partial w^2}\frac{\varphi(\lambda w^{2q})}{\psi(w^{2q})} \ . \end{array} \right. \qquad (46)$$

As a consequence of the first equation, the value of the variable λ is constantly increasing toward its maximum value $\lambda_\star$. As a consequence of this and of the second equation, for every value of the variable λ, the dynamical system (37) searches for the (unique) value of w^2 that maximizes the criterion function J_X. Therefore, the dynamical system (37) is asymptotically convergent to the equilibrium points derived in the **Corollary 7** and **Corollary 8**. □

7 Complements and discussion

In the preceding sections, we have reported a detailed analysis of the Uncertainty Maximization Theory and of the properties of the corresponding GSA system.

The present section is devoted to the presentation of some complements to the above analysis of the RUT differential system. In particular, in the following we briefly consider three aspects:

- The relationship among the plain GSA differential learning system and the natural gradient theory: We show that pseudo-orthonormality in this specific case is natural-gradient invariant.

- The possibility of introducing a control law for the dynamics of the RUT learning systems, that allows to control the converge speed and possibly the stability of these systems.
- A brief discussion of some theoretical points that are just touched by the present contribution and would deserve deeper investigation in the future.

7.1 Relationships with natural gradient theory

Shortly speaking, the "natural gradient" theory [2] arises by considering as network parameter space the set of proper-size full-rank matrices and by endowing it with specific geometrical features. This amounts to the post-multiplication of the Euclidean gradient of a learning criterion by $W^T W$, which gives the Riemannian or "natural" gradient.

From a different perspective, we may observe that, since the product matrix $W^T W$ is positive-definite when W is "tall-skinny" and full rank, the gradient in the equation (32) may be post-multiplied by $W^T W$, obtaining the new system:

$$\frac{dW}{dt} = \frac{\gamma}{\zeta(W)}[\Sigma_x W (W^T \Sigma_x W)^{-1} W^T - a(W) I_p] W \ . \tag{47}$$

Such subspace learning system has some interesting properties, too. First, it is simpler than the original one, because its use does not involve the inverse $(W^T W)^{-1}$. Moreover, it is straightforward to show that it allows a pseudo-orthonormal evolution of the state-matrix W.

Theorem 10 *Let Σ_x be a constant real-valued symmetric matrix in $\mathcal{M}(p,p)$, and W a variable matrix in $\mathcal{M}(p,q)$. Given the system (47), where $a = a(W)$ and $\eta = \eta(W)$ are real scalar functions of W, if the initial state $W(0)$ is pseudo-orthonormal, then for all $t > 0$ the matrix W stays pseudo-orthonormal.*

Proof.

Pre-multiplying both members of the dynamical equation (47) by W^T yields:

$$W^T \frac{dW}{dt} = \frac{(1-a)\gamma}{\zeta} W^T W \ ,$$

therefore it holds true that:

$$\frac{d(W^T W)}{dt} = \frac{2(1-a)\gamma}{\zeta} W^T W \ . \tag{48}$$

From the hypothesis, the product $W^T(0)W(0)$ is diagonal, and the structure of the equation (48) is such that it preserves the diagonality, therefore the product $W^T(t)W(t)$ stays diagonal for every $t > 0$. □

From the above Theorem it follows that $W^T(t)W(t) = w^2(t) I_q$: In other terms, if $W(0) \in \mathbf{\Omega}_{w_0}$(p,q) then $W \in \mathbf{\Omega}_w$(p,q). Moreover, with the special

choice $\varphi(x) := ln(x)$ the expression (48) reduces to a scalar differential equation for $w^2(t)$, that reads:

$$\frac{dw^2}{dt} = 2\gamma \frac{w^2(1-a)}{\psi(w^{2q})} \ , \tag{49}$$

which plays the same role of the equation (36) for the corresponding matrix system.

7.2 A 'control law' for the differential equations

Although it is known *a priori* from the previous section that, at the equilibrium $W_\star$, the property $a(W_\star) = 1$ holds, so far we have not been able to take advantage of this knowledge. Now, observing equations (34) and (47), we argue that the variable a has the important role to weight the second term into the square brackets with respect to the first term. This simple consideration suggests that we could replace in the above-mentioned differential learning equations the function $a(W)$ as defined in (27) with a chosen function *only* of the time parameter, $a(t)$, chosen so that at least the condition:

$$\lim_{t \to +\infty} a(t) \ = \ 1 \ , \tag{50}$$

holds and such that the second terms in the square brackets in (34) and (47) are weighted in a proper way. We name the function $a(t)$ a *control law* for the learning differential equations.

Actually, by replacing the true expression of the parameter a with a chosen time-function we do not change the structure of the learning systems, but we simply force the temporal evolution of the variables to a non-self-controlled course. An equivalent viewpoint is that, by using the notation introduced in the **Lemma 3**, the relationship:

$$\psi'[W(t)]\varphi[W(t)]d_{I_p}[W(t)] = a(t)\psi[W(t)]\varphi'[W(t)]d_{\Sigma_x}[W(t)] \tag{51}$$

is forced to hold. As a consequence of this assumption, the systems (32) and (47) become, respectively:

$$\frac{dW}{dt} \ = \ \frac{2\gamma}{\zeta(W)}[\Sigma_x W(W^T\Sigma_x W)^{-1} - a(t)W(W^TW)^{-1}] \ , \tag{52}$$

$$\frac{dW}{dt} \ = \ \frac{2\gamma}{\zeta(W)}[\Sigma_x W(W^T\Sigma_x W)^{-1}W^T - a(t)I_p]W \ . \tag{53}$$

The main consequence of the explained modification is that, as $a(t)$ is a known function only of the parameter t, the scalar differential equations (36) and (49) can be solved (at least in principle) by calculating the following integrals:

$$\int_{w_0^2}^{w^2(t)} \psi(u^{2q})du^2 = 2\gamma \int_0^t [1 - a(\theta)]d\theta \ , \tag{54}$$

$$\int_{w_0^2}^{w^2(t)} \frac{\psi(u^{2q})du^2}{u^2} = 2\gamma \int_0^t [1 - a(\theta)]d\theta \ , \tag{55}$$

under the usual condition of the orthonormal initial states for the learning systems and $\varphi(x) = ln(x)$.

7.3 Brief discussion of further arguments

Some concluding observations about the presented analysis are briefly reported in the following:

- PSA appeared historically first, because the method of extracting principal directions looked difficult, but extracting the principal subspace looked easier. Currently, there exist several good algorithms and theories to extract the principal components directly. However, only the subspace can be extracted for the eigen-directions corresponding to a multiple or degenerated eigenvalue. The case that the eigenvalues are multiple has not been treated in the present contribution: Care should be taken of such pathological cases, in order to treat the general case.

- There are many cost functions to be optimized in the problem of PCA, MCA, PSA and MSA [19]. A possible criticism to the criterion considered in the present contribution is that it includes arbitrary functions (φ and ψ).

- The later idea of Brockett and Xu himself is to use $tr(D\Sigma)$, where D is a non-negative diagonal matrix and Σ is a covariance matrix. By using $tr(D\Sigma)$, it is possible to obtain PCA and MCA as well as PSA and MSA in a unified manner at the same time. When D is the identity matrix, the analysis reduces to PSA (or MSA), so it includes PSA and MSA as special examples. A unified general theory on this aspect was presented by Chen and Amari in [8, 9].

- The Stiefel manifold is implicit in the presented learning algorithm, but it could be emphasized more explicitly. For instance, the space of matrices $\mathbf{\Omega}(p, q; D)$ could be introduced, in which all the columns are orthogonal and their square length are d_i, the diagonal elements of D (see preceding points). The present case coincides to $D = r^2 I_q$, and $D = I_q$ gives the Stiefel manifold. In this case, any matrix can be decomposed as $X\sqrt{D}$, where X is in the Stiefel manifold. When D is fixed, the dynamics takes place in the Stiefel manifold, which has a Lie-invariant Riemannian metric. If we like to make D free, the dynamics separates into two parts, one in the Stiefel manifold and the other in the set of diagonal matrices.

- Chen and Amari not only studied the stability analysis inside the Stiefel manifold, but also studied the stability of the Stiefel manifold in the general matrix space. If the Stiefel manifold is not an attractor, because of the accumulation of numerical errors, the algorithms which are working inside the Stiefel manifold might no longer work in practice. It is worth noting that the numerical errors strongly depend on how the continuous-time differential equations that describes network's learning are integrated. A general discussion on this aspect are going to appear in [7]. The conclusion drawn in [7] is that care should be taken in order to choose a proper numerical integration algorithm. Fortunately, there exist effective and computationally-efficient techniques that allow us to integrate matrix-type differential equations on the orthogonal group and on the Stiefel manifold.

8 Conclusions

In this work we have dealt with a special class of matrix-type differential learning systems that have the set $\mathbf{\Omega}_r(p,q)$ of the orthonormal matrices as base manifold, providing that their initial states belong to $\mathbf{\Omega}_r(p,q)$, too.

As a case study, we have taken the Xu's RUT theory, which allows finding principal subspaces associated to an input random process. We have proven here that the original system and a derived one share common properties. In particular, their dynamical features may be studied by analyzing the behavior of single scalar differential equations. This implies that the rotational stability properties of the systems, for instance, can be inferred by these equations. Moreover, we have shown that often an $\mathbf{\Omega}_r(p,q)$-evolving system may be modified without loosing its desirable features nor affecting its steady-state behavior, conversely improving their dynamical and computational characteristics.

But, mainly, we have pointed out the existence of a class of optimization systems that *inherently* fulfill some constraints which, consequently, do not need to be explicitly taken into account when constructing the objective (cost, loss) learning function.

Another potentially interesting outcome of the presented analysis is that there exist some functions whose Euclidean gradient appear as Riemannian gradient on a curved manifold. It would be particularly interesting to discover some general result about this outcome because the existence of such functions (generally termed 'potential functions') would substantially help computing in a easy way some geometric quantities related to the (approximate) solution of matrix-type differential learning equations on manifolds, such as geodesic branches. To this respect, the present contribution treated a single case that may enhance the knowledge of such important analytical problem.

9 Acknowledgments

The present research work was initiated in September 1996 and was completed when the author was a short-term visitor of the Mathematical Neuroscience Laboratory of the Brain Science Institute (BSI) at RIKEN (Japan), in February-March 2004. The author wishes to express his gratitude to the BSI director, Prof. Shun-ichi Amari, for the interesting and useful comments to the manuscript (that have been summarized in Section 7.3) and to the laboratory members for the kindest and warmest hospitality.

The Author also wishes to express his gratitude to the anonymous reviewers whose careful proofreading of the manuscript and detailed comments greatly contributed to improving the paper quality and readability.

References

[1] S. Affes and Y. Grenier, *A process subspace tracking algorithm for speech acquisition and noise reduction with a microphone array*, Proc. of IEEE/IEE Workshop on Signal Processing Methods in Multipath Environments, pp. 64 – 73, 1995

[2] S.-i. Amari, *Natural gradient works efficiently in learning*, Neural Computation, Vol. 10, pp. 251 – 276, 1998

[3] S.-i., Amari, *Mathematical foundations of neurocomputing*, Proc. of IEEE, Vol. 78, No. 9, pp. 1443 – 1463, Sept. 1990

[4] D.P. Bertekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, 1982

[5] P. Bloomfield and W.L. Steiger, *Least Absolute Deviation: Theory, Applications and Algorithms*, Birkäuser, 1993

[6] R.W. Brockett, *Dynamical systems that sort lists, diagonalize matrices and solve linear programming problems*, Linear Algebra and its Applications, Vol. 146, pp. 79 – 91, 1991

[7] E. Celledoni and S. Fiori, *Neural learning by geometric integration of reduced 'rigid-body' equations*, Journal of Computational and Applied Mathematics (JCAM). Accepted for publication

[8] T.-P. Chen, S.-i. Amari and Q. Lin, *A unified algorithm for principal and minor components extraction*, Neural Networks, Vol. 11, No. 3, pp. 385 – 390, 1998

[9] 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

[10] S. Costa and S. Fiori, *Image compression using principal component neural networks*, Image and Vision Computing Journal (special issue on "Artificial Neural Network for Image Analysis and Computer Vision"), Vol. 19, No. 9-10, pp. 649 – 668, Aug. 2001

[11] R. Courant, *Variational methods for the solution of problems of equilibrium and vibrations*, Bull. American Mathematical Society, Vol. 49, pp. 1 – 23, 1943

[12] K.D. Diamantaras and S.-Y. Kung, *Principal Components Neural Networks: Theory and Applications*, New York: Wiley, 1996

[13] S.C. Douglas, S.-i. Amari and S.Y. Kung, *On gradient adaptation with unit-norm constraints*, IEEE Trans. on Signal Processing, Vol. 48, No. 6, pp. 1843 – 1847, June 2000

[14] A. Edelman, T.A. Arias and S.T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM Journal on Matrix Analysis Applications, Vol. 20, No. 2, pp. 303 – 353, 1998

[15] A. Einstein, *Die Grundlagen der allgemeinen Relativitätstheorie*, Annalen der Physik, Vol. 4, No. 49, pp. 769 – 822, 1916

[16] S. Fiori, *A theory for learning by weight flow on Stiefel-Grassman manifold*, Neural Computation, Vol. 13, No. 7, pp. 1625 – 1647, July 2001

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

[18] S. Fiori, *Unsupervised neural learning on Lie group*, International Journal of Neural Systems, Vol. 12, No.s 3 & 4, pp. 219 – 246, 2002

[19] S. Fiori, *A minor subspace algorithm based on neural Stiefel dynamics*, International Journal of Neural Systems, Vol. 12, No. 5, pp. 339 – 350, 2002

[20] R. Fletcher, *Practical Methods of Optimizations*, J. Wiley - Interscience, 1986

[21] G.H. Golub and C.F. Van Loan, *Matrix Computations*, Third edition, Baltimore, MD: Johns Hopkins University Press, 1996

[22] F.R. Hampel, E.N. Ronchetti, P. Rousser and W.A. Stachel, *Robust Statistics - The Approach Based on Influence Functions*, J. Wiley, 1987

[23] S. Haykin, *Neural Network*, Ed. MacMillan College Publishing Company, 1994

[24] M.R. Hestens, *Multiplier and gradient methods*, Journal Optimization Theory and Applications, Vol. 4, pp. 303 – 320, 1969

[25] G. Hori, *A new approach to joint diagonalization*, Proc. of Second International Workshop on Independent Component Analysis and Blind Signal Separation (ICA'2000), pp. 151 – 155, Helsinki, Finland, June 19-22, 2000

[26] P.J. Huber, *Robust Statistics*, J. Wiley, 1981

[27] J. Karhunen and J. Joutsensalo, *Learning of robust principal component subspace*, Proc. of International Joint Conference on Neural Networks (IJCNN), pp. 2409 – 2412, 1993

[28] M. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Vol. 1 (Distribution theory), Ed. Charles Griffin & Co, 1977

[29] K.H. Knuth, *Bayesian source separation and localization*, SPIE'98 Proceedings: Bayesian Inference for Inverse Problems, San Diego, pp. 147 – 158, 1998

[30] B. Laheld and J.F. Cardoso, *Adaptive source separation with uniform performance*, Signal Processing VII: Theories and Applications (EUSIPCO), Vol. 1, pp. 183 – 186, 1994

[31] G. Lebanon, *Computing the volume element of a family of metrics on the multinomial simplex*. Technical report CMU-CS-03-145, Cernegie Mellon University, 2003

[32] R.-w. Liu, *Blind signal processing: An introduction*, Proc. of International Symposium on Circuits and Systems (IEEE-ISCAS), Vol. 2, pp. 81 – 84, 1996

[33] X. Liu, A. Srivastava and K. Gallivan, *Optimal linear representations of images for object recognition*, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 26, No. 5, pp. 662 – 666, May 2004

[34] B.C. Moore, *Principal component analysis in linear systems: Controllability, observability and model reduction*, IEEE Trans. on Automatic Control, Vol. AC-26, No. 1, pp. 17 – 31, 1981

[35] E. Moreau and J.C. Pesquet, *Independence/decorrelation measures with application to optimal representations*, Proc. of International Conference on Acoustic, Speech and Signal Processing, 1997, pp. 1935 – 1938

[36] V.A. Morozov, *Methods for Solving Incorrectly Problems*, Springer-Verlag, 1984

[37] H. Niemann and J.-K. Wu, *Neural network adaptive image coding*, IEEE Trans. on Neural Networks, Vol. 4, No. 4, pp. 615 – 627, July 1993

[38] E. Oja, *Subspace Methods of Pattern Recognition*, Research Studies Press, England, 1983

[39] E. Oja, A. Hyvärinen and P. Hoyer, *Image feature extraction and denoising by sparse coding*, Pattern Analysis and Applications Journal, Vol. 2, No. 2, pp. 104 – 110, 1999

[40] D.P. O'Leary, *Robust regression computation using iteratively reweighted least-squares*, SIAM Journal Matrix Analysis Applications, Vol. 11, pp. 466 – 480, 1990

[41] A.L. Peressini, F.E. Sullivan and J.J. Uhl, *The Mathematics of Nonlinear Programming*, Springer-Verlag, 1988

[42] F. Palmieri and J. Zhu, *A comparison of two eigen-network*, Proc. of International Joint Conference on Neural Network (IJCNN), Vol. II, pp. 193 – 199, 1991

[43] M.D. Plumbley, *Algorithms for nonnegative independent component analysis*, IEEE Trans. on Neural Networks, Vol. 14, No. 3, pp. 534 – 543, May 2003

[44] M.J.D. Powell, *A method for nonlinear constraints minimization problems*, in Optimization (Ed. R. Fletcher), Academic Press, pp. 283 – 258, 1969

[45] S.S. Rao, *Optimization: Theory and Applications*, J. Wiley - Eastern, 1978

[46] P. Saisan, G. Doretto, Y.N. Wu and S. Soatto, *Dynamic texture recognition*, Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, pp. 58 – 63, Dec. 2001

[47] A. Srivastava, A.B. Lee, E.P. Simoncelli and S.-C. Zhu, *On advances in statistical modeling of natural images*, Journal of Mathematical Imaging and Vision, Vol. 18, No. 1, pp. 17 – 33, 2003

[48] A.N. Tikhonov and V.Y. Arsenin, *Solutions of Ill-Posed Problems*, W.H. Winston, 1977

[49] J.M. Varah, *A practical examination of some numerical methods for linear discrete ill-posed problems*, SIAM Rewiev, Vol. 21, No. 1, pp. 100 – 111, Jan. 1979

[50] L. Xu, *Theories for unsupervised learning: PCA and its nonlinear extension*, Proc. of International Joint Conference on Neural Networks (IJCNN), 1994, pp. 1252 – 1257

[51] B. Yang, *Projection approximation subspace tracking*, IEEE Transaction on Signal Processing, Vol. 43, 1995, pp. 1247 – 1252

[52] H.H. Yang and S.-i. Amari, *Adaptive online learning algorithms for blind separation: maximum entropy and minimal mutual information*, Neural Computation, Vol. 9, pp. 1457 – 1482, 1997

[53] K. Zhang and T.J. Sejnowski, *A theory of geometric constraints on neural activity for natural three-dimensional movement*, Journal of Neuroscience, Vol. 19, No. 8, pp. 3122 – 3145, 1999

[54] P. Zhang, J. Peng and C. Domeniconi, *Kernel Pooled Local Subspaces for Classification*, IEEE Trans. on Systems, Man and Cybernetics, Part B: Cybernetics. (Special Issue on Computer Vision and Pattern Recognition.). To appear (2004)

[55] L.-Q. Zhang, S.-i. Amari and A. Cichocki, *Semiparametric model and superefficiency in blind deconvolution*, Signal processing, Vol. 81, pp. 2535 – 2553, 2001

[56] L.-Q. Zhang, A. Cichocki and S.-i. Amari, *Geometrical structures of FIR manifold and multichannel blind deconvolution*, Journal of VLSI for Signal processing Systems, Vol. 31, pp. 31 – 44, 2002