Neural Learning and Weight Flow on Stiefel Manifold

Simone Fiori, Aurelio Uncini, and Francesco Piazza [*]
Dipartimento di Elettronica e Automatica,
Università di Ancona, Ancona – Italy
{simone,aurel,upf}@eealab.unian.it

Abstract

The aim of this paper is to present a new class of learning models for linear as well as non-linear neural layers called *Orthonormal Strongly-Constrained* (SOC or Stiefel). They allow to solve orthonormal problems where orthonormal matrices are involved. After general properties of the learning rules belonging to this new class are shown, examples derived independently or by reviewing learning theories known from the literature are presented and discussed.

1 Introduction – Stiefel learning

Learning without supervision in a neural layer means adapting a weight matrix so that a predefined target is achieved. The columns of the weight matrix individuate directions in the weight-space, thus learning a weight matrix means also discovering *interesting* directions [6]. This way, an index is established over the set of all possible directions, so that any 'interesting' direction is quantitatively defined as the one maximizing the index.

In the present work we deal with the special case of the constrained search where the constraint is that of orthonormality of the columns of the neural layer's weight-matrix representing the state-variable. The associated learning (searching) algorithms are called *Orthonormal Strongly–Constrained* (SOC or Stiefel) and they allow to solve the orthonormal problems. Orthonormal problems (ONP) arise in several contexts, as those of the Optimal Linear Compression (Principal Component/Subspace Analysis, [4, 10, 11]), Whitening [7], Blind Separation of Sources [1, 4, 7, 8], DOA estimation [1]. In a ONP, the target of the adaptation rule for a neural network is to learn an orthonormal weight matrix related in some a way to the input signal. Since it is a-priori known that the final state must belong to the subset $\mathcal{N}$ of the whole space of searching containing orthonormal matrices, the evolution of the weight matrix may be strongly bounded to *always* belong to $\mathcal{N}$.

To denote an orthonormal matrix the set $\mathcal{N}_r^{p\times q}$ of the *r–orthonormal* $p \times q$ *matrices* as $\mathcal{N}_r^{p\times q} \overset{\text{def}}{=} \{\mathbf{W} \in \mathcal{R}^{p\times q} | \mathbf{W}^T\mathbf{W} = r^2\mathbf{I}\}$ is defined. By definition r

[*] This work was supported by the Italian MURST. **In Proc. of X Italian Workshop on Neural Networks, May 1998, Vietri sul Mare, Salerno – Italy**

must be non-null and $q \leq p$. The set $\mathcal{N}_r^{p\times q}$ is known as a *Stiefel manifold*. Formally we can then define a class of learning algorithms as:

$$\mathcal{C} \stackrel{\text{def}}{=} \{\mathcal{A}_J(\mathbf{W}) | \forall t \in \mathcal{T} : \mathbf{W}^T(t)\mathbf{W}(t) = w^2(t)\mathbf{I}\} \ , \tag{1}$$

where $\mathcal{A}_J(\mathbf{W})$ is a generic searching-algorithm with state-variable $\mathbf{W}$, $\mathcal{T}$ is a time interval the algorithm runs within, and $w(t)$ is a function differentiable at least once and non-null. J is the objective function whose extrema are searched. Each algorithm contained in $\mathcal{C}$ is called SOC (Orthonormal Strongly-Constrained) or Stiefel. Any algorithm in $\mathcal{C}$ is characterized by a fundamental property: Differentiating both members of $\mathbf{W}^T(t)\mathbf{W}(t) = w^2(t)\mathbf{I}$ with respect to the time yields:

$$\frac{d(\mathbf{W}^T\mathbf{W})}{dt} = \dot{\mathbf{W}}^T\mathbf{W} + \mathbf{W}^T\dot{\mathbf{W}} = 2\dot{w}w\mathbf{I} \ ,$$

which can be rewritten as:

$$\dot{\mathbf{W}}^T\mathbf{W} + \mathbf{W}^T\dot{\mathbf{W}} = \pi\mathbf{I} \ , \text{ with } 2\dot{w}w = \pi \ , \tag{2}$$

where $\dot{\mathbf{W}} = d\mathbf{W}/dt$. Hereafter, examples of SOC learning rules are presented and discussed.

2 A Xu's Principal Subspace algorithm

As a first example of an algorithm belonging to $\mathcal{C}$, here the one proposed by L. Xu in [11] is recalled. It allows the extraction of the Principal Subspace of fixed dimension from a multivariate random signal $\mathbf{x}$ with covariance matrix $\mathbf{C}_x$ having distinct and positive eigenvalues.

2.1 Xu's Subspace Rule is SOC

Xu proposed as an objective function the following criterion:

$$\varrho(\mathbf{W}) \stackrel{\text{def}}{=} \frac{\varphi[\det(\mathbf{W}^T\mathbf{C}_x\mathbf{W})]}{\psi[\det(\mathbf{W}^T\mathbf{W})]} \ , \tag{3}$$

where $\varphi(z)$ and $\psi(z)$ are positive monotonically increasing functions defined for $z > 0$, and $\det(\cdot)$ returns the *determinant* of the matrix contained within. Xu proved [11] that under some conditions the above function maximizes when the columns of $\mathbf{W}$ span a Principal Subspace of $\mathbf{x}$.

The Gradient Steepest Ascent (GSA) algorithm $\mathcal{A}_J(\mathbf{W})$ based on the above objective function is SOC. This property is proved in the following result.

Theorem 1. *If for the system* $\dot{\mathbf{W}} = \partial\varrho/\partial\mathbf{W}$ *an initial condition* $\mathbf{W}(0) \in \mathcal{N}^{p\times q}$ *is assumed, then for all* $t > 0$ *the flow* $\mathbf{W}(t) \in \mathcal{N}^{p\times q}$.

Proof. By handling the plain expression of the gradient $\partial\varrho/\partial\mathbf{W}$ and defining the functions:

$$\beta(\mathbf{W}) \stackrel{\text{def}}{=} \frac{\varphi'[\det(\mathbf{W}^T\mathbf{C}_x\mathbf{W})]\det(\mathbf{W}^T\mathbf{C}_x\mathbf{W})}{\psi[\det(\mathbf{W}^T\mathbf{W})]} \ ,$$

$$\alpha(\mathbf{W}) \stackrel{\text{def}}{=} \frac{\psi'\varphi\cdot\det(\mathbf{W}^T\mathbf{W})}{\psi\varphi'\cdot\det(\mathbf{W}^T\mathbf{C}_x\mathbf{W})} \ ,$$

the system $\dot{\mathbf{W}} = \partial\varrho/\partial\mathbf{W}$ reads:

$$\beta\frac{d\mathbf{W}}{dt} = \mathbf{C}_x\mathbf{W}(\mathbf{W}^T\mathbf{C}_x\mathbf{W})^{-1} \Leftrightarrow \alpha\mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1} \ . \tag{4}$$

Pre-multiplying the preceding equation by $\mathbf{W}^T$ yields:

$$\beta\mathbf{W}^T\frac{d\mathbf{W}}{dt} = (\mathbf{W}^T\mathbf{C}_x\mathbf{W})(\mathbf{W}^T\mathbf{C}_x\mathbf{W})^{-1} \Leftrightarrow \alpha\cdot(\mathbf{W}^T\mathbf{W})(\mathbf{W}^T\mathbf{W})^{-1} \ . \tag{5}$$

Moreover $\beta\mathbf{W}^T\dot{\mathbf{W}} = (1 \Leftrightarrow \alpha)\mathbf{I} \ \Rightarrow \ \beta\dot{\mathbf{W}}^T\mathbf{W} = (1 \Leftrightarrow \alpha)\mathbf{I}$, and their hand-by-hand addition gives:

$$\mathbf{W}^T\frac{d\mathbf{W}}{dt} + \frac{d\mathbf{W}^T}{dt}\mathbf{W} = \frac{d(\mathbf{W}^T\mathbf{W})}{dt} = \frac{2(1 \Leftrightarrow \alpha)}{\beta}\mathbf{I} \ , \tag{6}$$

therefore the SOC strong property $\mathbf{W}^T(t)\mathbf{W}(t) = w^2(t)\mathbf{I}$ holds for all $t \geq 0$, whereby the thesis follows. □

The quantity $w(t)$ is a real function of the time and of the initial value $w^2(0) = \frac{1}{q}\mathrm{tr}[\mathbf{W}^{\mathrm{T}}(0)\mathbf{W}(0)]$. Its evolution is governed by the function π_X given by $\beta\pi_X = 2(1 \Leftrightarrow \alpha)$. It would be interesting to note that with the choice $\varphi(z) = \log(z)$, the preceding expression simplifies, since $\beta[\mathbf{W}(t)] = 1/\psi(w^{2q}(t))$. Likely, the simplest form of α is obtained when $\psi(z) = z$, indeed in that case $\alpha(\mathbf{W}) = \det(\mathbf{W}^T\mathbf{C}_x\mathbf{W})$.

2.2 A 'learning control law' $\bar{\alpha} = \bar{\alpha}(t)$

From equation (6) (see also [11]), it is known that at the equilibrium the property $\alpha(\mathbf{W}) = 1$ must hold. We now propose to replace in the algorithm (4) the function $\alpha(\mathbf{W})$ with an arbitrariy function of the time only, $\bar{\alpha}(t)$, chosen so that at least condition:

$$\lim_{t\to t_\star}\bar{\alpha}(t) \ = \ 1 \ , \tag{7}$$

holds true, and such that the second term in (4) is weighted in a proper way. The function $\bar{\alpha}(t)$ is called here *learning control law*. Time $t_\star$ may not be bounded.

Actually, by replacing the true expression of the parameter α with an arbitrarily chosen time-function we do not change the structure of the algorithm, but simply force the temporal evolution of the variables to a non-self-controlled course, and control the speed the state $\mathbf{W}(t)$ travels with along the Stiefel manifold $\mathcal{N}^{p\times q}$. As a consequence of this assumption the system (4) becomes:

$$\beta\frac{d\mathbf{W}}{dt}=\mathbf{C}_x\mathbf{W}(\mathbf{W}^T\mathbf{C}_x\mathbf{W})^{-1}\Leftrightarrow\bar{\alpha}(t)\mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\ .$$

The main consequence of the earlier assumption is that, as $\bar{\alpha}(t)$ is a known function of the parameter t only, the scalar differential equation in (2) can be easily tackled (at least in the intentions) by solving the following integral equation:

$$\int_{w_0}^{w(t)}\frac{u\,du}{\psi(u^{2q})}=\int_0^t[1\Leftrightarrow\bar{\alpha}(\theta)]d\theta\ , \tag{8}$$

under the usual condition of the orthonormal initial state for the algorithm, providing that $\varphi(x)=\log(x)$.

3 A 2^{nd}-order differential learning equation

As a second example of algorithms belonging to $\mathcal{C}$, consider the following second order non-linear differential learning rule for the neural layer described by the input-output relation $\mathbf{y}=\mathbf{G}(\mathbf{W}^T\mathbf{x}+\mathbf{w}_0)$:

$$\mathbf{W}^T\ddot{\mathbf{W}}\Leftrightarrow\ddot{\mathbf{W}}^T\mathbf{W}=2\sigma\rho\mathbf{W}^T(\mathbf{H}\Leftrightarrow\mathbf{H}^T)\mathbf{W}\ , \tag{9}$$

where $\ddot{\mathbf{W}}=d^2\mathbf{W}/dt^2$, σ and ρ are positive constants and $\mathbf{H}$ is an arbitrarily variable $p\times p$ matrix, which controls the learning. $\mathbf{G}(\cdot)$ is a non-linear diagonal activation operator and $\mathbf{w}_0$ is a biasing vector arbitrarily adapted.

Let us prove that algorithm (9) is SOC if $\mathbf{W}(0)\in\mathcal{N}_{w_0}^{p\times q}$. To start with, it is useful to define the following coupled systems:

$$\dot{\mathbf{W}}=\sigma\mathbf{P}\mathbf{W}\ \text{with}\ \mathbf{W}^T(0)\mathbf{W}(0)=w_0^2\mathbf{I}\ , \tag{10}$$

$$\dot{\mathbf{P}}=\rho(\mathbf{H}\Leftrightarrow\mathbf{H}^T)\ \text{with}\ \mathbf{P}^T(0)=\Leftrightarrow\mathbf{P}(0)\ . \tag{11}$$

Due to the structure of the sub-system (11), it is straightforward to see that for all $t\geq 0$ the property $\mathbf{P}^T=\Leftrightarrow\mathbf{P}$ (skew-symmetry) holds true. Besides, the following result holds:

Theorem 2. *Consider the dynamical system $d\mathbf{W}/dt=\sigma\mathbf{P}\mathbf{W}$ where $\mathbf{P}$ is skew-symmetric and σ is a real-valued scalar, if $\mathbf{W}(0)\in\mathcal{N}_{w_0}^{p\times q}$ for some w_0 real constant, then the flow $\mathbf{W}(t)$ is Stiefel.*

Proof. The matrix $\mathbf{P}$ is always skew-symmetric, that means $\mathbf{P}^T=\Leftrightarrow\mathbf{P}$. With some algebra it can be found that $d(\mathbf{W}^T\mathbf{W})/dt=\sigma\mathbf{W}^T(\mathbf{P}^T\Leftrightarrow\mathbf{P})\mathbf{W}=\mathbf{0}$, therefore $\mathbf{W}^T(t)\mathbf{W}(t)=\mathbf{W}^T(0)\mathbf{W}(0)=w_0^2\mathbf{I}$. □

From the above Theorem it follows that the algorithm (10) together with (11) is a SOC one. It remains to show that systems (10)+(11) and (9) are equivalent.

Differentiating equation (10) with respect to the time yields $\ddot{\mathbf{W}} = \sigma\dot{\mathbf{P}}\mathbf{W} + \sigma\mathbf{P}\dot{\mathbf{W}}$. Replacing equations (10) and (11) in the preceding formula gives:

$$\mathbf{W}^T\ddot{\mathbf{W}} = \sigma\rho\mathbf{W}^T(\mathbf{H}\Leftrightarrow\mathbf{H}^T)\mathbf{W} + \sigma^2\mathbf{W}^T\mathbf{P}^2\dot{\mathbf{W}} \ .$$

Finally, by subtracting $(\mathbf{W}^T\ddot{\mathbf{W}})^T$ from $\mathbf{W}^T\ddot{\mathbf{W}}$ (observe that $\mathbf{P}^2$ is symmetric), the rule (9) arises.

3.1 Mechanics-type control

In [3] it was shown how to interpret the (10) and (11) as equations describing the dynamics of a rigid system of masses moving in an abstract space under a potential energy field and a viscous braking effect, providing that:

$$\mathbf{H} = \Leftrightarrow\left(\kappa\frac{\partial U}{\partial\mathbf{W}} + \mu\mathbf{P}\mathbf{W}\right)\mathbf{W}^T \ , \tag{12}$$

where U is a function bounded below to be minimized under the restriction $\mathbf{W} \in \mathcal{N}^{p\times q}_{w_0}$ and $\mu \geq 0$, $\kappa > 0$.

It would be interesting to study the existence of special equilibrium points for the second-order system (10)+(11) when the control matrix (12) is used, in those cases when $U = U(\mathbf{W})$ only.

Theorem 3. *Let $\mathcal{S}$ be the dynamic system (10)+(11) where the control matrix is assumed as in (12), the initial state is chosen so that $\mathbf{W}(0) \in \mathcal{N}^{p\times q}_{w_0}$ and $\mathbf{P}(0)$ is skew-symmetric. Define the matrix function $\mathbf{F}(\mathbf{W}) \overset{\text{def}}{=} \Leftrightarrow\kappa\frac{\partial U}{\partial\mathbf{W}}$. A matrix $\mathbf{W}_\star$ is a stationary point of $\mathcal{S}$ if $\mathbf{F}^T(\mathbf{W}_\star)\mathbf{W}_\star$ is diagonal and $\mathbf{P}(\mathbf{W}_\star) = \mathbf{0}$.*

Proof. Using the Lagrangian function:

$$J_1(\mathbf{W}) \overset{\text{def}}{=} J(\mathbf{W}) + \tfrac{1}{2}\mathrm{tr}[(\mathbf{W}^T\mathbf{W}\Leftrightarrow w^2\mathbf{I})\mathbf{E}] \ ,$$

where $J = \kappa U$ and $\mathbf{E}$ is a diagonal matrix containing Lagrange multipliers, recalling that at each instant $\mathbf{W}^T\mathbf{W} = w_0^2\mathbf{I}$, it can be proved that the standard Kuhn-Tucker condition for the existence of an (orthonormal) extreme of U reads $\Leftrightarrow\mathbf{F} + \mathbf{W}\mathbf{E} = \mathbf{0}$, where $w_0^2\mathbf{E} = \mathbf{W}^T\mathbf{F}$, therefore $w_0^2\mathbf{F} = \mathbf{W}(\mathbf{W}^T\mathbf{F})$. This equation implies:

$$w_0^2(\mathbf{F}\mathbf{W}^T\Leftrightarrow\mathbf{W}\mathbf{F}^T) = \mathbf{W}(\mathbf{W}^T\mathbf{F}\Leftrightarrow\mathbf{F}^T\mathbf{W})\mathbf{W}^T \ .$$

By hypothesis $\mathbf{P}(\mathbf{W}_\star) = \mathbf{0}$ and $\mathbf{D}_\star \overset{\text{def}}{=} \mathbf{F}^T(\mathbf{W}_\star)\mathbf{W}_\star = \mathbf{D}_\star^T$, thus from equation (12) it follows that matrix $\mathbf{H}_\star = \mathbf{F}(\mathbf{W}_\star)\mathbf{W}_\star^T$ is symmetric. □

Another fundamental feature of the system (10)+(11)+(12) is its asymptotic stability. Note that system (9) is represented by the pair of coupled systems (10) and (11), whose state-matrices $\mathbf{P}$ and $\mathbf{W}$ form an unique extended state $\mathbf{X} = (\mathbf{P}, \mathbf{W})$.

Theorem 4. *Be U a real-valued, bounded below function of $\mathbf{W}$, with a minimum in $\mathbf{W}_\star$. Then the equilibrium state $\bar{\mathbf{X}} = (\mathbf{0}, \mathbf{W}_\star)$ of the system $\mathcal{S}$ obtained by gathering (10)+(11)+(12) is asymptotically stable.*

Proof. Due to space limitations, the proof, based on the Lyapunov stability theory, is here omitted. The proof of a slightly less general result may be found in [5]. □

Note that function $U(\mathbf{W})$ may have more than one minimum (local minima) corresponding to local maxima of $\Leftrightarrow U(\mathbf{W})$. Besides, note that *the local extreme for which $\mathbf{P} = \mathbf{0}$ depends not only upon $\mathbf{W}(0)$, but also on $\mathbf{P}(0)$.*

4 Relationships with other theories

In this part, relationships among the SOC theory and other learning theories known from the literature are briefly discussed.

4.1 The Moreau-Pesquet's theory

The main theoretical instrument for developing a Principal Component Analysis theory in connection with the SOC one, is the definition of *semicontrast* functions specialized for orthonormal mixtures. Said $\mathcal{S}_d$ the set of *uncorrelated* random source vectors with n entries, and $\mathcal{M}_d$ the set of random vectors $\mathbf{x} = \mathbf{R}\mathbf{s}$ where $\mathbf{s} \in \mathcal{S}_d$ and $\mathbf{R} \in \mathcal{N} \stackrel{\text{def}}{=} \mathcal{N}_1^{n \times n}$. Denote with $\mathcal{P}$ the subset of $\mathcal{N}$ containing matrices $\mathbf{R}$ satisfying the *independence property*: $\mathbf{R} = \mathbf{D}\mathbf{Q}$ where $\mathbf{D} = \mathrm{diag}(\pm 1, \ldots, \pm 1)$ and $\mathbf{Q}$ is a permutation matrix. Then, the concept of *semicontrast* is defined [8] as follows:

Definition 5. *A semicontrast on $\mathcal{M}_d$ is a multivariate mapping φ from the set $\mathcal{M}_d$ to $\mathcal{R}$ which satisfies the following three requirements:*

M1. $\forall \mathbf{s} \in \mathcal{M}_d,\ \forall \mathbf{R} \in \mathcal{P},\ \varphi(\mathbf{R}\mathbf{s}) = \varphi(\mathbf{s})$;

M2. $\forall \mathbf{s} \in \mathcal{S}_d,\ \forall \mathbf{R} \in \mathcal{N},\ \varphi(\mathbf{R}\mathbf{s}) \leq \varphi(\mathbf{s})$;

M3. $\forall \mathbf{s} \in \mathcal{S}_d,\ \forall \mathbf{R} \in \mathcal{N},\ \varphi(\mathbf{R}\mathbf{s}) = \varphi(\mathbf{s}) \Leftrightarrow \mathbf{R}\mathbf{s} \in \mathcal{S}_d$.

Suppose that a neural network described by $\mathbf{y} = \mathbf{W}^T\mathbf{x}$ is used for decorrelating random signals drawn by orthogonal mixtures of uncorrelated source waves. An example of semicontrast function evaluated on $\mathbf{y}$ is given in [8] $\varphi_{\mathbf{y}}^f \stackrel{\text{def}}{=} \sum_{i=1}^n f(\sigma_{y_i}^2)$, where f is a strictly convex function from $\mathcal{R}_+$ to $\mathcal{R}$, and the $\sigma_{y_i}^2$ are the variances of the y_i. This function was already used by Watanabe with $f(z) = z\log(z)$ [8].

It is worth to note that a function φ^f depends uniquely upon the matrix variable $\mathbf{W}$, and requirements M1 ⇔M2 ⇔M3 ensures that it has a unique *maximum* (but possible sign and permutation). This example allows to show that in order to perform decorrelation a function $U \propto \Leftrightarrow \varphi$ should be used in (12), where φ is a semicontrast. Since U depends uniquely on the variable $\mathbf{W}$, the obtained system is autonomous, hence Theorems 3 and 4 may be used.

4.2 The Deco-Brauer's theory

In [2] Deco and Brauer introduced the concept of *volume-conserving* neural networks. A volume-conserving 'square' architecture performs a mapping $\mathbf{y} = \mathbf{S}(\mathbf{x})$ between a set of input patterns and a set of output ones characterized by the property $\det\left(\frac{\partial \mathbf{S}}{\partial \mathbf{x}}\right) = 1$. For a linear network with weight matrix $\mathbf{W}$ the condition above becomes $\det(\mathbf{W}) = 1$. The same condition ensures the bijectivity of the mapping. Volume conservation allows for example the preservation of the entropy $H(\mathbf{x})$ when the input signal is mapped into the output signal whose entropy is $H(\mathbf{y})$, that means $H(\mathbf{x}) = H(\mathbf{y})$, in fact:

$$H(\mathbf{y}) = H(\mathbf{x}) + E[\log|\det(\partial\mathbf{S}/\partial\mathbf{x})|] \ .$$

Symbol $E[\cdot]$ represents the *mathematical Expectation* operator. A SOC algorithm with matrix $\mathbf{W} \in \mathcal{N}_1^{n\times n}$ for a linear network is always volume-conserving, indeed $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ implies $\det(\mathbf{W}) = \pm 1$.

4.3 The Laheld-Cardoso's 'PFS' approach

System (10)+(11) allows *equivariant* square blind source separation by orthonormal mixtures. Strictly speaking, 'equivariant' [7] means that the performance of the separation is independent of the mixing matrix, especially of its conditional number (eigenvalues' spread). This means that a *parameter free separator* (PFS) can manage well-conditioned problems as well as ill-conditioned ones. Formally, let $\mathbf{A}$ be the orthonormal square mixing matrix of the mixing model $\mathbf{x} = \mathbf{A}\mathbf{s}$, where $\mathbf{s}$ contains the source signals, and $\mathbf{T} \overset{\text{def}}{=} \mathbf{W}\mathbf{A}$ the source-to-output transference matrix, where $\mathbf{W}$ is the weight matrix of the neural net $\mathbf{y} = \mathbf{W}\mathbf{x}$. It is easy to check that:

Property 6. *If the control matrix* $\mathbf{H}$ *depends uniquely on the output* $\mathbf{y} = \mathbf{W}\mathbf{x}$ *then the system (10)+(11) is PFS.*

Proof. In this case equations rewrite:

$$\dot{\mathbf{T}} = \sigma\mathbf{P}\mathbf{T} \ , \ \dot{\mathbf{P}} = \rho[\mathbf{H}(\mathbf{T}\mathbf{s}) \Leftrightarrow \mathbf{H}^T(\mathbf{T}\mathbf{s})] \ , \tag{13}$$

thus the system's behavior depends on $\mathbf{T}$ and $\mathbf{s}$ only. □

Changing $\mathbf{A}$ is equivalent to change the starting point $\mathbf{T}(0)$ (and $\mathbf{P}(0)$) of the algorithm. For knowing details about the PFS approach, readers please refer to [7] and references therein.

5 Re-derivation of the Oja's Subspace Rule

The Oja's Principal Subspace Rule [9] has been studied extensively for many years. Probably, the deepest and most complete and lighting results about it have been obtained by Yan, Helmke and Moore in [12]. Here we aim to re-derive the same subspace rule by the Xu's algorithm discussed above.

Start by post-multiplying gradient $\partial \varrho / \partial \mathbf{W}$ in (4) by $\mathbf{W}^T \mathbf{C}_x \mathbf{W}$. The obtained system is described by the following differential equation:

$$\frac{d\mathbf{W}}{dt} = \frac{1}{\beta}[\mathbf{I} \Leftrightarrow \alpha \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T]\mathbf{C}_x\mathbf{W} \ . \tag{14}$$

The new algorithm (14) no longer allows the orthonormal evolution of the state-matrix $\mathbf{W}$. This consequence can be easily shown by observing that:

$$\mathbf{W}^T \frac{d\mathbf{W}}{dt} = \frac{1}{\beta(\mathbf{W})}[1 \Leftrightarrow \alpha(\mathbf{W})](\mathbf{W}^T\mathbf{C}_x\mathbf{W}) \ ,$$

whereby it is easily seen that:

$$\frac{d(\mathbf{W}^T\mathbf{W})}{dt} = 2\frac{1 \Leftrightarrow \alpha}{\beta}(\mathbf{W}^T\mathbf{C}_x\mathbf{W}) \ .$$

Even if for $t = 0$ the $\mathbf{W}^T\mathbf{W} = w_0^2\mathbf{I}$ holds true, for $t > 0$ the product matrix $\mathbf{W}^T\mathbf{C}_x\mathbf{W}$ does not result a scaled identity nor a diagonal. Thus rule (14) is not interesting to our concerns.

We now propose a useful modification of the system (14) that leads to a surprising result. Firstly, rewrite equation (14) in a different way, that is:

$$\frac{d\mathbf{W}}{dt} = \frac{\alpha}{\beta}[\alpha^{-1}\mathbf{C}_x \Leftrightarrow \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{C}_x]\mathbf{W} \overset{\text{def}}{=} \hat{\gamma}\mathbf{Z}\mathbf{W} \ , \tag{15}$$

where $\mathbf{Z}(\mathbf{W}) \overset{\text{def}}{=} \alpha^{-1}(\mathbf{W})\mathbf{C}_x \Leftrightarrow \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{C}_x$ and $\hat{\gamma}(\mathbf{W}) \overset{\text{def}}{=} \alpha(\mathbf{W})/\beta(\mathbf{W})$. The *new* system defined by:

$$\frac{d\mathbf{W}}{dt} = \sigma(\mathbf{Z} \Leftrightarrow \mathbf{Z}^T)\mathbf{W} = \sigma\mathbf{P}\mathbf{W} \ , \tag{16}$$

is Stiefel, moreover, recalling that $\mathbf{C}_x$ is symmetric, there holds:

$$\begin{aligned} \mathbf{P} &= \alpha^{-1}\mathbf{C}_x \Leftrightarrow \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{C}_x \Leftrightarrow \alpha^{-1}\mathbf{C}_x + \mathbf{C}_x\mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T \\ &= \mathbf{C}_x\mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T \Leftrightarrow \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{C}_x \ . \end{aligned}$$

If for the algorithm an initial condition $\mathbf{W}(0)$ such that $\mathbf{W}^T(0)\mathbf{W}(0) = w_0^2\mathbf{I}$ is assumed for some real constant w_0, it always remains $\mathbf{W}^T\mathbf{W} = w_0^2\mathbf{I}$. Replacing this result in the equation (16), its properties do not change, since $\mathbf{P}$ still remains skew-symmetric, therefore the above equation reads $w_0^2\mathbf{P} = \mathbf{C}_x\mathbf{W}\mathbf{W}^T \Leftrightarrow \mathbf{W}\mathbf{W}^T\mathbf{C}_x$. Post-multiplying both members of the latter equation by $\mathbf{W}$ yields:

$$w_0^2\mathbf{P}\mathbf{W} = \mathbf{C}_x\mathbf{W}\mathbf{W}^T\mathbf{W} \Leftrightarrow \mathbf{W}\mathbf{W}^T\mathbf{C}_x\mathbf{W} = (w_0^2\mathbf{I} \Leftrightarrow \mathbf{W}\mathbf{W}^T)\mathbf{C}_x\mathbf{W} \ , \tag{17}$$

therefore the *equivalent* form of the system (16) is:

$$\frac{d\mathbf{W}}{dt} = \frac{\sigma}{w_0^2}(w_0^2\mathbf{I} \Leftrightarrow \mathbf{W}\mathbf{W}^T)\mathbf{C}_x\mathbf{W} \ . \tag{18}$$

This closely recalls the continuous-time version of the well-known Oja's Principal Subspace estimation rule [12]. In turn, learning rule (18), attempts to maximize a simple quadratic cost function under the constraint of orthonormality. In fact, it is well-known that the Oja's rule (18) arises from the GSA maximization of the function:

$$T(\mathbf{W}) = \mathrm{tr}[\mathbf{W}^T \mathbf{C}_x \mathbf{W}] + \tfrac{1}{2}\mathrm{tr}[(\mathbf{W}^T \mathbf{W} \Leftrightarrow w_0^2 \mathbf{I})\mathbf{E}] \ ,$$

where $\mathbf{E}$ is a diagonal matrix containing the Lagrange multipliers for the constraints. The initial state has to belong to $\mathcal{N}^{p \times q}$.

References

[1] P. Comon, *Independent Component Analysis, A New Concept ?*, Signal Processing, Vol. 36, pp. 287 – 314, 1994

[2] G. Deco and W. Brauer, *Nonlinear Higher-Order Statistical Decorrelation by Volume-Conserving Neural Architectures*, Neural Networks, Vol. 8, No. 4, pp. 525 – 535, 1995

[3] S. Fiori, *Unsupervised Neural Artificial Learning Models for Blind Signal Processing*, M.Sc. Dissertation, Dept. of Electronics and Automatics, Univ. of Ancona (Italy), July 1996 (in italian)

[4] S. Fiori, A. Uncini and F. Piazza, *Mechanical Learning Applied to Principal Component Analysis and Source Separation*, Artificial Neural Networks, pp. 571 – 577, 1997, Springer-Verlag

[5] S. Fiori and F. Piazza, *Orthonormal Strongly-Constrained Neural Learning*, Proc. of IJCNN, pp. 2332 – 2337, 1998

[6] C. Fyfe, *A General Exploratory Projection Pursuit Network*, Neural Processing Letters, Vol. 2, No. 3, 1995

[7] J.F. Cardoso and B. Laheld, *Equivariant Adaptive Source Separation*, IEEE Trans. on Signal Processing, Vol. 44, No. 12, pp. 3017 – 3030, Dec. 1996

[8] E. Moreau and J.C. Pesquet, *Independence/Decorrelation Measures with Application to Optimized Orthonormal Representations*, Proc. of ICASSP, pp. 3425 – 3428, 1997

[9] E. Oja, *Neural Networks, Principal Components, and Subspaces*, International Journal of Neural System, Vol. 1, pp. 61 – 68, 1989

[10] F. Palmieri, J. Zhu and C. Chang, *Anti-Hebbian Learning in Topologically Constrained Linear Networks: A Tutorial*, IEEE Trans. Neural Networks, Vol. 4, No. 5, pp. 748 – 761, 1993

[11] L. Xu, *Theories for Unsupervised Learning: PCA and Its Nonlinear Extension*, Proc. of IJCNN, pp. 1252 – 1257, 1994

[12] W.-Y. Yan, U. Helmke, and J.B. Moore, *Global Analysis of Oja's Flow for Neural Networks*, IEEE Trans. Neural Networks, Vol. 5, No. 5, pp. 674 – 683, 1994