A NEW UNSUPERVISED NEURAL LEARNING RULE FOR ORTHONORMAL SIGNAL PROCESSING

Simone Fiori[‡], *Paolo Campolucci, Aurelio Uncini, Francesco Piazza* *

Dipartimento di Elettronica e Automatica – Università di Ancona
Via Brecce Bianche, 60131, An (Italy)
[‡]E-mail:simone@eealab.unian.it

ABSTRACT

We derive a new class of neural unsupervised learning rules which arises from the analysis of the dynamics of an abstract mechanical system. The corresponding algorithms can be used to solve several problems in Digital Signal Processing area, where orthonormal matrices are involved. We present an application which deals with blind separation of sources, i.e. a new method to perform efficient Independent Component Analysis (ICA) of random signals.

1. INTRODUCTION

Some neural learning processes can be viewed as searching processes in a proper state–space. For instance, if we examine the adaptation behavior of linear neural networks, as those commonly used to perform whitening [6], linear compression [7] and blind separation [1, 2, 3], with input $x \in \mathcal{R}^p$, output $y \in \mathcal{R}^n$, and described by $y = W^t x$, we can view it as a search in a matrix state–space M(p,n) in which it moves a point represented by the weight–matrix W. Let us suppose that we know in advance that the steady–state solution (or the target of searching) *belongs to a subset* Ω(p,n) of M(p,n). Let us indicate with $\mathbf{T} = \{W(t)\}$ a generic trajectory in the state–space, corresponding to a generic searching process, for t belonging to a proper temporal interval $[\tau_1, \tau_2]$.

It is clear that $\mathbf{T}$ entirely lies on $\mathbf{M}$, but not necessarily all elements of $\mathbf{T}$ belong to $\mathbf{\Omega}$. It means that it exists a non–empty set $\mathbf{D_T} = (\mathbf{M} - \mathbf{\Omega}) \cap \mathbf{T}$ containing many wrong searching–step. The nonnegative power of $\mathbf{D}$ measures the waste of time–steps in $[\tau_1, \tau_2]$.

It follows that a learning rule, which a trajectory search $\mathbf{T}$ such that $\mathbf{D_T} = \emptyset$ pertains to, is more efficient than a learning rule that does not show this property. Now, the problem is to find such a strongly–constrained adaptation rule.

We present a solution to this problem in the special case where Ω is a set of *orthonormal* matrices (i.e. rectangular matrices N satisfying $N^t N = I$) based on a new neural learning theory derived from the analysis of the dynamics of an abstract mechanical system.

NOTATIONS. The following notations are used in this paper: the superscript ^{t} denotes the usual transposition; I is an identity matrix; $E[\cdot]$ denotes the statistical expectation of the argument; $tr(\cdot)$ denotes the trace, i.e. the sum of the in-diagonal elements of the square matrix in the argument; $diag(d_1, d_2, \ldots, d_n)$ returns a $n \times n$ matrix whose i:th in-diagonal element is d_i; finally, the column rank of a rectangular $m \times n$ matrix, with $m > n$, is denoted by $rk[\cdot]$.

2. THE MEC 'NEURAL ENGINE'

In this section we shall derive some equations describing the dynamics of a *rigid* system of masses, and then we shall show how to use it in the neural network context.

Let $\mathcal{S}^* = \{[2m_i, w_i]\}$ be a system of n masses $2m_i$ in an abstract space $\mathcal{R}^p$, with $n \leq p$. Masses lie on mutually orthogonal axes and are positioned at a fixed (unity) distance from the origin, so that the position vectors w_i are such that $w_i^t w_k = \delta_{i,k}$, for $i, k \in \{1, 2, \ldots, n\}$ where $\delta_{i,k}$ is the Kronecker's delta, while we suppose negligible the entity of the masses of the rigid axes. Masses $2m_i$ are positive, distinct, real numbers. The space $\mathcal{R}^p$ contains an isotropic and homogeneous fluid characterized by a non–null viscosity coefficient μ. Moreover, the abstract space contains a free point $\mathcal{P}$ with a negligible mass, whose coordinates are the p components of a vector x, and which exerts the forces F_i on the masses i:th, $i = 1, 2, \ldots, n$, producing the global motion of the system. Let the origin $\mathcal{O}$ of the axes be a *fixed* point in the space $\mathcal{R}^p$.

*This research was supported by the Italian MURST. Please send comments and suggestions to the first author.

The earlier assumptions imply that the masses can only rotate around $\mathcal{O}$ and they never can translate with respect to it. It follows that the system $\mathcal{S}^*$ is dinamically equivalent to the adjoint *symmetric* system $\mathcal{S} = \{[m_i, w_i], [m_i, -w_i]\}$ that is symmetric in the sense that if an external force F_i is applied to the mass m_i in w_i then a force $-F_i$ is applied to the mass m_i in $-w_i$. In fact, the resulting force acting to the system is null, therefore it never translates.

Thanks to the preceding observations, in the following we can treat the adjoint system $\mathcal{S}$ instead of the original system $\mathcal{S}^*$. Dynamic equations describing the motion of such an abstract mechanical system are derived in the following:

Theorem 1 *Let $\mathcal{S}$ be the system of* n *masses in $\mathcal{R}^p$ described above, and let:*

$$F = \begin{bmatrix} F_1 & F_2 & \cdots & F_n \end{bmatrix},$$
$$W = \begin{bmatrix} w_1 & w_2 & \cdots & w_n \end{bmatrix},$$
$$M = diag(m_1, m_2, \ldots, m_n).$$

At any time the matrix of the instantaneous positions W *satisfies:*

$$\frac{d\,W}{d\,t} = \Omega W \tag{1}$$

where $\Omega \in M(p,p)$ plays the role of an angular speed tensor and its evolution is described by:

$$\frac{d\Omega}{dt} = \frac{1}{4}(F+P)M^{-1}W^t - \frac{1}{4}WM^{-1}(F+P)^t . \tag{2}$$

where:

$$P = -\mu\Omega W . \tag{3}$$

Proof. The proof here is omitted because of space limitations. □

The main point is that Ω is a skew–symmetric matrix, so, from equation (1), it follows that $W(t)$ remains an orthonormal matrix (i.e. it holds that $W^tW = I$) at any time t if it was at time $t = 0$.

Equations (1), (2) and (3) are fixed but the forcing terms contained in F can be arbitrarily chosen. We assume F derived by a *potential energy function* (p.e.f.) $\mathcal{U}$, which depends on W and on x, in the sense that:

$$F := -2\frac{\partial\,\mathcal{U}}{\partial\,W} . \tag{4}$$

Equations (1 – 2) can be directly interpreted as a couple of adaptation rules for a linear neural network described by $y = W^tx$, with p inputs and n outputs, able to perform a generic orthonormal signal processing. For this reason we call it '*neural engine*' and we refer to it as *MEC–net.*

It is here important to notice that the choice of a potential energy function with which customize the above neural system results widely free, therefore we can refer to the set of learning rules that can be obtained by particularizing the above fixed and free adaptation equations as a whole *class of neural learning rules.*

3. EQUILIBRIUM PROPERTIES – DISCRETIZATION

Due to its dissipative–nature, the equilibrium of such a system is described by the following:

Proposition 2 *If the potential energy $\mathcal{U}$ is chosen so that the dynamics is asymptotically stable and the initial state of the system is orthonormal, then the system reaches its equilibrium states when $\mathcal{U}$ is at its orthonormal-constrained minima.* □

It follows that such property can be used to properly chose the form of $\mathcal{U}$ to obtain a desired behavior for the MEC–net. For instance, $\mathcal{U}$ can be assumed as a properly convex cost function to be minimized.

For software–like application purposes, continuous–time equations (1 – 2) must be discretized. This operation must be conducted with some cautions. We found an appropriate method to do this, i.e. a discretization which maintains unaltered the property of orthonormality of W at any computation step.

4. APPLICATION TO BLIND SEPARATION

Although the MEC–net approach can be used in many digital processing problems, in this paper we present an application of the new algorithm to the blind separation of source signals [3]. The problem is to separate n unknown independent source signals from their linear full–rank over–determined mixture observed by a set of sensors. Formally, if s is a n-components vector containing such independent source signals and z is a p-components vector containing observed signals from sensors, it holds that $z = A^ts$, where A^t is said *mixing matrix* and it is such that $rk[A] = n$. It is well known [2, 6, 7] that pre-processing z with a so-called *standardizing* (or *whitening*) stage, described by $x = V^tz$, such that $E[xx^t] = \sigma^2 I$, with $\sigma > 0$ fixed, the original source signals can be recovered from the new mixture x by means of an *orthonormal* separation matrix W^t, i.e. at the equilibrium it results that:

$$y = W^tx = DPs , \tag{5}$$

with D an indeterminable scaling matrix and P an indeterminable permutation matrix [3]. Such an indeterminateness is due to the blind-nature of the problem. In practice, without any a-priori information about sources, we cannot recover the exact ordering or the exact power of the source signals.

Probably, the simplest way to standardizing the random signal z is to adopt the following whitening matrix: $V := \Pi\Delta^{-1}$, where the pair (Π, Δ) is a PCA ([2]) of z such that $\Pi^t\Pi = I$. The *Principal Component Analysis* is a well-known statistical analysis technique, and in literature several algorithms are known to perform it. As we shall explain later, such a linear transformation can be easily realized with a MEC-net sub-system, too.

Clearly, the last separation operation can be performed using a MEC network when an appropriate energy function is chosen. For this purpose we state the following, based on the basic concept of *discriminant contrast function* defined by P. Comon in [2],

Proposition 3 *If Ψ is a discriminant contrast function over a set $\mathcal{H}$ of random signals, choosing $\mathcal{U} = \Psi$, the MEC algorithm can be able to separate any n–tuple of source signal in $\mathcal{H}$ from an arbitrary mixture with mixing matrix A, providing that the condition stated in the* Proposition 2 *and the resolvability condition $rk[A] = n$ are verified.* □

For instance, providing that all the source signals are either leptokurtic or platikurtic (except for one allowed to be gaussian) and are endowed with a symmetrical probability density function, the discriminant contrast function:

$$\Psi(W, x) := \pm\frac{1}{4}\sum_{i=1}^{n} y_i^4 \ , \tag{6}$$

is adequate. Notice that such a function is convex; it is called *Comon's discriminant contrast function.*

Therefore, it can be assumed as potential energy function $\mathcal{U}$ for the MEC-net, allowing the blind separation of the sources as stated by the Proposition 2. The corresponding force F has the following expression:

$$F = \mp 2xCb^t[W^t x] \ , \tag{7}$$

where the $Cb[\cdot]$ operator is defined by:

$$Cb\left(\begin{bmatrix} a & b & c & \cdots \end{bmatrix}^t\right) = \begin{bmatrix} a^3 & b^3 & c^3 \cdots \end{bmatrix} \ .$$

We call this forcing term *C–Force.*

As we mentioned earlier, it is here interesting to notice that the standardization of a random vector can be viewed as the composition of two elementary operations: a PCA optimal compression followed by a simple scaling. Moreover, a PCA processing can be performed by a linear transformation which is orthonormal [5,7], therefore, in a neural context, it can be implemented by a MEC-net endowed with a proper p.e.f.

We found a simple adequate quadratic p.e.f. which leads to a forcing term that we call *H–Force* because it recalls the well-known *hebbian adapting term.* Even if it is not the aim of the paper to present the whole MEC-net's theory, for the sake of completeness we explain the above result in the following.

Let N_p be a linear neural network with n outputs described by $v = T^t u$, that must be trained so that, at the equilibrium, the pair $(T, \cdot)$ is a PCA of the input signal u and let (Π, Δ) an orthonormal PCA of u. By definition of PCA, if v has less components than u, the transformed signal v must retain the major fraction of the power of u compatibly with the loss in dimensionality. Therefore, a proper learning rule for the weight-matrix T can be obtained by maximizing the functional:

$$\Phi(T, u) := +\frac{1}{2}\sum_{i=1}^{n} v_i^2 \ , \tag{8}$$

with respect to T, under a proper consistency constraint, for instance under the constraint that $T^tT = I$.

Unfortunately, the non-strongly constrained maximization of the above objective function leads only to an orthonormal linear combination of the eigenvectors in Π or, in other words, the network N_p becomes a *Principal Subspace Analizer* ([4]) of the input signal, instead of a Principal Component Analyzer.

Into the MEC environment this drawback disappears, and assuming the function (8) with the sign changed, $\mathcal{U} := -\Phi$, as a p.e.f. it allows the algorithm to determine an orthonormal PCA of the signal (with eigenvalues *not* ordered as in Δ). The corresponding forcing term has the following expression:

$$F = 2uv^t \ , \tag{9}$$

and this explains because we refer us to it as an Hebbianic Force.

About the convergence properties of the MEC with the above p.e.f., more formally we state that the following proposition holds:

Proposition 4 *Let* u *be a real-valued, zero-mean random signal with a finite covariance, and $v = T^t u$ a neural network with the MEC learning rule with the expression (9) as forcing term. If the matrix of the masses M is composed of all-distinct in-diagonal values and the initial state $T(0)$ is orthonormal, at the statistical equilibrium the network becomes a Principal Component Analyzer, i.e., in the mean $(T_\infty, \cdot)$ is an orthonormal PCA of* u. □

5. EXPERIMENTAL RESULTS

As an application, we can use our algorithm to solve a simple blind separation problem with two sources ($n = 2$) and three sensors ($p = 3$). We assume the following mixing matrix:

$$A^t = \begin{bmatrix} 1 & -2 \\ 0 & 3 \\ 1 & -1 \end{bmatrix},$$

and we use, as components of s, a couple of wide–band source signals (zero-mean white noise with flat probability density function) and a couple of narrow–band source signals (sinusoid by arbitrarily chosen frequency), both normalized to have unity powers. The resulting $z = A^t s$ signal at the sensors must be firstly standardized and then orthonormally separated.

The first operation can be performed using a two-stage neural network, described by $x = V^t z$, composed by a linear compressor (e.g. a PCA network [5, 7] or a MEC-net provided with a properly chosen forcing term) and a simple post-scaler; the second can be performed by a MEC-net with a proper forcing term.

Both in the wide–band and in the narrow–band case, we found through simulations that the standardization can be performed by a linear transformation represented by the operator:

$$V^t \cong \begin{bmatrix} \frac{1}{\sqrt{1.30}} & 0 \\ 0 & \frac{1}{\sqrt{14.67}} \end{bmatrix} \begin{bmatrix} 0.448 & -0.570 \\ 0.612 & 0.757 \\ 0.652 & -0.313 \end{bmatrix}^t$$

so that for the standardized signal x the $E[xx^t] = I$ holds true. In particular, we found the second matrix in the product, by running a MEC algorithm provided by an *H–Force*.

Now, since the resolvability condition $rk[A] = 2$ holds, the two source signals can be separated from x by means of a pure rotation through a linear neural transformation described by $y = W^t x$. From calculus we found that such a transformation should be:

$$W^t \cong \begin{bmatrix} 0.96 & 0.25 \\ -0.23 & 0.97 \end{bmatrix}.$$

Because the signals to be separated are either platikurtic we can use the *C–Force* for the MEC-net with the proper sign. Running our algorithm we found the averaged matrix:

$$W^t \cong \begin{bmatrix} 0.97 & 0.24 \\ -0.24 & 0.97 \end{bmatrix}.$$

To those values it corresponds the following averaged *separating product*:

$$DP \cong \begin{bmatrix} 0.984 & 0.019 \\ 0.007 & 1.001 \end{bmatrix}.$$

So, the relative interference residuals are, respectively, of –43dB and –34dB.

6. CONCLUSION

In the paper we presented a new class of semi–general purpose neural learning algorithms, that can provide an efficient method to solve several orthonormal signal processing problems, based on the dynamics of an abstract mechanical system: the MEC environment. Then, we presented an application to blind source signals separation based on a known problem decomposition: this allow us to show two possible ways to use our MEC-net environment, i.e., in linear compression and in orthonormal separation.

7. REFERENCES

[1] A.J. BELL AND T.J. SEJNOWSKI, *An information maximisation approach to blind separation and blind deconvolution.* Neural Computation, Vol.7, No. 6, pp. 1129 – 1159, 1995

[2] P. COMON, *Independent Component Analysis, a new concept ?*, Signal Processing 36, pp. 287 – 314, 1994

[3] P. COMON, C. JUTTEN AND J. HERAULT, *Blind Separation of Sources. Part II: Problems statement*, Signal processing 24, pp 11 – 20, 1991

[4] J. KARHUNEN AND J. JOUTSENSALO, *Learning of Robust Principal Component Subspace*, Proc. of IJCNN, Vol. 3, pp. 2049 – 2412, 1993

[5] J. KARHUNEN. L. WANG AND R. VIGARIO, *Nonlinear PCA type learning for source separation and Independent Component Analysis*, Proceedings of IEEE International Conference on Neural Network (ICNN), pp. 995 – 1000, 1995

[6] 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

[7] L. WANG, J. KARHUNEN AND E. OJA, *A bi-gradient optimization approach for robust PCA, MCA and source separation*, Proceedings of International Joint Conference on Neural Network (IJCNN), pp. 1684 – 1689, 1995