SOME PROPERTIES OF BELL–SEJNOWSKI PDF-MATCHING NEURON

Simone Fiori

DIE–UNIPG, University of Perugia, Italy
E-MAIL: SFR@UNIPG.IT

ABSTRACT

The aim of the present paper is to investigate the behavior of a single-input single-unit system, learning through the maximum-entropy principle, in order to understand some formal property of Bell-Sejnowski's PDF-matching neuron. The general learning equations are presented and two case-study are discussed with details.

1. INTRODUCTION

The analysis of the behavior of adaptive activation function non-linear neurons is a challenging research field in the neural network theory, which may require analyzing non-linear differential equations of neuron's parameters. Especially in signal processing applications, the external excitations are not deterministic but stochastic, and the aim is to find a statistical description of the neural system's response and of system features. The formal techniques known in the scientific literature for studying such systems benefit from cross-fertilization among artificial neural networks, information theory and signal processing and neurobiology.

Recently, several researchers have focused their attention on this class of stochastic learning theories, with applications to blind separation of sources by the independent component analysis [1, 2, 3, 4, 5, 6, 16, 20], probability density estimation [1, 18, 7], self-organizing classification [19], and blind system deconvolution [2, 8, 9]. Also, some studies on neurobiological mechanisms have suggested interesting non-linear models and information-theoretic based learning theories [10, 11, 13, 14, 15].

Following the pioneering work of Linsker, Plumbley, Bell and Sejnowski [2, 12, 17], in recent papers, we presented some results related to the use of flexible non-linear units, termed FANs, trained in an stochastic way by means of an entropy-based criterion: In [7] we proposed some general structures and adapting frameworks for FAN non-linear unit, while papers [4, 5, 6] have been devoted to the application of these neurons to blind signal processing tasks, such as blind source separation by the independent component analysis and blind signal flattening; in these works we also compared the proposed structures to other flexible topologies known in the scientific literature, as e.g. the mixture-of-kernel, showing that the new approach may exhibit better estimation/approximation ability at a lower complexity burden.

The aim of our preceding work was to introduce the new adaptive-activation-function structures and adapting theories and to assess their features through numerical experiments on real-world data; however, due to the strong non-linearity of the involved equations we did not present any theoretical considerations about the mathematical structure and properties of the adapting equations. In the present paper we recall the basic adapting formulas and present the closed-form expressions of them for some special cases; our main goal is to discuss their features in an analytical way, in order to gain a deeper insight into the behavior of the non-linear differential equations governing information-theoretic FAN non-linear unit adapting, and to better explain the previous numerical results. In particular, the aim is to discuss some properties of Bell-Sejnowski probability density function matching neuron.

2. NEURON MODEL AND PDF-MATCHING LEARNING EQUATIONS

In the present paper we consider the simple neuron model depicted in the Figure 1, which may be formally described by the input-output equation:

$$y = s(wx + b) \ , \tag{1}$$

where $x(t) \in \mathcal{R}$ and $y(t) \in \mathcal{R}$ denote the neuron's input stimulus and the neuron's response signal, respectively, $w \in \mathcal{R}$ denotes the neuron's connection strength and $b \in \mathcal{R}$ stands for the bias; the non-linear function $s(\cdot)$ represents a bounded (saturating) squashing activation function, which meets the monotonicity condition $s'(\cdot) > 0$.

Both the input and output signals are treated as stationary stochastic signals, described by the probability density functions (pdfs) $p_x(x)$ and $p_y(y)$. We do not make any particular hypothesis about the stimulus' statistical distribution, but for requiring a sufficient regularity, namely $p_x(x)$ should be a smooth function endowed with sufficiently-high-order moments.

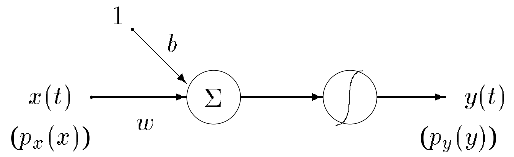

Fig. 1. Structure of single-input neuron.

The statistical distribution of the neuron's response depends upon the distribution of the stimulus through the neuron's non-linear transfer function; formally, the relationship among the two statistics write:

$$p_y(y) = \frac{p_x(x)}{D(x)} \ , \ D(x) \overset{\text{def}}{=} \left| \frac{dy}{dx} \right| = |w|s'(wx+b) \ . \quad (2)$$

Bell-Sejnowski's neuron should learn to respond a maximum-entropy signal, which gives rise to a learning theory based on entropy maximization. The entropy of the excitation and of the response define, respectively, as:

$$\begin{aligned} H_x &\overset{\text{def}}{=} -\int_{\mathcal{R}} p_x(\xi)\log p_x(\xi)d\xi \ , \\ H_y &\overset{\text{def}}{=} -\int_{\mathcal{R}} p_y(\eta)\log p_y(\eta)d\eta \ . \end{aligned}$$

Clearly these entropies are not independent, as they may be related though the statistics transformation formula (2), which gives the relationship:

$$H_y(w,b) = H_x + \int_{\mathcal{R}} p_x(\xi)\log[|w|s'(w\xi+b)]d\xi \ . \quad (3)$$

The neuron's parameters w and b may be learnt through an optimization principle which has the target to maximize neuron's response entropy, namely, through an entropy-gradient learning rule. In order to derive such learning equations the partial derivatives of the response entropy are necessary. Straightforward calculations give:

$$\begin{aligned} \frac{\partial}{\partial w}\int_{\mathcal{R}} p_x(\xi)\log[|w|s'(w\xi+b)]d\xi &= \\ \frac{1}{w} + \int_{\mathcal{R}} p_x(\xi)\frac{s''(w\xi+b)}{s'(w\xi+b)}\xi d\xi \ , & \quad (4) \\ \frac{\partial}{\partial b}\int_{\mathcal{R}} p_x(\xi)\log[|w|s'(w\xi+b)]d\xi &= \\ \int_{\mathcal{R}} p_x(\xi)\frac{s''(w\xi+b)}{s'(w\xi+b)}d\xi \ . & \quad (5) \end{aligned}$$

Note that the entropy of the stimulus does not depend on neuron parameters' values, thus its derivatives are not required.

An interesting observation about maximum-entropy neuron learning is that the neuron tries to *align* its transfer function to the stimulus' pdf [2]. This may be proven formally in the following way. We postulate that the first-order derivative of neuron's transfer function, namely $D(x)$ in (2) tends to match the stimulus pdf $p_x(x)$: To show this it is necessary to define a mismatch measure and to show that it gets minimized. A pseudo-distance among pdfs is the Kullback-Leibler informational divergence, which in this case writes:

$$\begin{aligned} A &\overset{\text{def}}{=} \int_{\mathcal{R}} p_x(\xi)\log\frac{p_x(\xi)}{D(\xi)}d\xi \\ &= -H_x - \int_{\mathcal{R}} p_x(\xi)\log D(\xi)d\xi \ . \end{aligned}$$

From equation (3) it is easily recognized that $A(w,b) = -H_y(w,b)$, therefore as H_y gets maximized, also A gets minimized, and this proves that the neuron-dependent function $D(x)$ tends to approach $p_x(x)$.

The aim of the present paper is to elucidate some properties of the pdf-matching neuron, which descend form the mathematical properties of the above learning equations. In order to carry out our analytical considerations, it is necessary to choose a neuron's structure, namely, to define the shape of the squashing function $s(\cdot)$.

The following expression, which may be regarded as a kind of sigmoidal function, proves to generate tractable mathematics:

$$s(u) = C + B\int_0^u \exp(-v^{n+1})dv \ , \ n \in \mathcal{N} \ ; \quad (6)$$

in the above formula $C \in \mathcal{R}$ and $B > 0$ are arbitrary constants, and n is an odd integer. The Figure 2 shows three examples of the shape of function $s(u)$ for different values of the integer n (the constants C and B have been chosen so that the function always ranges in $[0,1]$).

It is worth noting that the learning equations (4) and (5) do not depend explicitly on the sigmoidal function, but on the ratio $s''(u)/s'(u)$. From equation (6) we find:

$$\frac{s''(u)}{s'(u)} = -(n+1)u^n \ .$$

Ultimately, neuron's learning equations read:

$$\begin{aligned} \frac{dw}{dt} &= \frac{1}{w} - (n+1)\int_{\mathcal{R}} p_x(\xi)(w\xi+b)^n\xi d\xi \ , \quad (7) \\ \frac{db}{dt} &= -(n+1)\int_{\mathcal{R}} p_x(\xi)(w\xi+b)^n d\xi \ . \quad (8) \end{aligned}$$

As anticipated, the selected squashing function gives rise to tractable mathematics, in fact, the required integrals may be computed as follows. To start with, let us define the

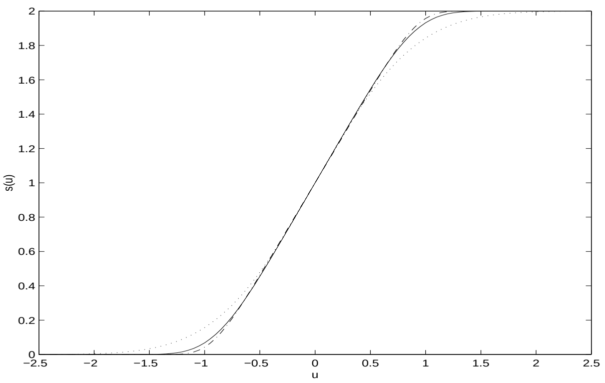

Fig. 2. Sigmoidal function $s(u)$ for three different values of the integer n: Dotted line: $n = 1$; Solid line: $n = 3$; Dot-dashed line: $n = 5$.

stimulus moments:

$$\begin{aligned}\mu_1 &\overset{\text{def}}{=} \int_{\mathcal{R}} p_x(\xi)\xi d\xi \ , \\ \breve{\mu}_m &\overset{\text{def}}{=} \int_{\mathcal{R}} p_x(\xi)(\xi - \mu_1)^m d\xi \ , \ m \geq 0 \ .\end{aligned}$$

Note that, by the hypotheses made on the stimulus, the moments exist at least for some value of m; it is important to remark that $\mu_1 \neq \breve{\mu}_1$ and, in particular, $\breve{\mu}_0 = 1$ and $\breve{\mu}_1 = 0$, always.

By replacing the term $(w\xi + b)^n$ in the integrals with the equivalent term $[w(\xi - \mu_1) + b + w\mu_1]^n$, and by making use of the binomial expansion formula, we have:

$$\int_{\mathcal{R}} p_x(\xi)(w\xi + b)^n \xi d\xi = \sum_{m=0}^{n} T_m^n (\breve{\mu}_{m+1} + \breve{\mu}_m \mu_1) w^m (b + w\mu_1)^{n-m} \ ,$$

and:

$$\int_{\mathcal{R}} p_x(\xi)(w\xi + b)^n d\xi = \sum_{m=0}^{n} T_m^n \breve{\mu}_m w^m (b + w\mu_1)^{n-m} \ ,$$

where T_m^n denotes the binomial coefficient $\frac{n!}{m!(n-m)!}$.

The binomial expansion may be used again in the above formulas, which allows writing the neuron's learning equations in the friendly form:

$$\begin{aligned}\frac{dw}{dt} &= \frac{1}{w} - (n+1) \sum_{m=0}^{n} \sum_{\ell=m}^{n} T_{\ell-m}^{n-m} T_m^n \\ &\quad \times (\breve{\mu}_{m+1} + \breve{\mu}_m \mu_1) \mu_1^{\ell-m} w^\ell b^{n-\ell} \ , \qquad (9) \\ \frac{db}{dt} &= -(n+1) \sum_{m=0}^{n} \sum_{\ell=m}^{n} T_{\ell-m}^{n-m} T_m^n \\ &\quad \times \breve{\mu}_m \mu_1^{\ell-m} w^\ell b^{n-\ell} \ . \qquad (10)\end{aligned}$$

It is also interesting to write explicitly the neuron's entropy gap, the learning criterion defined as $\Gamma_h(w,b) \overset{\text{def}}{=} H_y(w,b) - H_x - \log B$, that reads:

$$\begin{aligned}\Gamma_h(w,b) = \log(|w|) - \sum_{m=0}^{n+1} \sum_{\ell=m}^{n+1} T_{\ell-m}^{n-m+1} T_m^{n+1} \\ \times \breve{\mu}_m \mu_1^{\ell-m} w^\ell b^{n-\ell+1} \ . \qquad (11)\end{aligned}$$

The presence of the term $\frac{1}{w}$ in the learning equation for the connection strength has the meaning of creating a barrier line in the phase-plane $w - b$ of the differential learning system, which precludes the change of the sign of w: If the dynamics starts with $w(0) > 0$ then the connection weight remains excitatory, while $w(0) < 0$ makes the weight being always inhibitory. Of course, the barrier also makes the line $w = 0$ unstable for the system.

3. ANALYSIS FOR A GAUSSIAN EXCITATION

The signal $x(t)$ might be conceived as the 'net' input to the neuron, namely as the linear combination of many input signals, or as the only input to the neuron. In both cases it makes sense to study the learning equations when $x(t)$ has a Gaussian distribution, namely when:

$$p_x(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \ . \qquad (12)$$

In this case we have $\mu_1 = \mu$, $\breve{\mu}_m = 0$ for m odd, and $\breve{\mu}_{2m} = 1 \cdot 3 \cdots (2m-1)\sigma^{2m}$.

For $n = 1$, the sigmoidal function of the neuron coincides to the lifted error-function ('erf'), which has been investigated in a theoretical way in [7] and whose powerful in blind source separation by the independent component analysis has been numerically proven in the papers [5, 6].

The case $n = 3$ has not been considered before and is interesting to study. By particularizing the learning equations (9)+(10), we obtain the differential system governing the neuron's learning phase:

$$\begin{aligned}\frac{dw}{dt} &= \frac{1}{w} - 4[\mu b^3 \mu + 3(\mu^2 + \sigma^2)wb^2 + 3\mu(\mu^2 + \\ &\quad 3\sigma^2)w^2 b + (\mu^4 + 6\sigma^2\mu^2 + 3\sigma^4)w^3] \ , \qquad (13) \\ \frac{db}{dt} &= -4[b^3 + 3\mu w b^2 + (\mu^2 + \sigma^2)w^2 b + \\ &\quad \mu(\mu^2 + 3\sigma^2)w^3] \ . \qquad (14)\end{aligned}$$

One of the purposes of the present analysis is to find, in closed form, the equilibrium points of the above learning equations; this may be achieved by solving the system of two equations that arises by vanishing the right-hand side of

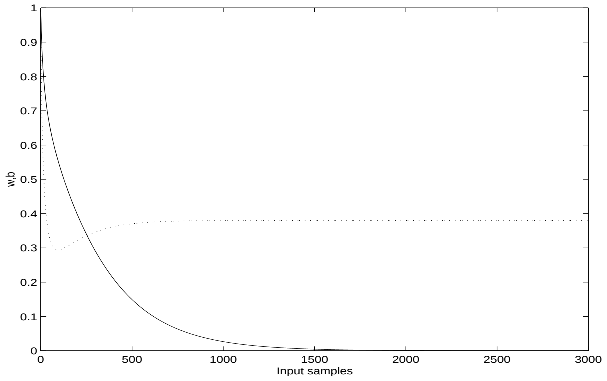

Fig. 3. Case $\mu = 0$: Learning curves for the parameters w (dashed-line) and b (solid-line).

the above learning equations to zero. In this way we obtain:

$$\begin{aligned} 1 &= 4[\mu w b^3 \mu + 3(\mu^2 + \sigma^2) w^2 b^2 + 3\mu(\mu^2 + \\ &\quad 3\sigma^2) w^3 b + (\mu^4 + 6\sigma^2\mu^2 + 3\sigma^4) w^4] \,, \qquad (15) \\ 0 &= b^3 + 3\mu w b^2 + (\mu^2 + \sigma^2) w^2 b + \\ &\quad \mu(\mu^2 + 3\sigma^2) w^3 \,, \qquad (16) \end{aligned}$$

where it is understood that $w \neq 0$ and that $w > 0$.

We have been able to identify two special cases when the above equilibrium equations may be exactly solved.

3.1. Case $\mu = 0$

When the mean value of the input Gaussian excitation is zero, the equilibrium system noticeably simplifies into:

$$\begin{aligned} 1 &= 4(3\sigma^2 w^2 b^2 + 3\sigma^4 w^4) \,, \\ 0 &= b(b^2 + 3\sigma^2 w^2) \,. \end{aligned}$$

Clearly the second equation has $b = 0$ as only feasible solution, because the sub-equation $b^2 + 3\sigma^2 w^2 = 0$ would lead to complex-valued solutions; by vanishing b in the first equation we also find $w = \frac{1}{\sqrt[4]{12}\sigma}$.

As a numerical example, let us consider the case that $\sigma = \sqrt{2}$ and the learning equations have been discretized in time with sampling-step $\Delta t = 0.001$. The Figure 3 shows the course of the slope w and bias b of the neuron's activation during the learning phase; it may be readily verified that $b \to 0$ and $w \to \frac{1}{\sqrt[4]{12}\sqrt{2}}$. Also, Figure 4 shows the true cdf of the input signal and the non-linear transference function of the neuron: They look nearly superimposed and, mainly, the neuron's activation is just aligned to the true cdf.

It would also be interesting to investigate the shape of the entropy-gap as a function of the learnable parameters. The entropy-gap surface and contour plot for the present case are depicted in the Figures 5. The symmetry of the gap $\Gamma_h(w, b)$ about the line $b = 0$, as well as the fact that

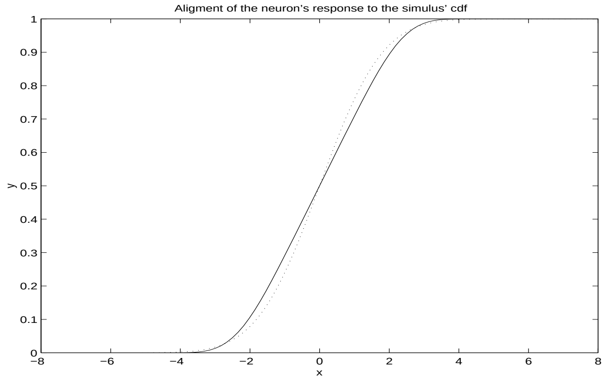

Fig. 4. Case $\mu = 0$: Alignment of the input stimulus cdf (dashed-line) and the neuron's activation function (solid-line).

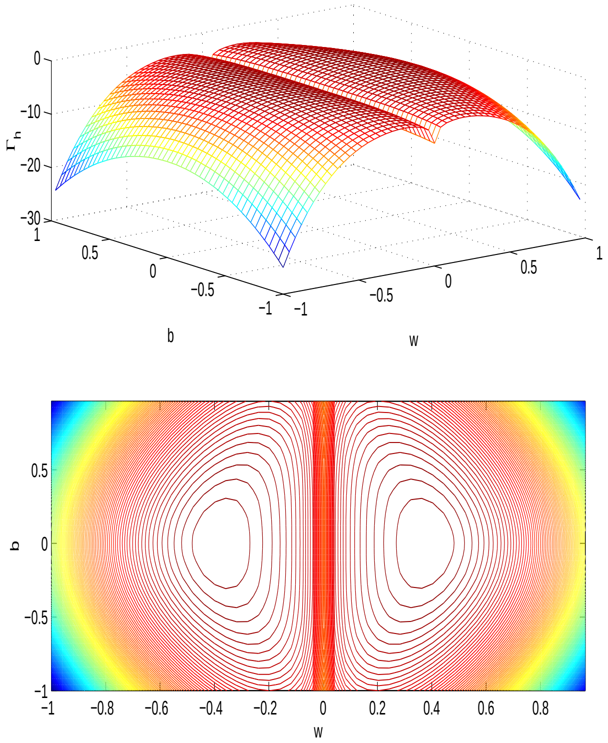

Fig. 5. Entropy-gap surface (top) and contour-plot (bottom) for the case $\mu = 0$.

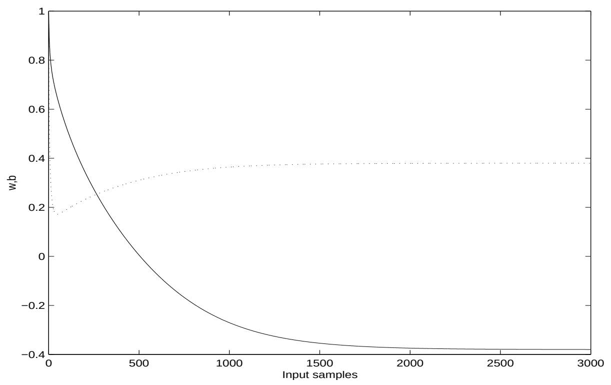

Fig. 6. Case $\mu = 1$: Learning curves for the parameters w (dashed-line) and b (solid-line).

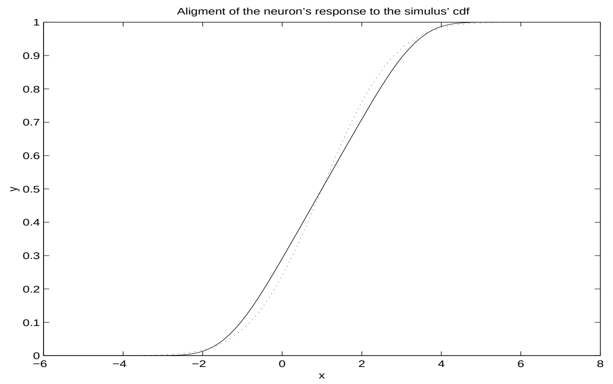

Fig. 7. Case $\mu = 1$: Alignment of the input stimulus cdf (dashed-line) and the neuron's activation function (solid-line).

the minima lie on this line, is quite apparent, confirming the conclusions of the theoretical analysis. The entropy-barrier in correspondence of $w = 0$ is also clearly visible.

3.2. Case $\mu = 1$

The result pertaining to an unitary mean value is really non-trivial and interesting. In fact, when $\mu = 1$ the equation (16) becomes an identity over the line $w + b = 0$, and the equation (15) then gives $w = -b = \frac{1}{\sqrt[4]{12\sigma}}$.

As a numerical example, let us consider again the case that $\sigma = \sqrt{2}$. The Figure 6 shows the course of w and b during the learning phase; it may be readily verified that this time $b \to -\frac{1}{\sqrt[4]{12\sqrt{2}}}$ while $w \to \frac{1}{\sqrt[4]{12\sqrt{2}}}$. Figure 7 shows the cdf of the input signal and the neuron's activation function, which look again nearly superimposed, as predicted by the theory. The shape of the entropy-gap as a function of the learnable parameters, shown in the Figures 8, again confirms that the learning algorithm has found the right neural configuration. The black diagonal line in the figure represents the line of equation $w + b = 0$, and the fact that the minima of the entropy-gap lies on such line clearly emerges.

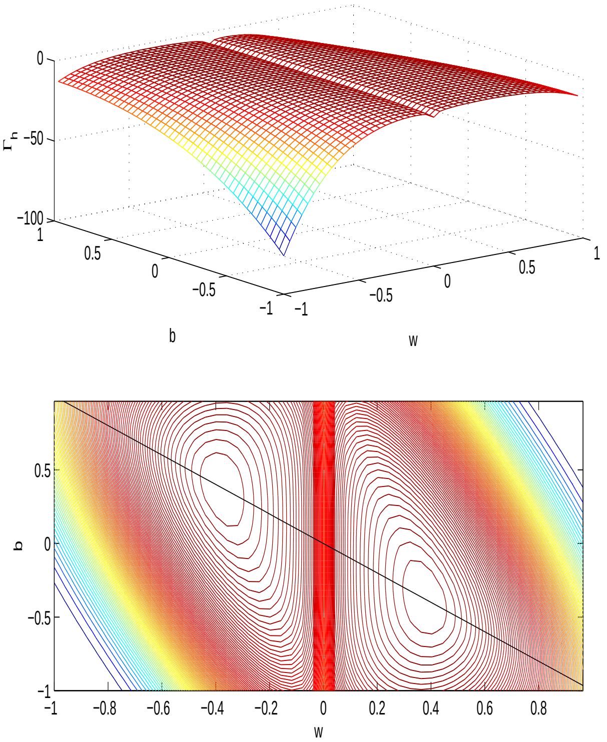

Fig. 8. Entropy-gap surface (top) and contour-plot (bottom) for the case $\mu = 1$.

4. CONCLUSIONS

The aim of the present paper was to discuss with some detail the behavior of a single-weight, single-bias non-linear neuron in presence of a stochastic excitation of known distribution, in order to analytically show that the maximum-entropy learning principle causes the neuron's transference to align to the input cdf as predicted by Bell and Sejnowski. It has been shown by chosen an activation function whose shape is very close to the standard sigmoids but which leads to tractable mathematics, and by considering a Gaussian excitation: In this case the closed-form expression of the learning equations can be analytically computed and their features can be investigated. Also, we illustrated the analytical results through numerical simulations.

5. REFERENCES

[1] S.-I. Amari, T.-P. Chen and A. Cichocki, *Stability Analysis of Learning Algorithms for Blind Source Separation*, Neural Networks, Vol. 10, No. 8, pp, 1345 – 1351, 1997

[2] A.J. Bell and T.J. Sejnowski, *An Information Maximization Approach to Blind Separation and Blind Deconvolution*, Neural Computation, Vol. 7, No. 6, pp. 1129 – 1159, 1996

[3] J.A. Feldman and D.H. Ballard, *Connectionist models and their perspectives*, Computer Science, Vol. 6, pp. 205 – 254, 1982

[4] S. Fiori, *Blind Source Separation by New M–WARP Algorithm*, Electronics Letters, Vol. 35, No. 4, pp. 269 – 270, Feb. 1999

[5] S. Fiori, *Entropy Optimization by the PFANN Network: Application to Independent Component Analysis*, Network: Computation in Neural Systems, Vol. 10, No. 2, pp. 171 – 186, May 1999

[6] S. Fiori, *Blind Signal Processing by the Adaptive Activation Function Neurons*, Neural Networks, Vol. 13, No. 6, pp. 597 – 611, Aug. 2000

[7] S. Fiori and P. Bucciarelli, *Probability Density Estimation Using Adaptive Activation Function Neurons*, Neural Processing Letters, Vol. 13 No. 1, pp. 31 – 42, Feb. 2001

[8] S. Fiori, *Notes on Cost Functions and Estimators for 'Bussgang' Adaptive Blind Equalization*, European Transactions on Telecommunications. Accepted for publication

[9] S. Fiori, *A Contribution to (Neuromorphic) Blind Deconvolution by Flexible Approximated Bayesian Estimation*, Signal Processing. Accepted for publication

[10] M.W. Spratling and G.M. Hayes, *Learning Synaptic Clusters for Nonlinear Dendritic Processing*, Neural Processing Letters, Vol. 11, No. 1, pp. 17 – 27, Feb. 2000

[11] S. Laughlin, *A Simple Coding Procedure Enhances a Neuron's Information Capacity*, Z. Naturforsch, Vol. 36, pp. 910 – 912, 1981

[12] R. Linsker, *Local Synaptic Rules Suffice to Maximize Mutual Information in a Linear Network*, Neural Computation, Vol. 4, pp. 691 – 702, 1992

[13] M.C. Mackey and L. Glass, *Oscilaltion and Chaos in Physiological Control Systems*, Science, pp. 287 – 289, July 1977

[14] V.J. Mathews, *Adaptive polynomial filtering*, IEEE Signal Processing Magazine, pp. 10 – 26, 1991

[15] B.W. Mel, *Information Processing in Dendritic Trees*, Neural Computation, Vol. 6, pp. 1031 – 1085, 1994

[16] D.T. Pham, *Blind separation of instantaneous mixtures of sources based on order statistics*, IEEE Trans. on Signal processing, Vol. 48, No. 2, pp. 363 – 375, Feb. 2000

[17] M.D. Plumbley, *Approximating Optimal Information Transmission Using Local Hebbian Algorithm in a Double Feedback Loop*, Artificial Neural Networks, pp. 435 – 440, Springer-Verlag, 1993

[18] Z. Roth and Y. Baram, *Multidimensional Density Shaping by Sigmoids*, IEEE Trans. on Neural Networks, Vol. 7, No. 5, pp. 1291 – 1298, Sept. 1996

[19] A. Sudjianto and M.H. Hassoun, *Nonlinear Hebbian Rule: A Statistical Interpretation*, Proc. of International Conference on Neural Networks (ICNN'94), Vol. 2, pp. 1247 – 1252, 1994

[20] 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