1996 IEEE INTERNATIONAL SYMPOSIUM ON CIRCUITS AND SYSTEMS
VOL. 2, 157–160. IEEE, PISCATAWAY, NJ, 1996. 96CH35876.

SELF–ADAPTIVE NEURAL NETWORKS FOR BLIND SEPARATION OF SOURCES

Andrzej CICHOCKI [1], *Shun-ichi AMARI*, *Masaharu ADACHI*, *Włodzimierz KASPRZAK*
Institute of Physical and Chemical Research, RIKEN
F.R. Program, Brain Information Processing Group
Hirosawa 2–1, Wako–shi, Saitama 351–01, JAPAN
E–mail: cia@kamo.riken.go.jp

ABSTRACT

Novel on–line learning algorithms with self adaptive learning rates (parameters) for blind separation of signals are proposed. The main motivation for development of new learning rules is to improve convergence speed and to reduce cross–talking, especially for non–stationary signals. Furthermore, we have discovered that under some conditions the proposed neural network models with associated learning algorithms exhibit a random switch of attention, i.e. they have ability of chaotic or random switching or cross–over of output signals in such way that a specified separated signal may appear at various outputs at different time windows. Validity, performance and dynamic properties of the proposed learning algorithms are investigated by computer simulation experiments.

1. INTRODUCTION

Most of known adaptive learning algorithms for blind separation of source signals assume that the global learning rate is a small positive constant, either fixed or exponentially decreasing to zero as time goes to infinity [2–10]. This approach leads usually to relative slow convergence speed or low performance (high cross–talking between outputs signals). Thus it is not suitable for non–stationary signals or time variable mixing parameters.

The objective of this paper is twofold. The main task is to develop on–line learning algorithms with local self adaptive learning parameters in order to provide high performance and high convergence speed for real–time separation of non–stationary source signals [1]. The second task is to investigate some dynamical properties of the learning models interesting from neuro–biological point of view. Especially, we are interested in the mechanism of selective attention and random (transient) switching of attention. In other words, any source signal can appear equally at any output and a specific source signal should appear in different outputs $y_i(t)$ at different time windows. For example in the case of a two channel separation it holds:

$$y_1(t) = \begin{cases} p_{11}s_1(t) & \text{for} \quad t_1 \le t < t_2 \\ p_{12}s_2(t) & \text{for} \quad t_3 \le t < t_4, \end{cases} \tag{1}$$

$$y_2(t) = \begin{cases} p_{22}s_2(t) & \text{for} \quad t_1 \le t < t_2 \\ p_{21}s_1(t) & \text{for} \quad t_3 \le t < t_4, \end{cases} \tag{2}$$

where the p_{ij}–s are scaling factors.

Such switching or cross–over property may not be a desired (satisfying) solution in many engineering applications. On the other hand, from neuro–biological point of view it can help to explain mechanisms in the brain that are associated with selective attention and switching of attention. We give here several examples.

In "cocktail party" problem many persons speak simultaneously. A human being has remarkable ability for selective attention, i.e. the ability to follow one speaker and to diminish surrounding noise. However, a listener may suddenly switch (change) attention from one person to another one and then back to the first person or other one.

Similar problem appears in visual perception. When we look at a composite figure, consisting of many patterns, we usually pay attention to one of them for a while and after that we switch to another pattern. This switching is sudden and usually it has some irregular transient behavior.

Closely related to "cocktail party" problem are models for the olfactory bulb, which perform separation and decomposition of mixed odor inputs from different sources [7]. It is well known that animals need to detect and recognize the odor properties and to localize them in space, in order to hunt for food or to flee from danger. Sometimes it is very crucial for the animals immediately to switch attention from one specific odor to another one, thus finding optimal food or recognizing a dangerous situation.

2. STATEMENT OF THE PROBLEM

Let us assume m stochastically independent source signals $s_j(t)$ $(j = 1,2,\ldots,m)$ are linearly and instantaneously combined via unknown mixing coefficients (parameters) $\mathbf{A} = [a_{ij}] \in R^{n\times m}, (m \le n)$ into n measured (observed) sensor signals [2–6]:

$$x_i(t) = \sum_{j=1}^{m} a_{ij}s_j(t), \tag{3}$$

or in matrix form

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t). \tag{4}$$

It is assumed that the mixing parameters $\{a_{ij}\}$ are fixed or slowly varying in time. Nothing more can be assumed about the unknown mixing matrix $\mathbf{A}$ except that it has the

[1] On leave from Warsaw University of Technology, Department of Electrical Engineering, Warsaw, POLAND.

full rank. The source signals are assumed to be statistical independent. Moreover, it is assumed that the number of source signals m, although it is in general also unknown, is not larger than the number of sensors n (i.e. $m \leq n$).

Our objective is to develop and to investigate adaptive learning algorithms which makes possible on–line generation of output signals $\mathbf{y}(t) = [y_1(t), y_2(t), ..., y_n(t)]^T$ which are proportional to primary source signals $\mathbf{s}(t) = [s_1(t), s_2(t), ..., s_m(t)]^T$. In general case for $m < n$ some output signals $y_i(t)$ could be zero or some output signals may have the same waveform shape at two or more output channels. For this problem, there is no way of knowing the original labeling of the sources, hence any permutation of the outputs is also a satisfying solution. Secondly, the output signals can be arbitrary scaled by non–zero scaling factors. This indeterminacy can be expressed mathematically by a matrix equation [11]

$$\mathbf{y}(t) = \mathbf{P}\mathbf{s}(t), \tag{5}$$

where $\mathbf{P} \in R^{n \times m}$ is a generalized permutation matrix in which each column and m rows ($m \leq n$) contain only one nonzero element. Remaining rows should have all elements equal to zero.

3. FEED–FORWARD MODEL AND ITS LEARNING RULES

3.1. Feed–forward NN model

Let us consider a linear feed–forward single layer neural network described by a set of equations [2,4,5,6]:

$$y_i(t) = \sum_{j=1}^{n} w_{ij} x_j(t), \quad (i = 1, 2, ..., n) \tag{6}$$

or in equivalent matrix form

$$\mathbf{y}(t) = \mathbf{W}(t)\mathbf{x}(t), \tag{7}$$

where $\mathbf{y}(t)$ is the vector of output signals, $\mathbf{W} = [w_{ij}] \in R^{n \times n}$ is the synaptic weight matrix, $\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$ is the vector of sensed input signals, $\mathbf{s}(t)$ is the vector of unknown, mutually independent source signals.

3.2. Associated learning algorithm

For the above described mixing and neural network models we have developed a novel learning rule described by a system of nonlinear differential equations (cf. [2], [6]):

$$\frac{dw_{ij}(t)}{dt} = \mu_{ij}(t) \left\{ w_{ij}(t) - f_i[y_i(t)] \sum_{p=1}^{n} w_{pj}(t) y_p(t) \right\}, \tag{8}$$

where $\mu_{ij}(t) > 0$ is a local adaptive learning rate, and $f_i(y_i)$ is a local activation function. Typically: $f(y) = y^{2k+1}$ or $f(y) = y^{2k} sign(y)$, $(k = 1 - 5)$. Usually $f(y) = y^3$.

The learning algorithm (8) can be written also in a vector form as:

$$\frac{d\mathbf{w}_j(t)}{dt} = \boldsymbol{\mu}_j(t) \left\{ \mathbf{w}_j(t) - \mathbf{f}[\mathbf{y}(t)] \mathbf{y}^T(t) \mathbf{w}_j(t) \right\}, \tag{9}$$

where $\boldsymbol{\mu}_j(t) = diag\{\mu_{1j}(t), \mu_{2j}(t), ..., \mu_{nj}(t)\}$, with $\mu_{ij} \geq 0$, $\mathbf{w}_j(t) = [w_{1j}(t), w_{2j}(t), ..., w_{nj}(t)]^T$ $(j = 1, 2, ..., n)$, and $\mathbf{f}[\mathbf{y}(t)] = [f_1[y_1(t)], f_2[y_2(t)], ..., f_n[y_n(t)]]^T$.

3.3. Learning algorithm for learning rates

In this paper we propose that each synaptic weight w_{ij} has its own (local) learning rate $\mu_{ij}(t)$ and that this rate is adjusted during the learning process according to a set of differential equations (cf. [1]):

$$\tau_1 \frac{dv_{ij}(t)}{dt} = -v_{ij}(t) + g_{ij}(t) \tag{10}$$

$$\tau_2 \frac{d\mu_{ij}(t)}{dt} = -\mu_{ij}(t) + \alpha |v_{ij}(t)|, \tag{11}$$

where $\tau_1 > 0, \tau_2 > 0$ are time constants, $\alpha > 0$ is the gain factor, $|x|$ means absolute value of x, and

$$g_{ij}(t) = w_{ij}(t) - f_i[y_i(t)] \sum_{p=1}^{n} w_{pj}(t) y_p(t). \tag{12}$$

Let us notice that this set of nonlinear differential equations can be easily implemented by electronic components using two simple (first order) low–pass filter (LPF) and a full–wave rectifier (absolute value nonlinear function). Of course, instead of first order filters we could use higher order filters in order to improve dynamic properties.

It is also interesting to observe that neural network together with its associated adaptive learning algorithm performs a separation task by self–adaptive system described by a set of nonlinear differential equations (9)–(12).

3.4. Simplified learning rule

The above algorithm can be modified assuming that the learning rate $\mu_i(t)$ is a positive scalar instead of a diagonal positive definite matrix. In this case the learning rate can be adjusted according to the following rule:

$$\tau_1 \frac{dv_{ij}(t)}{dt} = -v_{ij}(t) + g_{ij}(t), \quad \alpha > 0 \tag{13}$$

$$\tau_2 \frac{d\mu_i(t)}{dt} = -\mu_i^2(t) + \alpha \mu_i(t) \sum_{k=1}^{n} \psi[v_{ik}(t)]. \tag{14}$$

It should also be noted that instead of absolute value functions $|v_{ij}(t)|$ we can use another nonlinear symmetric functions, e.g. $\psi(v_{ij}) = \tanh(\beta |v_{ij}(t)|)$ with $\beta > 0$ or $\psi(v_{ij}) = v_{ij}^2(t)$ can be applied.

4. RECURRENT MODEL AND ASSOCIATED ON–LINE LEARNING ALGORITHM

4.1. Recurrent (feedback) neural network model

In previous section we have considered a linear feed–forward single layer neural network. In this section we consider a recurrent Amari–Hopfield type neural network described by a set of differential equations [3]:

$$\tau \frac{dy_i(t)}{dt} + y_i(t) = x_i(t) - \sum_{j=1}^{n} \widehat{w_{ij}}(t) y_j(t), \quad (i = 1, 2, ..., n) \tag{15}$$

where $\tau > 0$ is a time constant.

The above set of differential equations can be written in compact matrix form as

$$\tau \frac{d\mathbf{y}(t)}{dt} + \mathbf{y}(t) = \mathbf{x}(t) - \widehat{\mathbf{W}}(t)\mathbf{y}(t), \tag{16}$$

where $\widehat{\mathbf{W}} = [\widehat{w_{ij}}(t)] \in R^{n \times n}$.

Let us notice that the proposed neural network is fully connected, i.e. it contains self–loop connections with $\widehat{w_{ii}}(t)$ generally not equal to zero [3].

It can also be noted that if the time constant τ is negligible small, we can get a so called adiabatic approximation [3,10]

$$\mathbf{y}(t) = [\mathbf{I} + \widehat{\mathbf{W}}(t)]^{-1}\mathbf{x}(t) = \mathbf{W}(t)\mathbf{x}(t), \tag{17}$$

where $\mathbf{W}(t) = [\mathbf{I} + \widehat{\mathbf{W}}(t)]^{-1}$ is a synaptic matrix of an equivalent feed–forward neural network, under assumption that $\tau \simeq 0$ and all eigenvalues of the matrix have positive real parts. In other words for $\tau = 0$ both models are equivalent under assumption that the inverse matrix $[\mathbf{I} + \widehat{\mathbf{W}}(t)]^{-1}$ exists for any t, i.e. $\mathbf{W}(t) \neq -\mathbf{I}, \quad \forall t$.

4.2. Associated self-adaptive learning algorithm

For the fully recurrent model we have developed the following adaptive on–line learning algorithm, written in matrix form (cf. [3]):

$$\frac{d\widehat{\mathbf{W}}(t)}{dt} = \boldsymbol{\mu}(t) \; .* \left\{ (\widehat{\mathbf{W}}(t) + \mathbf{I}) \left[\mathbf{f}[\mathbf{y}(t)]\mathbf{y}^T(t) - \mathbf{I} \right] \right\}, \tag{18}$$

where $\boldsymbol{\mu}(t)$ is the $n \times n$ matrix with positive entries and $.*$ means element–wise multiplication.

Incorporating an auxiliary inter–neuron layer, defined as:

$$\mathbf{z}(t) = \widehat{\mathbf{W}}(t)\mathbf{f}[\mathbf{y}(t)] + \mathbf{f}[\mathbf{y}(t)], \tag{19}$$

this algorithm takes a particular simple form

$$\frac{d\widehat{\mathbf{W}}(t)}{dt} = \boldsymbol{\mu}(t). * \left[\mathbf{z}(t)\mathbf{y}^T(t) - \widehat{\mathbf{W}}(t) - \mathbf{I} \right]. \tag{20}$$

The above adaptive learning rule can be formulated in scalar form as:

$$\frac{d\widehat{w_{ij}}(t)}{dt} = \mu_{ij}(t) \left[z_i(t)y_j(t) - \widehat{w_{ij}}(t) \right], \text{for } i \neq j \tag{21}$$

$$\frac{d\widehat{w_{ii}}(t)}{dt} = \mu_{ii}(t) \left[z_i(t)y_i(t) - \widehat{w_{ii}}(t) - 1 \right], \tag{22}$$

where the output of the i-th neuron is defined as

$$z_i(t) = \sum_{j=1}^{n} \widehat{w_{ij}}(t) f_j[y_j(t)] + f_i[y_i(t)]. \tag{23}$$

4.3. Update of learning rates

The local learning rates are updated according to following rules:

$$\tau_1 \frac{dv_{ij}(t)}{dt} = -v_{ij}(t) + g_{ij}(t), \tag{24}$$

$$\tau_2 \frac{d\mu_{ij}(t)}{dt} = -\mu_{ij}(t) + \alpha\psi[v_{ij}(t)], \tag{25}$$

where $\tau_1 > 0, \tau_2 > 0$ are time constants of first order low pass filters, $\alpha > 0$ is the gain factor,

$$g_{ij}(t) = \begin{cases} z_i(t)y_j(t) - \widehat{w_{ij}}(t), & \text{for } i \neq j \\ z_i(t)y_i(t) - \widehat{w_{ii}}(t) - 1, & \text{for } i = j \end{cases} \tag{26}$$

and $\psi(v_{ij})$ is a symmetric non–negative function (e.g. $\psi(v_{ij}) = |v_{ij}|$, $\psi(v_{ij}) = v_{ij}^2$ or $\psi(v_{ij}) = \tanh(\beta|v_{ij}|)$).

The above differential equations can be written compactly in matrix form as:

$$\tau_1 \frac{d\mathbf{V}(t)}{dt} = -\mathbf{V}(t) + \mathbf{z}(t)\mathbf{y}^T(t) - \widehat{\mathbf{W}}(t) - \mathbf{I}, \tag{27}$$

$$\tau_2 \frac{d\boldsymbol{\mu}(t)}{dt} = -\boldsymbol{\mu}(t) + \alpha\boldsymbol{\psi}[\mathbf{V}(t)], \tag{28}$$

$$\mathbf{z}(t) = (\widehat{\mathbf{W}}(t) + \mathbf{I})\mathbf{f}[\mathbf{y}(t)]. \tag{29}$$

Instead of using the local learning rates $\mu_{ij}(t)$ individually for each synaptic weight $\widehat{w_{ij}}(t)$ we can use one global scalar learning rate $\mu(t) > 0$. This global learning rate can be updated as follows:

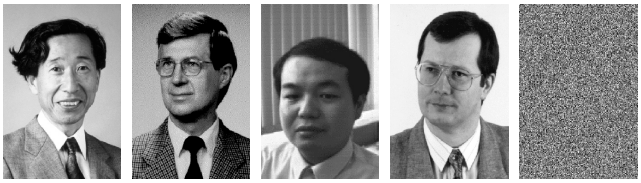
$$\tau_2 \frac{d\mu(t)}{dt} = -\mu^2(t) + \alpha\mu(t)||\mathbf{V}(t)||_F^2, \tag{30}$$

where $||\quad||_F$ means the Frobenius norm.

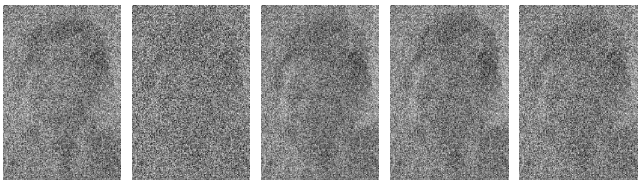
5. EXPERIMENTS

In Fig. 1 and Fig. 2 two simulation examples are given to demonstrate validity, performance and dynamics of learning of proposed algorithms.

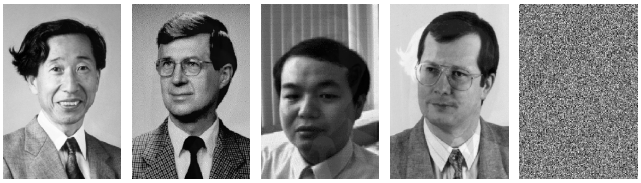
In first example four natural images (faces of the authors) were mixed together with large Gaussian noise (approximately 20 times stronger than the natural images) by using randomly chosen nonsingular mixing matrix $\mathbf{A}$. It was assumed that only mixed (sensor) images are available (see second row of Fig. 1).



Five original images (but unknown to the neural net)

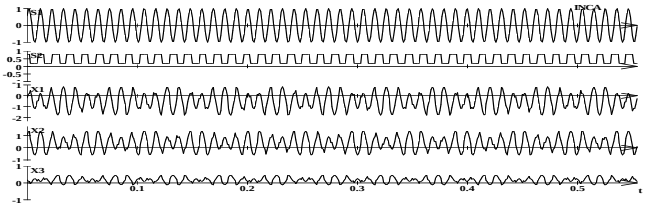


Five mixed images (used for separation)

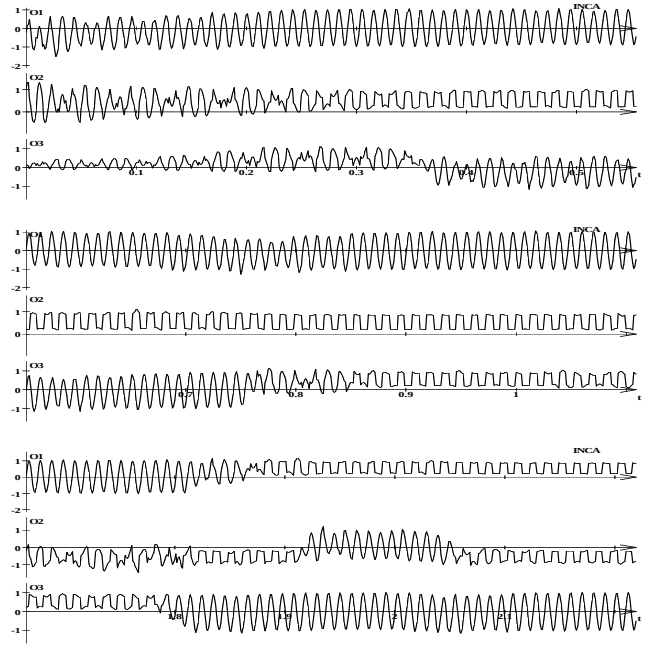


Final (stable states) of five separated images

Figure 1. Example of blind separation of image sources



Original (unknown) source signals: $S1, S2$ and their three mixtures: $X1, X2, X3$



Process of signal separation. $O1, O2$ and $O3$ represent the output signals $y_1(t), y_2(t), y_3(t)$. t is given in seconds.

Figure 2. One–dimensional source separation with switching effect in redundant sensor case.

Note that original faces are not visible from their mixture. The self–adaptive learning algorithm (8)–(12) (with parameters $f(y) = y^3, \tau_1 = \tau_2 = 0.1, \alpha = 1000$) was able to separate (extract) original images in time less than $600ms$ (for analog electronic circuit) or after 3 learning epochs (in discrete–time case). No switching (cross–over) behavior has been observed in the cases when $m = n$.

In the second example two unknown sources ($s_1(t) = \sin(600t)$ and $s_2(t) = 0.5 + 0.3sign(\cos(350t))$) were mixed by using randomly chosen matrix $\mathbf{A} \in R^{3\times 2}$ of the full rank $m = 2$. It was assumed that number of primary sources is also unknown and number of sensors is $n = 3$. Exemplary results of applying algorithm (18)–(20) (with parameters $f(y) = y^7, \alpha = 500, \tau_1 = \tau_2 = 0.2, \mu_{ij} \leq 10$) are shown in Fig. 2. From these plots it is seen that the algorithm is able successfully to separate original sources (sine-wave and rectangular signals). However, some random cross–over or switching between output signals may be observed.

6. CONCLUSIONS

In this paper we have proposed efficient on–line learning algorithms with self–adaptive local learning rates for blind separation of sources. The algorithms are developed both for feed–forward and feed–back (recurrent) neural network architectures and are described by a system of nonlinear differential equations. They can be easily converted to discrete–time (recursive) simple formulas, for example using the Euler formula. The proposed learning algorithms are especially suitable for non–stationary environment, i.e. when mixing parameters and/or stochastic distribution of signals are changing in time in unpredictable way. We have found by computer simulations that the proposed algorithms provide a switching of attention (cross–over of output signals) in the case when number of sensors is greater than number of sources. The open problem is to find a technique which enables control of such switching and ensures a stable selective attention.

REFERENCES

[1] Amari, S., Theory of adaptive pattern classifiers, *IEEE Trans. Electr. Comput.*, EC-16, 1967, pp. 299-307.

[2] Amari, S., Cichocki, A. and Yang, H.H., A new learning algorithm for blind signal separation, *Advances in Neural Information Processing Systems 8*, The MIT Press, Cambridge, 1996, (in print).

[3] Amari, S., Cichocki, A. and Yang, H.H., Recurrent neural networks for blind separation of sources, *Proceedings of NOLTA–95*, Las Vegas, USA, Dec.1995, vol.1, pp.37-42.

[4] Bell, A. J. and Sejnowski, T.J., An information–maximization approach to blind separation and blind deconvolution, *Neural Computation*, vol. 7, 1995, pp. 1129–1159.

[5] Cardoso, J.-F., Belouchrani, A. and Laheld,B., A new composite criterion for adaptive and iterative blind source separation, *Proceedings ICASSP–94*, vol.4, Adelaide, Australia, May 1994, pp.273–276.

[6] Cichocki, A., Unbehauen, R., Moszczynski, L. and Rummert, E., A new on–line adaptive learning algorithm for blind separation of source signals, *Proceedings of ISANN–94*, Taiwan, Dec. 1994, pp. 406–411.

[7] Hendin, O., Horn, D. and Hopfield, J.J., Decomposition of a mixture of signals in a model of the olfactory bulb, *Proc. Natl. Acad. Sci. USA*, 1994, pp. 5942–5946

[8] Jutten, C. and Herault, J., Blind separation of sources, Part I: An adaptive algorithm based on neuro-mimetic architecture, *Signal Processing*, vol. 24, 1991, 1–20.

[9] Oja, E. and Karhunen, J., Signal separation by nonlinear Hebbian learning, *Proceedings of ICNN–95*, Perth, Australia, Dec.1995, pp. 211–217.

[10] Matsuoka, K., Ohya, M. and Kawamoto, M., A neural net for blind separation of non-stationary signal sources, *Neural Networks*, vol.8, 1995, pp. 411–419.

[11] Tong, L., Liu, R., Soon, V.C. and Huang, Y.-F., Indeterminacy and identifiability of blind identification, *IEEE Transactions on Circuits and Systems*, vol.38, 1991, pp. 499–509.