

Hebbian Learning and Self-Association in Nonlinear Neural Networks

Francesco Palmieri, *Member IEEE*

Abstract— A self-organizing feature map, based on a Hebbian paradigm, is proposed as a universal adaptive memory. The learning paradigm, can be applied to arbitrary network topologies containing the standard sigmoidal nonlinearities at their nodes. The system generalizes the linear principal components by mapping the input space into a set of orthogonal nonlinear projections. Only localized learning rules are necessary for the adaptation. The size of the system is related to the desired accuracy and to the density of the examples.

I. INTRODUCTION

A self-organizing neural network has to provide a useful alternate, or compressed, representation of the input space (feature map). Previous contributions have shown how in artificial neural networks, simple learning rules, localized at the synaptic level, can lead to interesting patterns of organization (Linsker, 1986). These structures bear strong resemblance to patterns of activity observed in biological brains.

The emphasis of this work is on the search for suitable *principles of self-organization*, that applied to loosely constrained architectures, can lead to the emergence of representations useful for the applications. For example, in the linear framework, the *principal components* provide compression for sources exhibiting marked linear correlations. Our goal is to provide an analysis framework, solidly rooted in the signal processing and system theory, that generalizes the well-studied adaptive linear filter theory (Haykin, 1986).

We came to focus our attention in our previous papers (Palmieri, Zhu and Chang, 1993) (Palmieri and Zhu, 1993), on two main criteria for the Hebbian (excitatory) and the anti-Hebbian (inhibitory) synapses respectively. In particular, an Hebbian

synapse attempts to *excite* the node to which it is attached in a way that guarantees the preservation of the source information. This is achieved by requiring the input to be maximally reconstructible from the node value (Principle of Maximum Inverse Linear Reconstruction). An anti-Hebbian synapse instead, *inhibits* the node to which it is attached, by removing the correlation between its input and the node value.

These criteria are analyzed in detail for the linear and the nonlinear neural networks, made up of nodes containing the familiar sigmoidal nonlinearities (Palmieri, 1993).

The principle of self-association in linear networks has been proposed by Bourland and Kamp (1988), Baldi and Kornik (1989), Hrycej (1990), Xu (1993), Karhunen and Joutsensalo (1993), that did not consider the hybrid architecture containing both Hebbian and anti-Hebbian synapses. Also the constraints on the locality of the learning rules was not always imposed. Hybrid structures for the computation of the principal components have been proposed by Foldiak (1988) and by Kung and Diamantaras (1990), that first postulated learning equations, and then showed how the principal components emerge from the self-organization.

We start from the objective functions, and study the learning behavior of a linear and a nonlinear neural network that contains an arbitrary mixture of Hebbian and anti-Hebbian synapses. The nonlinear projections that emerge from the self-organization provide a potentially universal multidimensional representation of the input space. The remarkable feature of this paradigm is in the full locality of the learning algorithms. Even though the cost function (backward reconstruction) is formulated in global terms, the orthogonality that emerges from the anti-Hebbian connections only need Hebbian learning equations completely localized to the synaptic level.

II. THE LINEAR NEURAL NODE

A. The Anti-Hebbian Synapse

Figure 1(a) shows a linear neural node with an anti-Hebbian input x_a and another generic input z . The

The author is with the Dipartimento di Elettronica, Università degli Studi di Napoli "Federico II," via Claudio 21, 80125 NAPOLI (Italy) (E-mail: palmieri@nadis.dia.unina.it), and with the Department of Electrical and Systems Eng., The University of Connecticut, Storrs, CT 06269-3157 (USA).

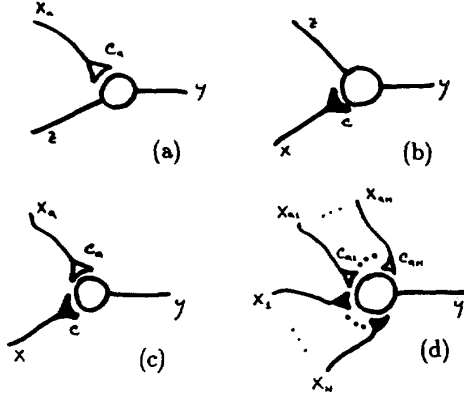


Figure 1: The neural node

output value is $y = c_a x_a + z$ where both x_a and z are two arbitrary random variables. As already discussed in (Palmieri, Zhu and Chang, 1993), the anti-Hebbian synapse “inhibits” the neuron by removing the correlation between x_a and y . The unique solution to $E[x_a y] = 0$, is clearly $c_a = -\frac{E[x_a z]}{E[x_a^2]}$. Note that the linear node with the anti-Hebbian synapse can be seen as a linear adaptive filter of order one, with the output y playing the role of the error in estimating $-z$ linearly from x_a . In fact, the same solution can be found by minimizing $E[y^2]$. The classical adaptive linear filter theory (Haykin, 1986) is based on this formulation with the *orthogonality principle* ensuring that the input x_a and the error y are orthogonal in optimal conditions. We prefer using the orthogonality condition as the starting point because, as we will see in the following, it becomes immediately applicable to the nonlinear scenario.

B. The Hebbian Synapse

Figure 1(b) shows the same linear neural node of Figure 1(a) with one Hebbian input x . The Hebbian synapse “excites” the neuron by forcing the neural node value to be “as related as possible” to x . We identified in our previous papers (Palmieri and Zhu, 1993; Palmieri, 1993) the following strategy for the Hebbian synapse: learn the value c that corresponds to the best linear backward reconstruction of x . That is, find the value of c that minimizes

$$\mathcal{E} = E[(x - cy)^2]. \quad (1)$$

The cost function can be re-written as

$$\mathcal{E} = E[x^2]c^4 + 2E[xz]c^3 + (E[z^2] - 2E[x^2])c^2 - 2E[xz]c + E[x^2]. \quad (2)$$

By computing its derivatives, it can be easily seen that for arbitrary z , \mathcal{E} it is not a purely convex func-

tion, but it may exhibit one global and one local minimum (Palmieri and Zhu, 1993). Typical cases arise when both inputs are orthogonal, $E[xz] = 0$, in which case there are two symmetric equivalent global minima, and when $z = 0$ which gives the two equivalent solutions $c = \pm 1$.

C. The Hybrid Node

A hybrid node is shown in Figure 1(c) where both excitatory and inhibitory synapses are present. The node value is $y = c_a x_a + cx$ with the anti-Hebbian synapse trying to inhibit the node, and the Hebbian synapse trying to excite it. Since the anti-Hebbian synapse has a stationary point given by $E[x_a y] = 0$ with $c_a = -c \frac{E[x_a x]}{E[x_a^2]}$, by substituting this expression in \mathcal{E} , we have

$$\mathcal{E} = (E[x^2] - \frac{E[x x_a]^2}{E[x_a^2]})(c^4 - c^2) + E[x^2]. \quad (3)$$

Since $E[x x_a]^2 \leq E[x_a^2]E[x^2]$, \mathcal{E} has a maximum at $c = 0$ and two equivalent global minima at $c = \pm 1$. The two stationary points are $(c, c_a) = \pm(1, -\frac{E[x_a x]}{E[x_a^2]})$.

A typical special case arise when x and x_a are orthogonal, $E[x_a x] = 0$. The anti-Hebbian synapse does not arise and the Hebbian synapse just copies the input to the output. Also interesting is the case with $x_a = kx$, with k being any real constant. The anti-Hebbian synapse becomes $c_a = -kc$ and $y = 0$. The anti-Hebbian synapse shuts off the neuron completely, with the Hebbian synapse left undefined (floating). The inverse reconstruction of x is ill-defined with $\mathcal{E} = E[x^2]$, $\forall c$.

The above analysis can be easily generalized to an hybrid node accepting M anti-Hebbian inputs $(x_{a1}, x_{a2}, \dots, x_{aM})^T = \mathbf{x}_a$ through M synapses $(c_{a1}, c_{a2}, \dots, c_{aM})^T = \mathbf{c}_a$, and N Hebbian inputs $(x_1, x_2, \dots, x_N)^T = \mathbf{x}$ through N synapses $(c_1, c_2, \dots, c_N)^T = \mathbf{c}$ as shown in Figure 1(d). The anti-Hebbian synapses guarantee orthogonality with $E[x_{ai} y] = 0$, $\forall i = 1, \dots, M$, with values given by

$$c_a = -E[\mathbf{x}_a \mathbf{x}_a^T]^{-1} E[\mathbf{x}_a \mathbf{x}^T] \mathbf{c}. \quad (4)$$

The principle of linear backward reconstruction of \mathbf{x} requires the simultaneous minimization of the cost function

$$\mathcal{E} = E[\|\mathbf{x} - \mathbf{c}\mathbf{y}\|^2], \quad (5)$$

which becomes using (4)

$$\mathcal{E} = E[\mathbf{x}^T \mathbf{x}] - 2\mathbf{c}^T \mathbf{R}' + \mathbf{c}^T \mathbf{c} \mathbf{c}^T \mathbf{R}', \quad (6)$$

with

$$\mathbf{R}' = \mathbf{R} - \mathbf{p}_a \mathbf{R}_a^{-1} \mathbf{p}_a^T, \quad (7)$$

and $\mathbf{R} = E[\mathbf{x} \mathbf{x}^T]$ and $\mathbf{p}_a = E[\mathbf{x} \mathbf{x}_a^T]$. We have already proven (Palmieri and Zhu, 1993), and it is

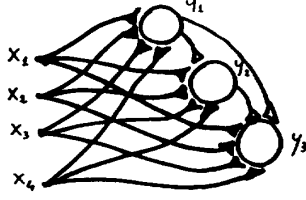


Figure 2: The architecture for orthogonalization

easy to verify it by direct substitution, that equation (6) has critical points at the eigenvectors of R' . The only two minima are at plus-or-minus the principal eigenvector. All the other ones are saddle points.

It is useful to examine a special cases that will be important in the study of the architecture of the next section. If the anti-Hebbian inputs are proportional to the projections of \mathbf{x} on a set of M eigenvectors of R ($M < N$), as $x_{a1} = k_1 \mathbf{q}_1^T \mathbf{x}$; $x_{a2} = k_2 \mathbf{q}_2^T \mathbf{x}$; ...; $x_{aM} = k_M \mathbf{q}_M^T \mathbf{x}$, with k_1, k_2, \dots, k_M , arbitrary constants, we have that from (7) \mathbf{c} will converge to plus-or-minus the principal eigenvector of the set $\{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_N\} - \{\mathbf{q}_{i_1}, \mathbf{q}_{i_2}, \dots, \mathbf{q}_{i_N}\}$. A more comprehensive analysis of the hybrid neural node would include a nonlearning input \mathbf{z} as in Figures 1(a) and 1(b). This case may be of interest if we want to assume that each node in the network has an independent noise source, or accepts nonlearning inputs, or inputs that learn at a much lower speed to be considered stationary. We leave the more general analysis out for brevity. It will be discussed somewhere else.

D. A Linear Architectures

The analysis of the general hybrid node allows us to establish a number of results for arbitrary architectures made up of linear neurons and mixtures of Hebbian and anti-Hebbian synapses. For example, the architecture of Figure 2 certainly gives the principal components of the input. The first output is the projection on the principal component. The second one, using (6) and (7) is the projection on the second one and so on. Note that the criterion of maximum inverse reconstruction is applied in a decoupled fashion. In other words, each neuron maximally reconstruct its inputs from its output independently from the others. The solution is the same as that obtained from the more general cost function $\mathcal{E} = E[\|\mathbf{x} - \mathbf{x}_b\|^2] = E[\|\mathbf{x} - C C^T \mathbf{x}\|^2]$, which would require coupling of the various outputs for \mathbf{x}_b (Palmieri and Zhu, 1993(b)). The equivalence is obvious if we observe that at equilibrium

the outputs are orthogonal and the minimization of \mathcal{E} is the same as the independent minimization of $\mathcal{E}_i = E[\|\mathbf{x} - \mathbf{c}_i \mathbf{y}_i\|^2]$, $\forall i = 1, \dots, M$

III. THE NONLINEAR NODE

The great interest we attribute to the paradigm studied above for the linear scenario, is in the fact that it can be immediately applied to neural nodes that have the common nonlinear structure with the sigmoidal nonlinearity. The sigmoid can be any strictly increasing function such that $\phi_\alpha(0) = 0$, $\lim_{h \rightarrow \infty} \phi_\alpha(h) = k_u$, and $\lim_{h \rightarrow -\infty} \phi_\alpha(h) = k_l$ with k_u and k_l being any two real constants. Even though most of results hold for these functions, in our calculations and simulations we adopt the following function:

$$\phi_\alpha(h) = \frac{1}{2} \tanh \frac{\alpha h}{2}, \quad (8)$$

where the parameter α is introduced to control the slope of the transition around zero. In fact for $\alpha \rightarrow \infty$, $\phi_\alpha(h) \rightarrow \frac{1}{2} \text{sgn}(h)$, and for small α , $\phi_\alpha(h) \simeq \frac{\alpha}{4} h$. According to the magnitude of the values at its input the sigmoid can behave either as a binary threshold, or as a linear unit, conveniently generalizing the linear case. The derivative is easily computed to be $\phi'_\alpha(h) = \alpha(\frac{1}{4} - \phi_\alpha^2(h))$.

A. The Anti-Hebbian Synapse

Let us begin by considering the simple architecture of Figure 1(a) with only one anti-Hebbian input and an extra input \mathbf{z} . The output is $\mathbf{y} = \phi(c_a \mathbf{x} + \mathbf{z})$. Just as in the linear case, the objective of the anti-Hebbian synapse is that of removing the correlation $E[\mathbf{x}_a \mathbf{y}]$. The following theorem guarantees that this is possible under the most general conditions.

Theorem: For any random variables x_a and \mathbf{z} , and any sigmoidal function $\phi(h)$ with $\phi(0) = 0$, there exist a unique value of c_a that satisfies the equation:

$$E[\mathbf{x}_a \phi_\alpha(c_a \mathbf{x}_a + \mathbf{z})] = 0. \quad (9)$$

Proof: Since

$$\frac{\partial E[\mathbf{x}_a \mathbf{y}]}{\partial c_a} = E[\mathbf{x}_a^2 \phi'_\alpha(c_a \mathbf{x}_a + \mathbf{z})] > 0, \quad (10)$$

$\forall c_a$, the function $E[\mathbf{x}_a \mathbf{y}]$ is strictly increasing in c_a . Therefore if a solution to the equation exists, it is unique. But c_a can take arbitrarily large values to bring $\phi_\alpha(c_a \mathbf{x}_a + \mathbf{z})$ to saturation. This means that there are values of c_a at which the quantity $E[\mathbf{x}_a \phi_\alpha(c_a \mathbf{x}_a + \mathbf{z})]$ is positive and values at which it is negative, which implies that there exist a value of c_a where $E[\mathbf{x}_a \phi_\alpha(c_a \mathbf{x}_a + \mathbf{z})] = 0$.

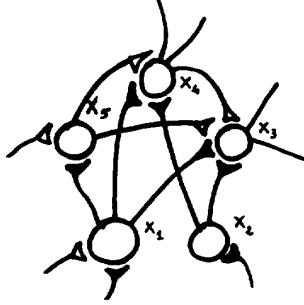


Figure 3: A generic hybrid architecture

B. The Hebbian Synapse

Refer to Figure 1(b). The cost function for the Hebbian synapse on the nonlinear node becomes

$$\mathcal{E} = E[(x - c\phi(cx + z))^2]. \quad (11)$$

The results on the linear case suggest that for arbitrary distribution of x and z the function \mathcal{E} is not necessarily a convex function of c . However, we verified for special cases, that the unicity of the solution (two equivalent solutions) remains in a mixed node, just like in its linear counterpart.

C. The Hybrid Node

For the mixed architecture of Figure 1(c), even though we cannot write down an explicit expression for the value of c_a , we know from the above theorem that for any value of c a solution to the equation

$$E[x_a \phi_a(c_a x_a + cx)] = 0, \quad (12)$$

exists and it is unique. The cost function for the Hebbian synapse becomes $\mathcal{E} = E[(x - c\phi(c_a x_a + cx))^2]$, where to c_a we should substitute the solution to (12). We have not yet been able to study rigorously the convexity of \mathcal{E} as in the linear case, but we conjecture from simulations that a result similar to that of the linear case, should hold with one maximum and two symmetric minima.

The same criteria can be applied to the more general node of Figure 1(d). The output is $y = \phi(c_a^T x_a + c^T x)$. For any c , the anti-Hebbian synapse will have a unique weight configuration given by the solution to the set of equations

$$E[x_a \phi(c_a^T x_a + c^T x)] = 0, \quad (13)$$

and the cost function for the Hebbian part will be

$$\mathcal{E} = E[||x - c\phi(c_a^T x_a + c^T x)||^2]. \quad (14)$$

D. The Nonlinear Neural Network

An arbitrarily complex architecture with both Hebbian and anti-Hebbian synapses can be studied with

the tools provided above. To fix the ideas, let us look at the network of Figure 3. The node values $\{x_i\}$ (after the sigmoids) are the result of arbitrary mixtures of excitatory and inhibitory inputs. From the above discussion, we know that if two nodes are connected through an anti-Hebbian synapse, at equilibrium their node values will be orthogonal. Simultaneously, for the Hebbian synapses the criterion of inverse reconstruction requires minimization of

$$\mathcal{E}_i = E[(x_i - x_{bi})^2] = E[(x_i - \sum_{j \in B_i} c_{ji} x_j)^2], \quad (15)$$

applied to all the nodes in the network, where B_i is the set of nodes excited by the i th node through Hebbian synapses.

This criterion does not appear to be properly local because it requires the computation of x_{bi} which is the *backpropagation* of the values $x_j \in B_i$ through the Hebbian connections. However, if every pair of $x_j \in B_i$ is orthogonal, the minimization of \mathcal{E}_i is equivalent to the separate simultaneous minimization of

$$\mathcal{E}_{ij} = E[(x_i - c_{ji} x_j)^2], \quad \forall j \in B_i. \quad (16)$$

Hence if each pair of nodes in B_i is connected through an anti-Hebbian synapse, the cost function is *completely local at the synaptic level!*

A neural architecture composed by arbitrarily interconnected nodes leads to the emergence of a representation of the inputs that guarantees the existence of a linear transformation for "retrieval" from the network. This is the meaning we attribute to this system as an *emerging physical memory*.

Note that in a multi-layered feedforward topology, a perfect backward reconstruction at each node, implies a perfect reconstruction at the inputs.

The network architecture that follows, already introduced in (Palmieri, 1993), demonstrates how to obtain a set of nonlinear projections of the input space, that go from purely linear principal components, to binary networks.

E. An Architecture for Nonlinear Principal Components

The architecture shown in Figure 2 implements a generalization of the linear principal components. If the sigmoids at each node are used in their linear region (small net variance at the input of each node), the network essentially behaves as a linear filter with the linear principal components emerging as the result of the self-organization. If instead the input energy is increased, or the sigmoidal functions are made increasingly sharp, the network begins to provide a progressively finer breakdown of the input space into orthogonal nonlinear projection.

The following simulation performed with the architecture of Figure 2, on a one-dimensional input should be sufficient to demonstrate the potential of this idea.

Figures 4 and 5 show the resulting mappings provided by the network of Figure 2 with $N = 1$, after self-organization with input examples uniformly distributed in the interval $[-0.5, 0.5]$. The learning rule for the anti-Hebbian synapses is the simple one:

$$\Delta c_{aij} = -\mu y_i y_j, \quad j = 1, \dots, M, i = j + 1, \dots, M, \quad (17)$$

where c_{aij} is the anti-Hebbian connection from the output j to the output i . A number of variations to the rule would work as well with slightly different convergence properties. The Hebbian synapses are trained using to the rule

$$\Delta c_{ij} = \mu (x_j - x_{bj}) y_i, \quad j = 1, \dots, N; i = 1, \dots, M, \quad (18)$$

where $x_{bj} = \sum_{i=1}^M c_{ij} y_i$ is the output y backpropagated to the j th input. As we pointed out above, even though all the simulations that we report are performed using this rule, a decoupled version of it (Oja's rule): $\Delta c_{ij} = \mu (x_j y_i - c_{ij} y_i^2)$, $j = 1, \dots, N; i = 1, \dots, M$, which is completely local at the synaptic level, would also work for the orthogonality of the outputs. Oja's rule would just exhibit worse missadjustment and generally slower convergence. A detailed analysis of the convergence behavior of these algorithms is beyond the purpose of this paper and it will be reported elsewhere.

Figure 4 shows the resulting orthogonal projections $\{y_1(x), \dots, y_4(x)\}$ for $M = 4$, $\alpha = 100$ and 1000 examples uniformly distributed in $[0.5, 0.5]$, presented 100 times with $\mu = 0.003$. Note how the sigmoid has still a linear region that is used by the various stages in the backward reconstruction of x . The last graph in the figure shows the almost perfect reconstruction on the whole range. Note that only 3 orthogonal projections, with an oscillatory behavior, emerge and are sufficient for an almost perfect reconstruction of x . Figure 5 shows the results of the same simulation with $M = 6$, $\alpha = 10^7$ and 1000 examples presented 300 times with $\mu = 0.00003$. The sigmoid is practically a step function and the orthogonal projections that emerge show the oscillatory behavior necessary to reconstruct the input. The last graph shows the reconstruction of the input. It is important to note that the accuracy of the solution is due mainly to the *density* of the examples in the input space. The self-organizing network *zooms-in* the details of the input space up to the resolution of the examples presented.

To demonstrate this effect more clearly we have performed two simulations with only 5 examples,

whose results are shown in Figure 6. The input space is now very sparse and the network converges to a configuration that guarantees almost perfect reconstruction of those points leaving the other regions filled in by the interpolation provided by the final network configuration. Figure 6(a) is the result of 5 examples presented 20000 times to a network with $M = 4$, $\alpha = 100$ and $\mu = 0.003$. The resulting functions, are orthogonal on the example set and are obviously less constrained providing good reconstruction only on the input points. The results of same simulation are repeated in Figure 5(b) for a very sharp sigmoid having $\alpha = 10^7$.

More simulations have been performed on multi-dimensional examples confirming what we see from the one dimensional case. The orthogonal projections that emerge guarantee an almost perfect reconstruction at the sample points. More projections can be obtained at each stage by increasing the "degree of nonlinearity" of the sigmoids.

IV. CONCLUSIONS

The framework provided by this paper provides a solid approach to the design and implementation of a self-organizing feature map. The network structure, based on standard linear combiners and sigmoidal functions, derives its learning algorithms from a clear definition of the cost functions, and uses only localized learning rules of the Hebbian type.

The network spans the whole spectrum ranging from linear principal components to nonlinear orthogonal projections. The slope of the sigmoidal functions, or the net energy of the data at each node, determines the degree of nonlinearity. By increasing the size of the network, the feature map "zooms in" the details of the input space up to density of the examples.

The feature map can be used as a pre-processing stage followed by a linear filter (Palmieri, 1993), for any adaptive nonlinear mapping and provides a promising alternative to backpropagation networks.

V. REFERENCES

- Baldi P. and K. Kornik, (1989), "Neural Networks and Principal Component Analysis: Learning from Examples Without Local Minima," *Neural Networks*, Vol. 2, pp. 53-58.
- Bourland H. and Y. Kamp (1988), "Auto-Association by Multi-Layer Perceptrons and Singular Value Decomposition," *Biological Cybernetics*, Vol. 59, pp. 291-294.
- Foldiak P. (1988), "Adaptive Network for Optimal Linear Feature Extraction," *Neural Networks*, Vol. 2, pp. 459-473.
- Haykin S. (1986), *Adaptive Filter Theory*, Prentice Hall.
- Hrycej T. (1990), "Self-Organization by Delta Rule," *Proceedings of the International Joint Conference on Neural Networks*, San Diego, CA, pp. 307-312.
- J. Karhunen and J. Joutsensalo (1993), "Nonlinear Generalization of Principal Component Learning Algorithms," *Proc.*

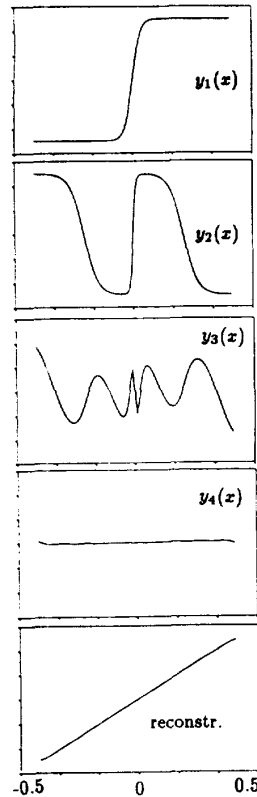


Figure 4: Resulting mappings for $M = 4$, $\alpha = 100$ and 1000 examples

of the International Joint Conference on Neural Networks, Nagoya, Japan, October.

Kung S. Y. and K. I. Diamantaras (1990), "A Neural Network Learning Algorithm for Adaptive Principal Component Extraction (APEX)," *Proceedings of International Conference on Acoustics Speech and Signal Processing*, Vol. 2, pp. 861-864.

Linsker R. (1986), "From Basic Network Principles to Neural Architectures: Emergence of Spatial-Opponent Cells," *Proc. National Academy of Science, USA, Neurobiology*, Vol. 83, pp. 7508-7512.

Palmieri F., J. Zhu and C. Chang (1993), "Anti-Hebbian Learning in Topologically Constrained Linear Networks: A Tutorial," *IEEE Trans. on Neural Networks*, Vol. 4, N. 5, Sept.

Palmieri F., J. Zhu (1993a), "The Behavior of a Single Self-Associative Neuron," *Proc. of the World Congress on Neural Networks*, Portland, OR, July.

Palmieri F., J. Zhu (1993b), "Hebbian Learning in Linear Networks: A Review," Tech. Rep. 5-93, Dept of Electrical and Systems Eng., The University of Connecticut, Storrs, CT 06269-3157.

Palmieri F. (1993), "Linear Self-Association for Universal Memory and Approximation," *Proc. of the World Congress on Neural Networks*, Portland, OR, July.

Xu L. (1993), "Least Mean Square Error Reconstruction Principle for Self-Organization," *Neural Networks*, Vol. 6.

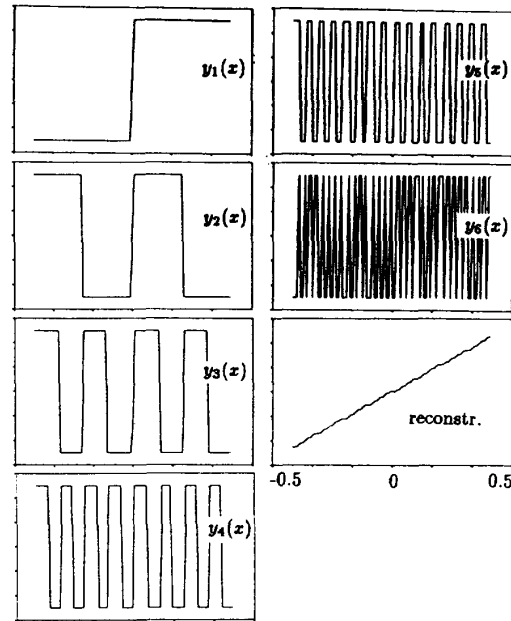


Figure 5: Resulting mappings for $M = 4$, $\alpha = 10^7$ and 1000 examples

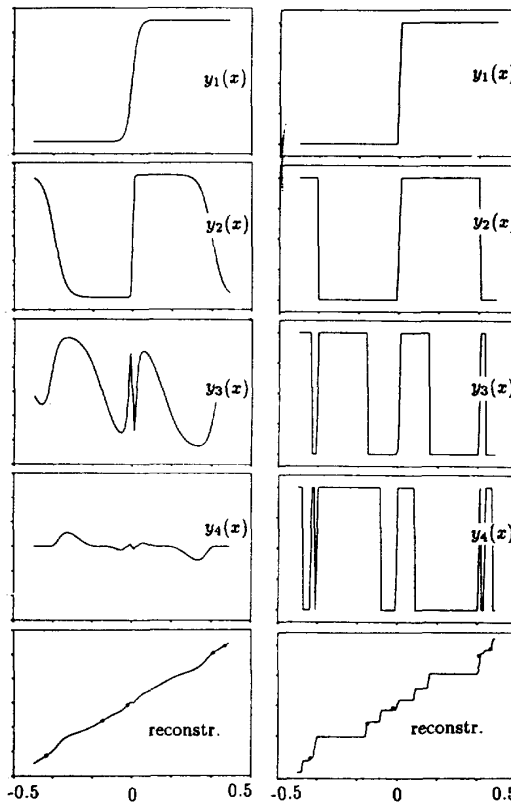


Figure 6: Resulting mappings for $M = 4$, $\alpha = 100$ (a) and $\alpha = 10^7$ (b), with 5 examples