

# Nonlinear Factorization in the Hippocampal Neural Structure.

Sirota A.M., Frolov A.A.\*, Husek D.\*\*

Moscow Institute of Physics and Technology, anton.sirota@usa.net

\*Institute of Higher Nervous Activity and Neurophysiology of the RAS

\*\*Institute of Computer Science AS of the CR, dusan@uivt.cas.cz

## Abstract

*Intrinsic factor analysis (factorization) framework for information redundancy elimination by means of Hebbian learning in sparsely encoded Hopfield-like neural network is presented. Computer simulations revealed that the information redundancy, which can be eliminated by factorization, is sparseness dependent. Due to strong similarity of Hopfield-like neural network to that of CA3 field of the hippocampus and following Marr's ideas we propose physiological mechanism for redundancy elimination (factorization) in CA3 and further replay to neocortex in the form of "classificatory units".*

## Introduction

In order to be efficient information storage mechanism should possess several main features. Firstly, encoding units should be used effectively i.e. maximum information should be encoded in each synapse. Secondly, internal informational redundancy in both spatial and temporal structure of raw data should be eliminated. One of the most important functions of human memory is the elimination of that redundancy. Due to regularity of our environment there exists a universal vocabulary of *factors* (combinations of coherently occurring features) that are invariant through individual life and even from species to species. In his influential papers David Marr [12,13] raised this problem and noted that redundant information may be stored efficiently if some collection of features that commonly occur were extracted from the input pattern and added to the vocabulary of brain's experience as a new entity (factor, or concept, in Marr's terminology). Through this vocabulary the brain later interprets and records its experience. Marr suggested that hippocampus serves as a processor performing the elimination of redundancy of the incoming complex sensory information for its efficient storage in the neocortex as a set of "classificatory units".

## Factor Analysis in neural network

The problem of information redundancy elimination is of purely statistical origin. One of the methods that solve it is Factor Analysis (factorization) that performs a decomposition of a complex vector signal into a set of simple factors basing on correlation between components of the former. One particular form of nonlinear factorization and most natural for the high level brain functions is a Boolean one, which implies that a complex vector signal (pattern) has a form of logical sum of weighted binary factors:  $X = \bigvee_{l=1,L} \mathbf{a}_l \mathbf{f}^l$ .

Each component of the factor corresponds to one of the environment features. The goal of Factor Analysis is to find the representation of the complex input signal in which the relations between a large number of correlated variables are reduced to relations between a few, usually non-correlated factors. This representation is based on the analysis of a covariance matrix of the input signal.

Hopfield-like neural network can naturally perform such function. The central idea is that Hopfield-like network can easily learn cross-correlations that underlie in incoming complex signal using Hebbian learning rule, which forms connections matrix as a covariance matrix for the set of stored patterns. Neurons that tend to fire together (represent one common factor) will be more correlated and corresponding connection strengths will be larger in respect to those neurons that belong to different factors. Hence, each group of neurons that forms a factor will be tightly connected via synaptic matrix and hence might correspond to the attractor of the network dynamics. Thus in order to perform the factorization one has to train the net with a set of complex patterns using Hebbian learning rule and search for the attractors it possesses. We investigated the conditions under which factors do form reliable attractors in sparsely encoded Hopfield-like neural network by means of computer simulation and analytical analysis as well.

## Model description

Let the neural network consist of  $N$  neurons of McCulloch-Pitts type (integrate-and-fire binary neurons) with gradually ranged synaptic connections between them. Only fully connected network is considered here. The number of active neurons  $n$  is kept constant for every network state vector. The ratio  $p = n/N$  characterizes the level of *sparseness*. Encoding is called sparse if  $p \ll 1$ . Detailed theoretical and computational analysis of sparsely encoded Hopfield-like neural networks was given elsewhere [1,6,7,8,14]. It was shown that sparsely encoded networks have a great advantage in informational capacity and the size of attraction basins compared to non-sparsely encoded ones.

Network was trained by a set of  $M$  patterns of the form

$\mathbf{X}^m = \bigvee_{l=1}^L \mathbf{a}_l^m \mathbf{f}^l$ , where  $\mathbf{f}^l \in B_n^{N-1}$  are  $L$  factors ( $N$  dimensional vectors) and for every  $m^{\text{th}}$  pattern  $\mathbf{X}^m \in B_C^L$  is the corresponding vector of factor scores.

As follows from the definition every factor consists of exactly  $n$  ones. In every complex pattern  $\mathbf{X}^m$  in turn the number of mixed factors is exactly equal to  $C$ , which is termed *complexity* of the input. The probability of one neuron in pattern be active is  $\mathcal{P}\{X_i^m = 1\} = 1 - (1-p)^C$ . For large  $C$  ( $pC \approx 1$ ) the level of input pattern activity is high, thus input signal is not sparsely encoded. We assumed factors and factor scores to be statistically independent. In a limit  $C=1$  patterns become pure factors and we obtain ordinary Hopfield case.

Connection matrix  $\mathbf{J}$  was formed using the correlational Hebbian rule:

$$\mathbf{J}_{ij} = \frac{1}{Np(1-p)} \sum_{m=1}^M (X_i^m - q\{X^m\})(X_j^m - q\{X^m\}), \quad (1)$$

$i \neq j, \mathbf{J}_{ii} = 0$

, where bias  $q\{X^m\} = \sum_{i=1}^N X_i^m / N$  is the total activity of the  $m^{\text{th}}$  pattern. Such form of bias enhances informational properties of the network and corresponds to the biologically plausible global inhibition being proportional to overall neuronal activity.

On the recall stage, on presentation of initial pattern,

---


$$^1 B_n^N = \{\mathbf{X} \mid X_i \in \{0,1\}, \sum_{i=1}^N X_i = n\}$$

the network was let to evolve until it stabilizes in some attractor. The evolution of the network's state is determined by the synchronous dynamics equation for activity  $\mathbf{X}$  in time:

$$\begin{aligned} X_i(t+1) &= \Theta(h_i(t) - T(t)) \\ X_i(0) &= X_i^{\text{in}}, i=1, \dots, N \end{aligned} \quad (2)$$

, where  $h_i(t) = \sum_{j=1}^N \mathbf{J}_{ij} X_j(t)$  is synaptic excitation,  $\Theta$  - step function, and  $T(t)$  - the activation threshold. The threshold  $T(t)$  is chosen at each time step in such a way that network activity is kept constant and equal to  $n$ . Thus, on each step  $n$  "winners" (neurons with greatest synaptic excitation) are chosen and only they are active on the next step. This procedure ensures that attractors are only fixed point or cyclic of length two. The stable pattern (point attractor or first pattern of cyclic attractor) was taken as a resulting pattern (further termed as final pattern  $\mathbf{X}^f$  of the recall process). In order to test the network's ability to perform factorization function i.e. to form attraction basins around factors encoded in the complex signal, we used factors as initial network states, i.e.  $\mathbf{X}^{\text{in}} = \mathbf{f}^l$ .

Aiming to analyze the factorization ability informational properties of the network will be considered in respect to factors and not to presented complex patterns. Thus recall quality is measured by overlap between recalled factor  $\mathbf{f}^l$  and final vector  $\mathbf{X}^f$

$$m^f = m(\mathbf{f}^l, \mathbf{X}^f) = \frac{1}{N} \sum_{i=1}^N (f_i^l - p) X_i^f.$$

As a measure of the relative informational loading we use entropy of set of stored factors relative to total number of modifiable synapses in the network.

$$\mathbf{a} = Lh(p)/N,$$

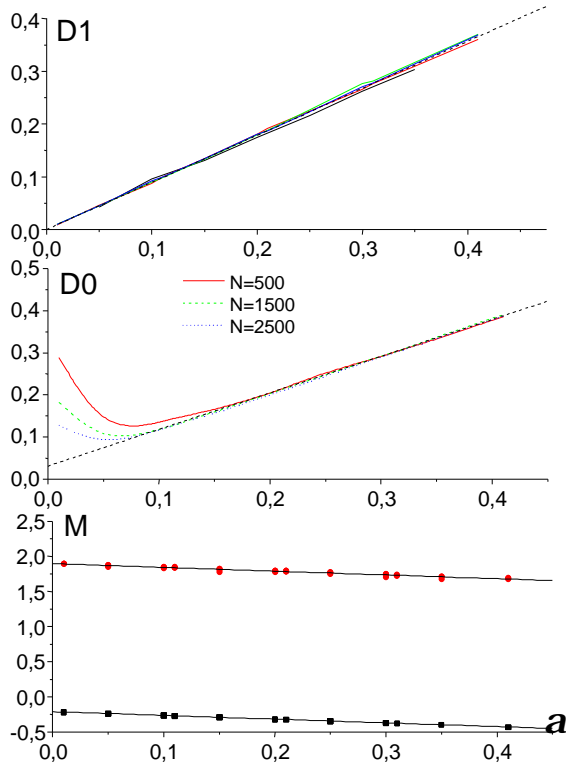
where  $h(p) = -p \log_2 p - (1-p) \log_2 (1-p)$  is the Shannon function. The informational capacity of a network  $\mathbf{a}_{cr}$  is a maximum  $\mathbf{a}$  for which stable states in the vicinities of stored factors still exist.

Since there is an analogy between conventional way of information encoding (pattern-by-neurons) and higher hierarchical (complex-pattern-by-factors, factors-by-neurons) scheme that we utilize here one could expect, the parameter  $p_f = C/L$  characterizing level of sparseness of pattern in respect to factors (relative complexity of the pattern) to be critical parameter for the network behavior. Surprisingly, this turned out not to be the case. By contrast, the absolute parameter  $C$  is most important. It will be shown later that this parameter frozen the increase of network size doesn't

result in sharp change of informational and dynamical properties of the network. Therefore, five major parameters  $p$ ,  $C$ ,  $\mathbf{a}$ ,  $N$ ,  $M$  can be treated as independent.

### Signal/noise analysis

One of the approximations used to study the network neurodynamics is the signal/noise analysis of the synaptic excitation  $h_i$  at the first step of network dynamics. The distribution of  $h_i$  has two distinct modes at  $M_1$  and  $M_0$ , that correspond to 1 and 0 in the testing pattern. Both modes are considered to have normal distribution with deviations  $\mathbf{s}_0$  and  $\mathbf{s}_1$  respectively. Since nonlinear dynamics of the network is provided by the threshold-like cutting of the fixed number of winners in synaptic excitation distribution it is useful to treat net deviation  $\mathbf{s}_0 + \mathbf{s}_1$  as a measure of noise, and a gap between modes  $\Delta M$  as a signal.

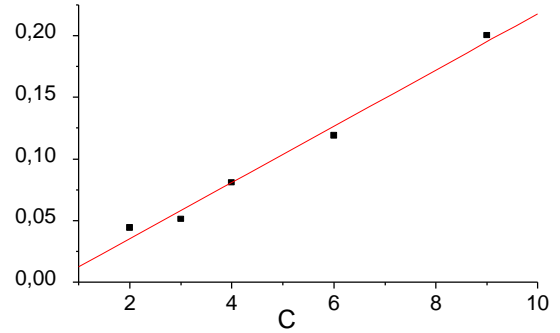


**Figure 1.** Experimental curves for dispersions and expectations of  $h_i$  for  $C=4$ ,  $p=0.1$  for  $N$  from 500 to 2500

Thus for the pattern to be recognized by the network, signal to noise ratio  $\Delta = \Delta M / (\mathbf{s}_0 + \mathbf{s}_1)$  must be high. Signal/noise ratio dependence on  $C$  and  $\mathbf{a}$  was

investigated theoretically and by means of computer simulation.

Results of the simulations of signal/noise are shown in Figure 1. By contrast to original Hopfield-like network two peculiarities can be noted: (1). For fixed  $N$  the signal/noise ratio will first increase when  $\mathbf{a}$  is small and then for larger  $\mathbf{a}$  will decrease, while traditionally it monotonically increases. (2) As  $N \rightarrow \infty$  and  $\mathbf{a} \rightarrow 0$   $D_0 = \mathbf{s}_0^2$  asymptotically approaches some nonzero value  $\tilde{D}_0$ .



**Figure 2** Dependence of noise/signal ratio  $\tilde{D}_0 / \Delta M$  ( $\tilde{D}_0 \xrightarrow{\mathbf{a} \rightarrow 0} 0$ ) for  $N \rightarrow \infty$  and  $\mathbf{a} \rightarrow 0$  on  $C$  for  $p=0.1$ .

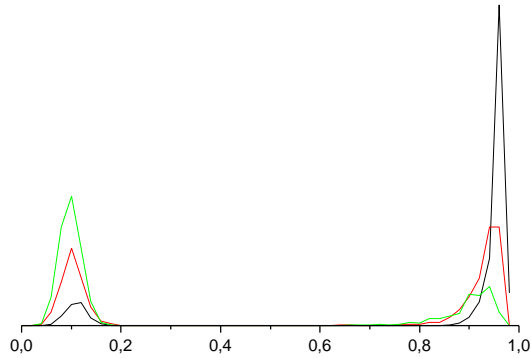
Growth of noise/signal ratio with respect to  $C$  (see Figure 2) suggests that there exist some critical level of complexity  $C_{cr}$ , at which there will be no factorization even for  $\mathbf{a} \rightarrow 0$ .

### Neurodynamics simulation

In order to analyze the dependency of  $\mathbf{a}_{cr}$  on parameters of the network the model was simulated on the computer. Particularly, the computer simulation was used to assess  $\mathbf{a}_{cr}$ . In order to do this we followed [2], and evaluated the probability  $P$  that a given factor has a stable state in its vicinity. The recall process ends when the activity either reaches a fixed point, a cycle of length 2. A stable pattern or the first pattern in a cycle was taken as the resulting pattern  $\mathbf{X}^f$  of the retrieval process. Calculations were performed for  $N$  from 500 to 10000, for  $p = 0.5, 0.1, 0.02$ . The program generated set of random factors and mixed them into the set of  $M$  patterns. Then network was trained with this set and tested with factors.

On the basis of results of computer simulation (number of trials was about 1000) the distribution of final overlaps  $m^f$  was plotted. See typical histogram of  $m^f$

in Figure 3. This distribution has two distinct modes:  $m^f \approx 1$  (“true”) and  $m^f \approx 0$  (“false”), that correspond to stabilization of the network in true and spurious attractors consistently. Besides these modes there exist a number of less evident modes which manifest at large  $N$ .



**Figure 3** Histogram of the final overlap for  $p=0.1$ ,  $C=5$ ,  $N=3000$  for  $\mathbf{a} = 0.15$  (black),  $0.2$  (red),  $0.25$  (green). Note that as  $\mathbf{a}$  increases activity redistributes from the “true” mode that corresponds to factors to “false” mode that corresponds to spurious states. The larger  $N$  the sharper this transition. It reflects the decrease of probability of correct recall  $P$ .

The probability of existence of stable attractors in the vicinities of factors was estimated by the probability of  $m^f$  to belong to the “true” mode:  $P = \mathcal{P}\{m^f > m_{thr}^f\}$ .

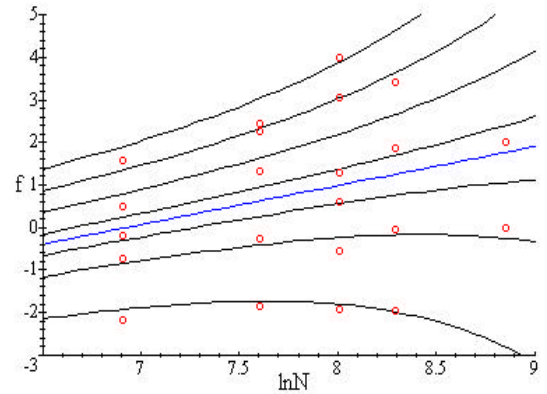
The threshold value  $m_{thr}^f$  used for separation was determined as a point of a minimum in case of balanced distribution between two modes. The obtained values of  $P$  corresponding to different  $p$  and  $C$  were approximated by the following logistic function:

$$P = 1/(1 + e^{-f}), \quad (3)$$

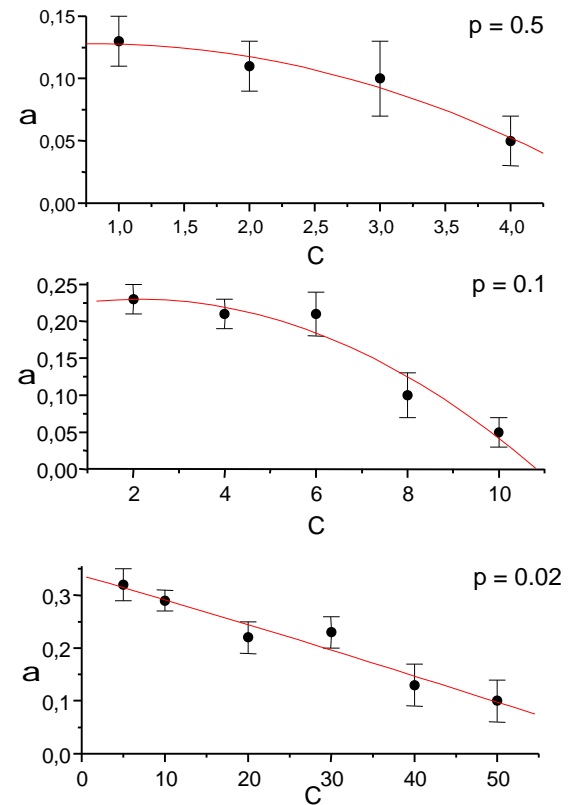
$$f = c_1 + c_2 \mathbf{a} + (c_3 + c_4 \mathbf{a} + c_5 \mathbf{a}^2)N + c_6 \ln N$$

The form of fitting function follows from early observation that the probability of correct recall falls down sharply at some value of  $\mathbf{a}$ . This effect is treated as a phase transition between correct recall phase and non-recall phase.

The family of curves is separated into two branches: concave (that tends to infinity) and convex (that tends to minus infinity) as  $N \rightarrow \infty$ . The separation line (in blue) in Figure 3 corresponds to  $\mathbf{a}$  that delivers zero value to the quadratic term at  $N$  of  $f$ . This value of  $\mathbf{a}$  we treated as  $\mathbf{a}_{cr}$ . Obtained evaluations of  $\mathbf{a}_{cr}$  were plotted for each  $p$  as a function of  $C$  (see Figure 4).



**Figure 3** Experimental data and fitting curves for  $p=0.02$ ,  $C=20$ . Each curve corresponds to some  $\mathbf{a}$ . Blue line indicates the point of phase transition from recall (upper) to non-recall (lower) phases of the network neurodynamics.



**Figure 4** Phase transition boundary curves in  $\mathbf{a} - C$  plane for different levels of sparseness  $p$ .

Notably for each  $p$  the value of  $\mathbf{a}_{cr}$  gradually declines as  $C$  approaches some critical value after which the factorization is impossible. This critical value is approximately inversely proportional to the level of activity in stored factors. We couldn’t find exact critical value of  $C$ , due to described above decrease of

signal/noise ratio as  $\mathbf{a}$  tends to zero for large  $C$ . To overcome this effect too large size of the network is required that is unreachable in computer simulations.

### Neurophysiological application

Pioneering work of Marr stimulated numerous theoretical and experimental studies of the hippocampus and its role in memory function. After Marr hippocampus was widely accepted to be intermediate-term memory storage and processing site [9,11]. Hippocampus seems to be responsible for complex associative memory tasks [13,16], formation of relationships among different memory items [5]. Due to its extensive collateral excitatory system and relatively sparse activity level the field CA3 of the hippocampus is thought to be natural autoassociator. [9,13,15]

We suppose that process of factorization proceeds in field CA3 of the hippocampus. Since polymodal incoming information is temporally and spatially redundant Marr suggested that some elementary factors must be extracted in hippocampus to be effectively stored as “classificatory units” in the neocortex. Following Buzsaki’s two-stage model of information recording [3,4], we consider “learning” and “extraction-recording” stages. On the first stage complex pattern of preprocessed theta-synchronized polymodal input from NC converges on tightly connected CA3 pyramids. Hebbian associative LTP-modification of synapses in CA3 forms sparsely bound cell ensembles (attractors) corresponding to factors that are encoded in incoming pattern. Synchronized gamma activity in CA3 and NC leads to weak modification of synapses on distal dendrites of neurons, which are currently active in NC. Thus, factors in CA3 are linked with future “classificatory units” in NC. On the next stage (during sleep) synchronized self-reactivation [17] of ensembles-factors in CA3 (SPW) and consequent “ripples” in CA1 result in strengthening of connections between coherently active “classificatory units” in NC as well as enhancement of these “replayed” factors.

### Conclusions

As far as we know, this is the first attempt to implement the idea of factorization (prompted by Marr) through physiological model and test its possibility on ANN. The possibility of nonlinear factorization in associative network was shown analytically and by computer simulation. The proposed framework can be easily checked in behavioral experiments. The idea of

factorization can be fruitfully applied to various memory tasks, e.g. spatial navigation.

### Acknowledgements

This work was done under grant from the Grant Agency of Czech republic No. 201/97/0885.

### References

1. Amari S, Maginu K (1988). Statistical neurodynamics of associative memory. *Neural Networks*, **1**, 63-73.
2. Amit DJ, et al (1987). Statistical mechanics of neural networks near saturation. *Annals of Physics*, **173**, 30-67.
3. Buzsaki, G (1989). A two-stage model of memory trace formation: a role for "noisy" brain states. *Neuroscience*, **31**, 551-570.
4. Buzsaki, G (1996). The hippocampo-neocortical dialogue. *Cerebral Cortex*, **6**, 81-92.
5. Eichenbaum H et al (1994). Two component functions of the hippocampal memory system. *Behav Brain Sci*, **17**, 449-518.
6. Frolov AA, Muraviev IP (1993). Informational characteristics of neural networks capable of associative learning based on Hebbian plasticity. *Network*, **4**,495-536.
7. Frolov AA, Husek D, Muraviev IP (1997). Informational capacity and recall quality in sparsely encoded Hopfield-like neural networks: analytical approaches and computer simulation. *Neural Networks*, **10** (5), 845.
8. Horner H, Bormann D, Frick M, Kinzelbach H, Schmidt A (1989) Transients and basins of attraction in neural network models, *Z. Phys.B – Condensed Matter* **76**, 381-398
9. Gluck MA (1997). Physiological models of hippocampal function in learning and memory. *Neurobiology of Learning and Memory*, **11**.
10. Hopfield JJ (1982). Neural networks and physical systems with emergent collective computational abilities. *PNAS*, **79**, 2544-2548.
11. Kavanau JL (1996). Memory, sleep and dynamic stabilization of neural circuitry: evolutionary perspectives. *Neurosci Biobehav Rev*, **20** (2), 289-311.
12. Marr D (1970). A theory for cerebral neocortex. *Proc R Soc Lond*, B **176**, 161-234.
13. Marr D (1971). Simple memory: a theory for archicortex. *Phil Trans R Soc Lond*, B **262**, 24-81.
14. Okada M (1996) Notions of associative memory and sparse coding. *Neural Networks*, **9**(8) 1429-

15. Rolls ET (1996). A theory of hippocampal function in memory. *Hippocampus*, **6**, 601-620.
16. Sutherland RJ, Rudy JW (1989) Configurational association theory: the role of the hippocampal formation in learning, memory, and amnesia. *Psychobiology*, *17*, 129-144
17. Wilson MA, McNaughton BL (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, **265**, 676-679.

Filename: IJCNN\_Full  
Directory: D:\ANTON\CONFERENCES\IJCNN'99  
Template: C:\Program Files\Microsoft Office\Templates\NORMAL.DOT  
Title: Nonlinear factorization in neural networks of the hippocampus  
Subject:  
Author: Anton Sirota  
Keywords:  
Comments:  
Creation Date: 01.05.99 15:07  
Change Number: 21  
Last Saved On: 10.05.99 18:08  
Last Saved By: Anton Sirota  
Total Editing Time: 1 537 Minutes  
Last Printed On: 10.05.99 18:09  
As of Last Complete Printing  
Number of Pages: 6  
Number of Words: 2 807 (approx.)  
Number of Characters: 16 004 (approx.)