Evaluation of the Informational Capacity of Hopfield Network by Monte Carlo Simulation

Húsek, Dušan
1994

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Evaluation of the Informational Capacity of Hopfield Network by Monte Carlo Method Simulation

Dušan Húsek   A.A. Frolov

Technical report No. V-606

September 25, 1994

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic
phone: 66-05-32-30 fax: 858-57-89
e-mail: dusan@uivt.uivt1.cas.cz

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Evaluation of the Informational Capacity of Hopfield Network by Monte Carlo Method Simulation

Dušan Húsek A.A. Frolov[1]

Technical report No. V-606
September 25, 1994

Abstract

In our work we investigate the informational capacity of Hopfield fully connected neural network (NN) with ± 1 neurons' states by computer simulation. The results of our computer experiments were approximated by formulae which are ordinarily used in analytical approaches. It allowed us to extrapolate the results of computer simulation to the case of thermodynamic limit when the number of NN neurons $N \to \infty$. Some results of such extrapolation were compared with those which have been obtained earlier. These two kinds of results were found to be close. Information amount which is extracted from the network due to a certain proposed decoding procedure was calculated. It was shown that this procedure allows to extract from the network amount of information $I \approx 0.12N^2$ that is close to the well known high limit of information capacity evaluated by thermodynamic approach on the base of spin-glass theory.

Keywords

Neural network, Hopfield neural network, computer simulation, theory of information, Shanon function, spin glass model

[1]This work was done under the grant from the Grant Agency of Czech Republic, Prague No.201/94/079

1 Introduction

Since the Hopfield paper appeared [1] Hopfield NN has become one of the most extensively investigated object of the NN theory. Of main interest has been its informational properties. It is beyond belief but the evaluation of the amount of information extracted from the Hopfield Network for any retrieval procedure has been absent up to now. Here we evaluate this amount for a certain natural decoding procedure described in the following.

Hopfield NN is one of the models of the autoassociative memory. During learning phase the set of patterns $X^l (l = 1, \ldots, L)$, which we will call etalons, are sequentially presented to NN and stored by modification of the connection matrix T.

During the retrieval phase some patterns X' are taken as NN initial states which are treated as NN inputs. Then NN dynamics tend to some stable states X'' which are treated as NN outputs. If input patterns X' are close to stored etalons (in the sense of Hamining distance) they are called *familiar*. In the opposite case they are called *novel*.

Due to some decision rule NN recognizes familiar and novel patterns among input patterns X'. If X' is recognized as familiar then output pattern X'' is assumed to be a correction of X'. The example of one of the decision rules is described below.

Thus decoding procedure consists of two functions: recognition and correction. Usually only information which is extracted from the network due to correction is taken into account [2]. But it has been shown [3] that it leads to a high underestimation of the total information which can be extracted from the memory. Here we calculate amounts of information extracted both by recognition and correction on the basis of computer simulation.

2 Model Description

In the case of fully connected Hopfield network the elements of connection matrix are defined by equations

$$T_{i,j} = \frac{1}{N} \sum_l X_i^l X_j^l, \quad T_{i,i} = 0 \;\; i,j = 1,..,N, \;\; l = 1,..,L$$

where X^l are bipolar vectors written into the memory. Let these vectors be called etalons. Their elements are assumed to be statistically independent and equal to 1 or -1 with probability $p = 0.5$.

The discrete time neural network dynamics is defined by equation

$$X_i(t+1) = \text{sgn}\,(\sum_j T_{i,j} X_j(t)) \qquad i,j = 1,\ldots,N$$

where $\text{sgn}(\eta) = \begin{cases} +1 & \text{if } \eta > 0 \\ \text{unchanged} & \text{if } \eta = 0 \\ -1 & \text{if } \eta < 0 \end{cases}$

It is known that for asynchronous mode (that means that at each time step only one neuron can change its state) this activity starting from some initial state X' reaches one of the stable states X''. In our model the sequence of neurons which may change their states is given.

When relatively few etalons are stored, the stable states X'' will correspond to these etalons. But there exists a large number of spurious stable states i.e. states which are far from written etalons. The measure of distance between two network states X_1 and X_2 is given by the formula

$$m(X_1, X_2) = \frac{1}{N}\sum_{j=1}^{N} X_{1,j}X_{2,j},$$

where m is called overlap and is ranged between -1 and 1.

It is evident that network informational properties depend on

i. probability P_1 that given etalon has in its vicinity a stable state

ii. mean distance between etalons and their nearest stable states m_s

iii. shape and size of etalons' basins of attraction

iv. number of spurious stable states M

v. shapes and sizes of spurious stable states basins of attraction

Most of these properties of the Hopfield network have already been investigated (see for example [4], [5]). The main goal of this work is to evaluate how these properties affect information which is extracted from the memory under considered decoding procedure and to find by computer experiments asymptotical equations for informational properties in the case of $N \to \infty$.

3 Mathematical Background

The simplest and natural decoding procedure which is based on the ability of the network to produce stable states near written etalons with rather large basins of attraction is the following. Let random network states X' be sequentially taken as its initial states. Let pattern X' be assumed as familiar for one of the stored etalons X^l if $m(X^l, X') \geq m_{\text{th}}$. Let pattern X' be recognized as familiar if $m(X', X'') \geq m_{\text{th}}$ where X'' is a stable state to which network activity tends from initial state X'. This decision rule can be easily realized by some auxiliary neural network.

To calculate information extracted from the network due to recognition of one of the familiar input pattern let us introduce two binary variables ω and ω' and assume that $\omega = 1$ if X' is familiar to one of the stored etalon $X^l, \omega = 0$ otherwise. Let $\omega' = 1$ if X' is recognized as familiar for X^l, $\omega' = 0$ if X' is recognized as unfamiliar for X^l.

Recognition	ω	ω'
Correct recognition of familiar pattern	1	1
Missing familiar pattern	1	0
Correct recognition of novel pattern	0	0
False recognition of novel pattern	0	1

Thus the sequence of ω assigns familiar and unfamiliar input pattern for X^l when nothing is known about the relations between X' and X'' and the sequence ω' assigns them under the condition that decision device recognized them as familiar or unfamiliar.

According to definition [6] the amount of information which is necessary to find the value $\omega = 1$ among the sequence of ω is equal to

$$I = \mathrm{ld}(1/P)$$

where P is the probability that $\omega = 1$ in the sequence of ω that is the probability that random pattern X' is familiar for X^l, $\mathrm{ld}(x)$ is the logarithm of x on the base 2. Since the number of neurons with equal activities for two patterns with overlap m is given by formula $(1+m)N/2$ then

$$P \cong C_N^{\frac{(1+m_{\mathrm{th}})N}{2}}/2^N$$

then

$$I = N(1 - h((1+m_{\mathrm{th}})/2))$$

where $h(x) = x\mathrm{ld}(1/x) + (1-x)\mathrm{ld}(1/(1-x))$ is the Shannon function. The amount of information which is necessary to find the value $\omega = 1$ among the sequence of ω corresponding to $\omega' = 1$ is equal to

$$I' = \mathrm{ld}(1/P')$$

where P' is the probability that $\omega = 1$ given that $\omega' = 1$, i.e. the probability that X' which is recognized as familiar is really familiar. Then information which is extracted from the network due to recognition of one of the familiar input patterns for one of the stored etalons is equal to

$$J_1 = I - I' = N(1 - h((1+m_{\mathrm{th}})/2) - \mathrm{ld}(1/P')/N).$$

The whole information which is extracted from the network due to recognition of familiar inputs for all etalons is equal to

$$I_1 = LP_1J_1 = LNP_1(1 - h((1+m_{\mathrm{th}})/2) - \mathrm{ld}(1/P')/N) \tag{1}$$

where P_1 is the probability that given etalon has a stable state in its vicinity. Then information which is extracted from one synaptical connection is given by

$$E_1 = I_1/N^2 = \alpha P_1(1 - h((1+m_{\mathrm{th}})/2) - \mathrm{ld}(1/P')/N) \tag{2}$$

where $\alpha = L/N$.

Let X' be familiar for X^l. The amount of information which is extracted from NN due to correction of one of the familiar inputs for one of the stored etalons is given by the formula

$$J_2 = I(X^l, (X', X'')) = H(X^l|X') - H(X^l|X', X'')$$

where

$$H(X^l|X') = Nh((1+m_{\rm th})/2)$$

and

$$H(X^l|X',X'') \cong H(X^l|X'') = Nh((1+m_{\rm st})/2)$$

where $m_{\rm st}$ is the average overlap between etalon and the stable state in its vicinity if it exists. We can put $H(X^l|X',X'') \cong H(X^l|X'')$ because usually $1-m_{\rm st} \ll 1-m_{\rm th}$ that is X'' is much closer to X^l than X'. If $m_{\rm th} \cong m_{\rm st}$ correction is not performed at all.

The whole information which is extracted from the network due to correction is equal to

$$I_2 = LP_1J_2 = LNP_1(h((1+m_{\rm th})/2) - h((1+m_{\rm st})/2)) \tag{3}$$

and

$$E_2 = I_2/N^2 = \alpha P_1(h((1+m_{\rm th})/2) - h((1+m_{\rm st})/2)) \tag{4}$$

per one synaptical connection.

The total information which is extracted from the memory due to both recognition and correction is equal to

$$I = I_1 + I_2 = LNP_1(1 - h((1+m_{\rm st})/2) - {\rm ld}(1/P')/N) \tag{5}$$

and

$$E = E_1 + E_2 = \alpha P_1(1 - h((1+m_{\rm st})/2) - {\rm ld}(1/P')/N) \tag{6}$$

per one synaptical connection.

Hence NN informational characteristics are expressed now in terms of P_1, P' and $m_{\rm st}$ whose dependence on α and $m_{\rm th}$ can be estimated by Monte-Carlo simulations.

4 Computer Simulation

4.1 Evaluation of P_1 and $m_{\rm st}$

To evaluate the probability P_1 that given etalon has a stable state in its vicinity and mean overlap $m_{\rm st}$ between etalon and such stable state similarly to [7] the etalons themself were used as initial network states. The final stable states were assumed to be unspurious if their overlaps with initial states were greater than m'. Computer experiments were carried with 10 different random sets of etalons for each values L and N. Each written etalon was tested for the presence of a stable state in its vicinity. Thus each value of P_1 was calculated by $10L$ trials and each value of $m_{\rm st}$ was calculated by $10PL$ trials.

Probability distribution of overlaps between etalons and stable states in the vicinity of $m = 1$ is shown in Fig. 1. It is seen that there exist two modes of this distribution with just different dependencies on α and N. The mode with the larger values of m decreases under increasing of α and N while the mode with the smaller values of m increases under increasing of α and does not change under increasing of N. Such separation of probability distribution of m into two modes has not been noticed earlier [7]. Thus may be due to an overly large step used for histograms drawing.

These two modes are separated by $m = 0.97$ so we have taken $m' = 0.97$. It means that stable states have been treated as unspurious only if $m \geq 0.97$.

Since a preliminary analysis showed that the dependence of P_1 on α should be approximated by logistic function it is more convenient to express it in the form

$$P_1 = 1/(1 + \exp(F(\alpha))) \tag{7}$$

where $F(\alpha)$ is supposed to be linear function. Dependence of $F = \ln(1/P_1 - 1))$ on α for different values of N is shown in Fig. 2. It is seen that dependence of F on α is really very close to linear. As shown in Fig. 2a all approximating straight lines cross at the point $\alpha = \alpha_{\mathrm{cr}} \cong 0.14$. Thus we may put

$$F = A(N)(\alpha \ - \alpha_{\mathrm{cr}}) + B \tag{8}$$

where $B \cong -2.2$. Dependence A on N is shown in Fig. 2b. Under increasing of N it tends to the straight line with the slope $C \cong 0.1$. In [7] dependence F on α and N has been approximated by the same formula (7) but with other coefficients: $C \cong 0.028$, $B \cong -3.5$. This discrepancy may be affected by the influence of the second mode in probability distribution of m which has not been separated from another mode in calculation of P_1 in [7]. As was mentioned above, this mode with smaller values of m has not decreased under increasing of N in the used region $N \leq 300$ and $\alpha > \alpha_{\mathrm{cr}}$. Probability P_2 that the overlap m between etalon and the correspondent stable state is in the region $0.92 \leq m \leq 0.97$ is shown in Fig. 3. It may be under $N \gg 300$ this probability decreases and its dependence on a and N results in the values of B and C obtained in [7]. But it must be noted that our results confirm the main conclusion made in [7] that under $N \to \infty$, the dependence of P_1 on α tends to the step function: $P_1 = 1$ for $\alpha < \alpha_{\mathrm{cr}}$ and $P_1 = 0$ for $\alpha > \alpha_{\mathrm{cr}}$.

The obtained dependence of m_{st} on α for different values of N is shown in Fig. 4. Limit dependence of m_{st} on α for $N \to \infty$ calculated in [7] with the use of replica symmetric theory is also shown in Fig. 4. It is seen that the results of simulation tend to limit curve under increasing N but yet they are far from it for $N = 300$.

4.2 Evaluation of basins of attraction

To evaluate probability $q(m_{\mathrm{th}})$ that given initial state X' which has overlap with given etalon X^l $m(X', X^l) > m_{\mathrm{th}}$ converges to stable state in its vicinity (if it exists) five different random noisy variants of each etalon with given initial overlap $m(X', X^l) = m_{\mathrm{th}}$ were sequentially used as initial network states. In other respects experiments were the same as described in the previous section. Thus each value of $q(m_{\text{th}})$ was calculated by $50PL$ trials.

The obtained dependence of q on α and N under the different values of m_{th} has the same form as similar dependence for P_1. This dependence also may be approximated by formula (7) and (8) where $A(N) \cong CN + D$. Dependence of α_{cr}, B, C and D on m_{th} are shown in Fig. 5. For $N \to \infty$ we may assume that

$$q = \begin{cases} 1 & \text{if } \alpha < \alpha_{\text{cr}}(m_{\text{th}}) \\ \\ 0 & \text{in the opposite case.} \end{cases}$$

The obtained dependence of α_{cr} on m_{th} is close to one obtained in the paper [5] by macrodynamic approximation of retrieval procedure. Inverse dependence which determines the border of basin of attractivity for given α is shown in Fig. 6. in comparison with the similar dependence obtained in [5].

4.3 Amount of information extracted from the network by correction

To evaluate this amount by formula (4) we use the limit dependencies of $m_{\text{st}}(\alpha)$ calculated in [7] under $N \to \infty$ and shown in Fig. 4, 6 and the dependence of $m_{\text{th}}(\alpha)$ which determines the basin of attractivity for given α, also shown in Fig. 6. It may be expected with high probability that for initial overlap exceeding $m_{\text{th}}(\alpha)$ for given stored pattern NN activity will converge to a stable state having the overlap $m_{\text{st}}(\alpha)$ with this pattern. In the opposite case NN activity will converge to some spurious stable state and no information will be extracted from the memory. Thus for $\alpha < \alpha_{\text{cr}}(m_{\text{th}})$ information E_2 must be calculated by formula (4) and for $\alpha > \alpha_{\text{cr}}(m_{\text{th}})$ it must be put to be zero.

Dependence of E_2 on α calculated by this way for $m_{\text{th}} = 0.4$ is shown in Fig. 7.

The maximal values $E_2(\alpha)$ calculated by formula (4) for minimal value $m_{\text{th}}(\alpha)$ corresponding to the border of basin of attractivity for given α are shown in Fig. 6. It is seen that maximal amount of information which can be extracted from one synapse due to correction reaches at $\alpha \cong 0.12$ and $m_{\text{th}} \cong 0.4$ value $E_2 \cong 0.1$.

4.4 Amount of information extracted from the network by recognition

This amount (see formulae (1) and (2)) is determined by the probability P' that input pattern X' which is recognized as familiar is really familiar. Probability P' was calculated for different values m_{th} by M trials for each given set of etalons which then was averaged by 10 different random sets of etalons for each combination of values N and L. Unfortunately increase of m_{th} causes exponential decrease of probability that given trial with X' as initial state will provide the stable state X'' such that $m(X', X'') > m_{\text{th}}$. That is, the number of trials M which is necessary for accurate evaluation of P' must exponentially increase. So we succeed in evaluation of P' only for $m_{\text{th}} \leq 0.4$.

Dependence of $I' = (1/N)\text{ld}(1/P')$ on α for different values of N is shown in Fig. 8. It is seen that there exist the range of α and m_{th} where the value of P' is close to 1 in spite of the large number of spurious states which is known to exist for used values of N and L. It means that in this range of α and m_{th} attractiveness for spurious states are rather small in comparison with the attractiveness of stored patterns, that is the shape of basins of attractivity for spurious states is very complex and as a rule NN activity beginning from some initial state tends not to the nearest spurious stable state but to one of the farther stable states having the overlap with initial state exceeding m_{th}. On the contrary the shape of basins of attraction for stored patterns is rather simpler and their borders are determined only by initial overlap m_{th} shown in Fig. 6.

It is also seen that dependence of I' on α can be approximated by formula

$$I' = \begin{cases} 0.2(\alpha - \alpha'(m_{\text{th}})) & \text{for } \alpha > \alpha'(m_{\text{th}}) \\ \\ 0 & \text{for } \alpha < \alpha'(m_{\text{th}}) \end{cases} \tag{9}$$

where dependence $\alpha'(m_{\text{th}})$ coincides approximately with the border of basins of attractivity for stored templates shown in Fig. 6.

The amount of information which can be extracted from one synaptic connection due to recognition in the limit case $N \to \infty$ was calculated by formula (2) where I' was calculated by (9) and $\alpha'(m_{\text{th}})$ was evaluated from curve in Fig. 6. Corresponding dependence E_1 on α for $m_{\text{th}} = 0.4$ is shown in Fig. 7. It is seen that information which is extracted due to correction is much greater than for recognition. This is typical for NN sequential dynamics. For parallel dynamics the relation is opposite [3].

At last we can evaluate the information amount which may be extracted from the network due to both correction and recognition. Dependence of $E = E_1 + E_2$ on α for $m_{\text{th}} = 0.4$ is shown in Fig. 7. The limit dependence of E on α calculated by formula (6) is shown in Fig. 6. It is seen that maximal amount of information which can be extracted from the memory is obtained at $\alpha \cong 0.12$ and $m_{\text{th}} \cong 0.4$ reaches approximately value 0.12. This value is close to one calculated in [8] by formula (6) under the implicit presumption that the last term in this formula is negligibly small. Here we have shown that for special decoding procedure such presumption is valid.

5 Conclusions

Computer simulation of Hopfield network informational properties confirmed those which were obtained earlier by various analytical approximations [4], [5], [7], [8]. Furthermore it gave evaluation of attractiveness for true and spurious stable states. It was shown that there exist the range of α and m_{th} where attractiveness of spurious states falls much faster than the attractiveness of true states. This property of Hopfield network allows for the proposal of a special decoding procedure which provides the extraction of 0.12 bit of information per one synaptical connection.

It was shown that in the optimal case when the total information which is extracted from the memory is close to its maximum the greater part of information is extracted due to correction of input patterns. This is typical for multi-step retrieval of stored patterns and opposed to the case of single-step retrieval when the greater part of information is extracted by recognition of input patterns.

Bibliography

[1] Hopfield J.J.: Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. USA, 79, 1982, 2554-2558.

[2] Palm G.: On storage capacity of an associative memory with randomly distributed elements. Biol. Cyber. 39, 1981, 125-127.

[3] Frolov A.A.: Information properties of bilayer neuron nets with binary plastic synapses Biophysika 34, 1989, 868-76 (in Russian).

[4] Amit D.J., Gutfreund H., Sompolinsky H.: Spin-glass model of neural networks. Phys. Rew. A., 32, 1985, 1007-1018.

[5] Amari S.I., Maginu K.: Statistical neurodynamics of associative memory. Neural Networks, 1, 1988, 63-73.

[6] Claude E. Shannon: The Mathematical Theory of Communication. The Univ. of Illinois Pres, Urbana, 1949, 117p.

[7] Amit D., Gutfreund H., Sompolinsky H.: Statistical mechanics of neural networks near saturation. Annals of Physics 173, 1987, 30-67.

[8] Amit D., Gutfreund H., Sompolinsky H.: Information storage in neural networks with low levels of activity. Phys. Rev. A 35, 1987, 2293-303.

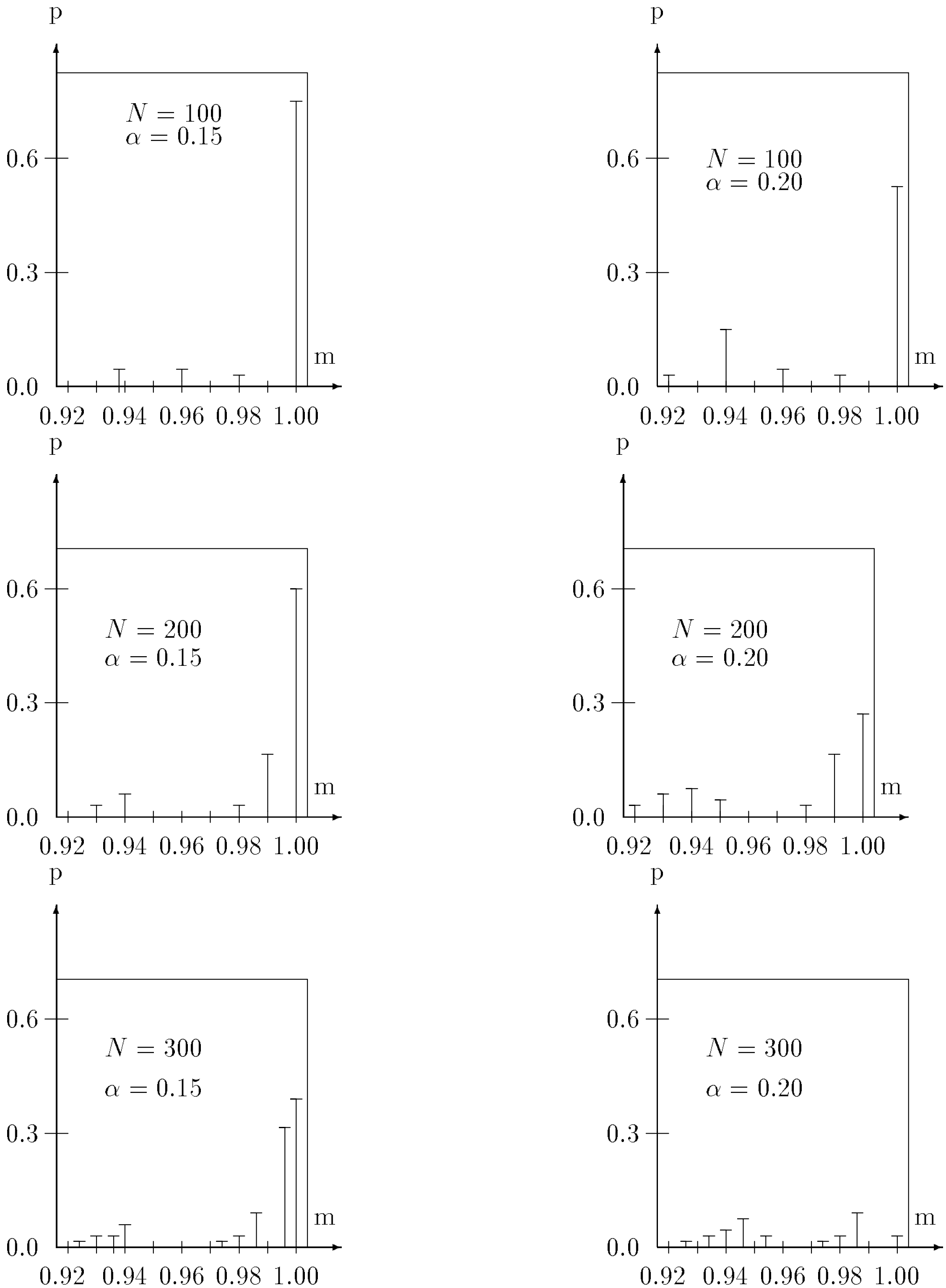


Fig. 1. *Probability distribution of overlaps* $m(X_i, X')$ *between stored patterns* X_i *and stable states in their vicinities for* $\alpha = 0.15$ *and 0.20 and* $N = 100, 200$ *and 300.*

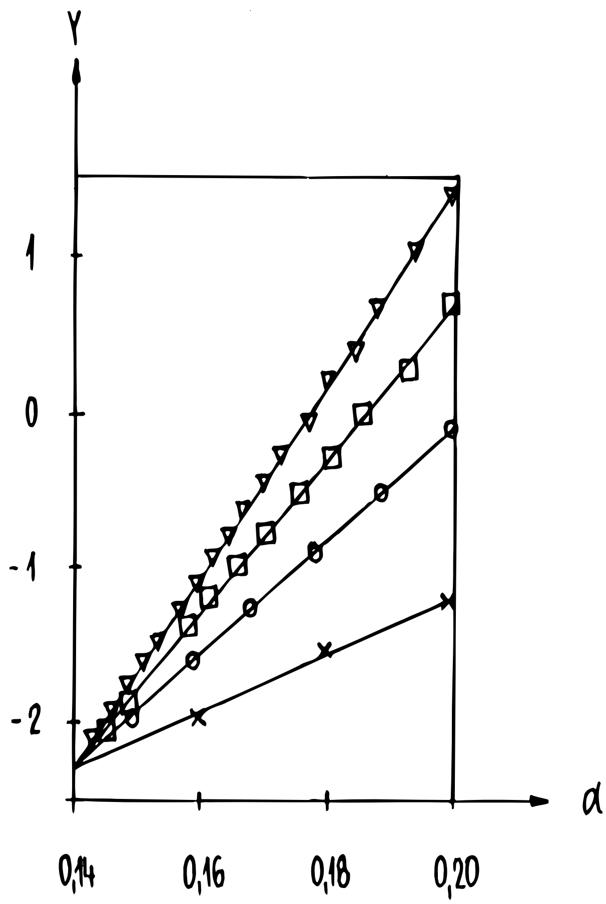


Fig. 2a. *Dependence of* $F = \ln(1/P_1 - 1)$ *on* α *for different values of* N*: a-, (circle) 100, (rectangle) 200 and (triangle) 300.*

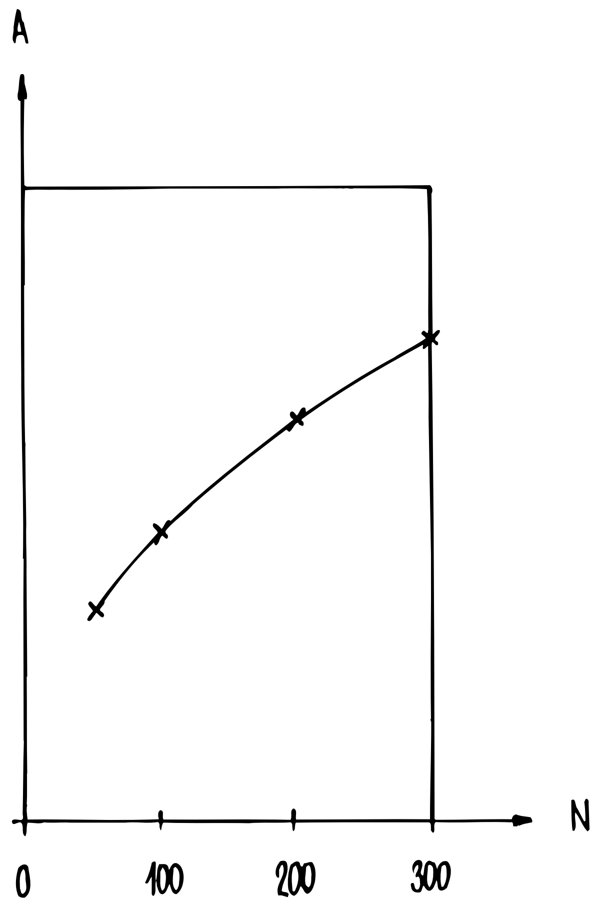


Fig. 2b. *Slopes of approximating straight lines for F in dependence on N.*

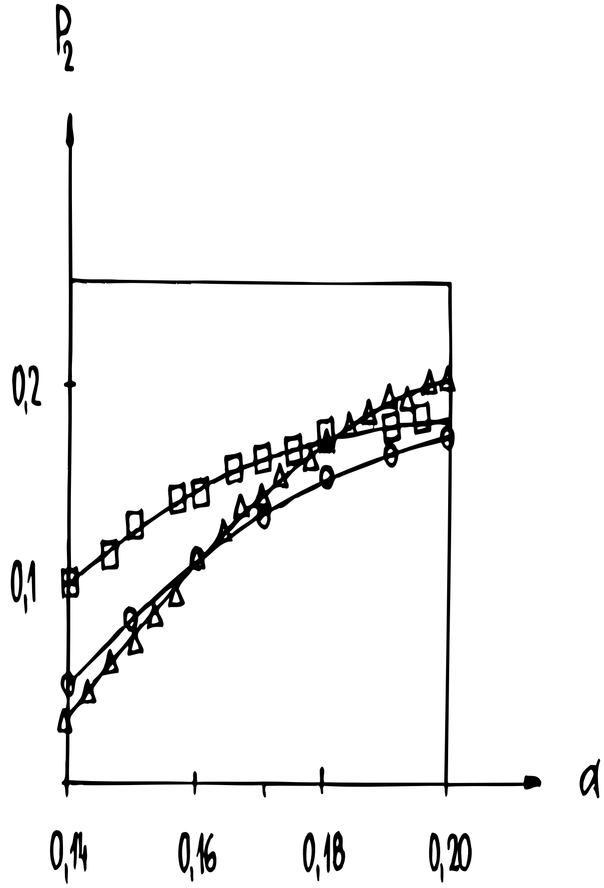


Fig. 3. *Dependence of P_2 on α for different values of N. (circle) 100, (rectangle) 200 and (triangle) 300.*

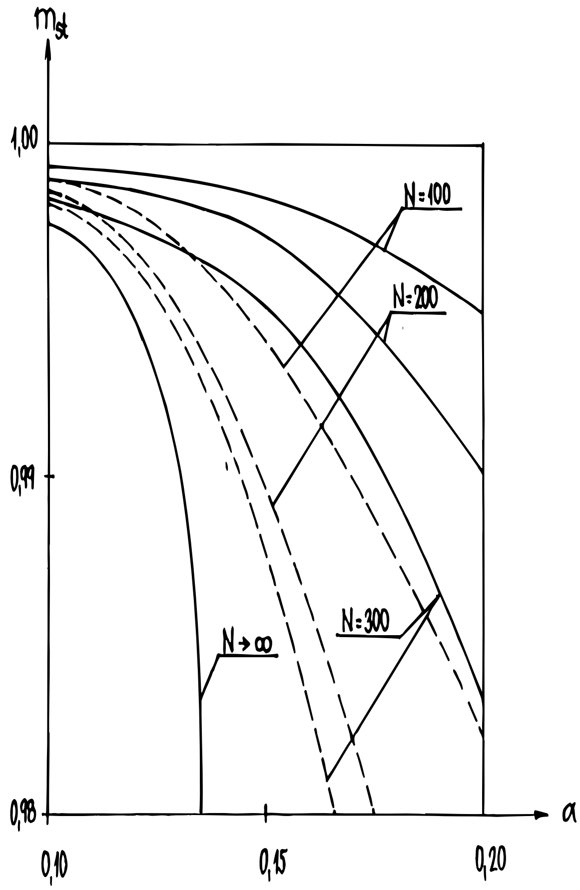


Fig. 4. *Mean overlaps between stored patterns and stable states in their vicinities in dependence on* α *and* N*. Solid lines for* $N = 100, 200$ *and 300 correspond to* $m' = 0.97$*, dashed lines correspond to* $m' = 0.92$*. A line for* $N \to \infty$ *is taken from [7].*

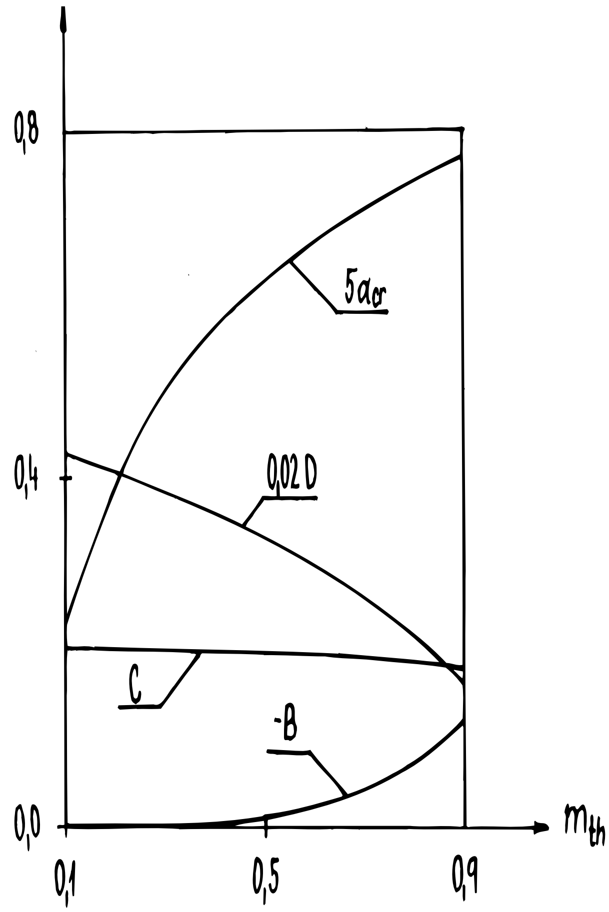


Fig. 5. *Parameters of basins of attractors for stored patterns in dependence on* α *(see text).*

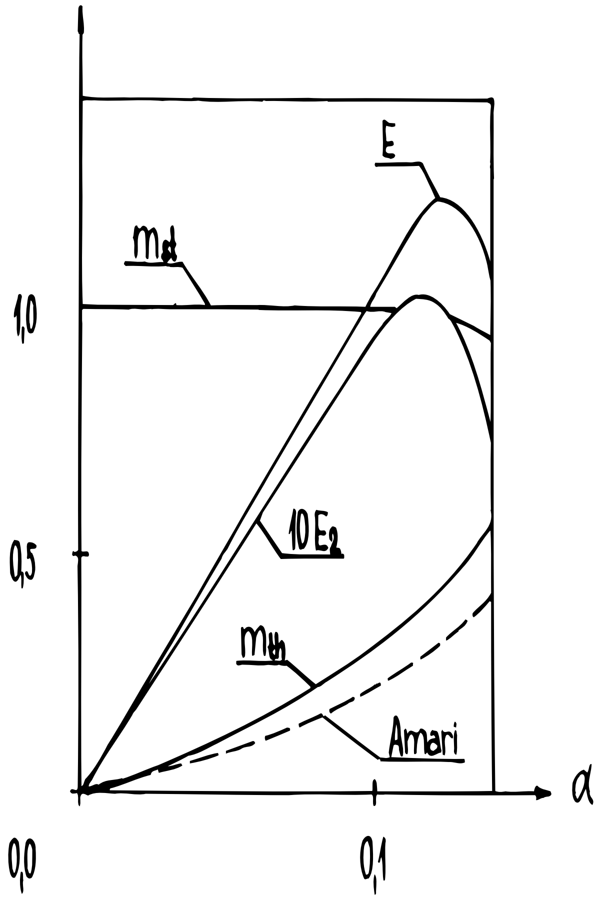


Fig. 6. *Mean overlap between stored patterns and stable states in their vicinities* $m_{\rm st}$*, the border of basins of attractors for stored patterns* $m_{\rm th}$*, limit information which can be extracted from one synapse due to correction* E_2 *and total limit information which can be extracted from one synapse due to both correction and recognition for* $N \to \infty$*. Dashed line is the border of basins of attractors taken from [5].*

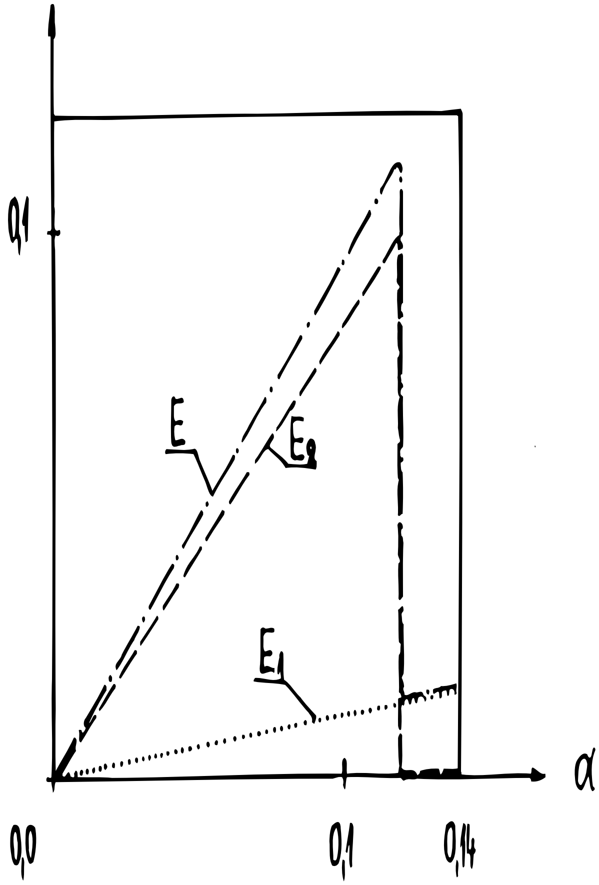
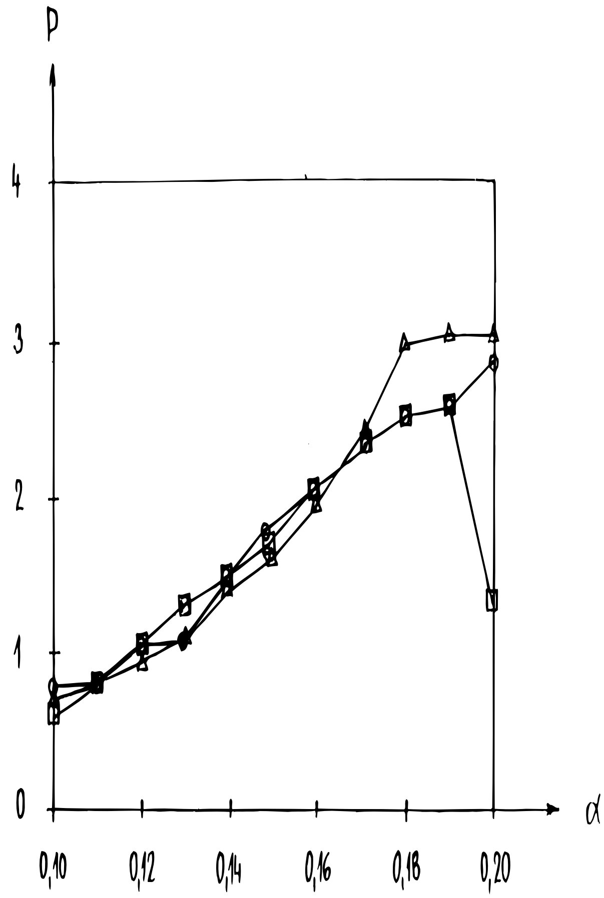
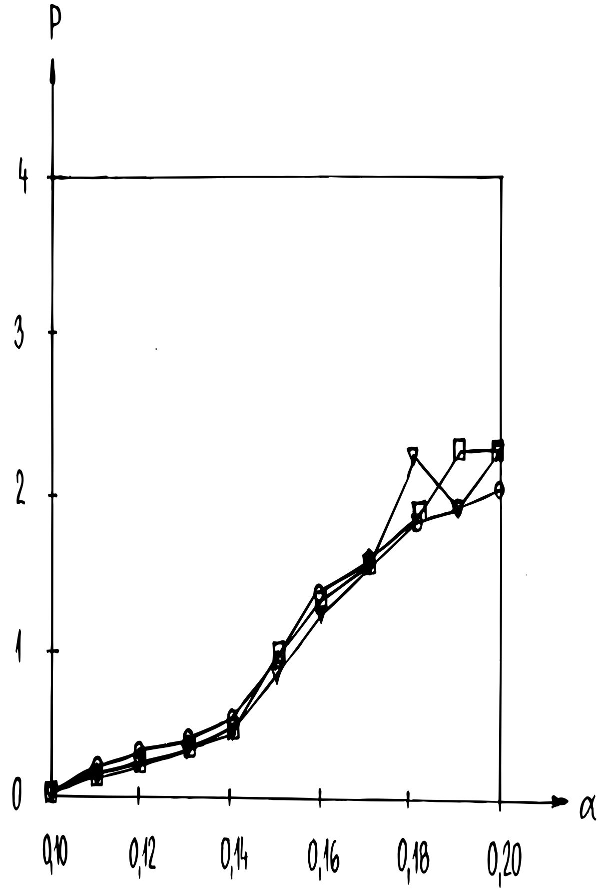


Fig. 7. *Information which is extracted from one synapse due to recognition* E_1, *due to correction* E_2 *and total* $E = E_1 + E_2$ *for* $m_{\rm th} = 0.4$.


P
4
3
2
1
0
α
0,10
0,12
0,14
0,16
0,18
0,20


P
4
3
2
1
0
α
0,10
0,12
0,14
0,16
0,18
0,20

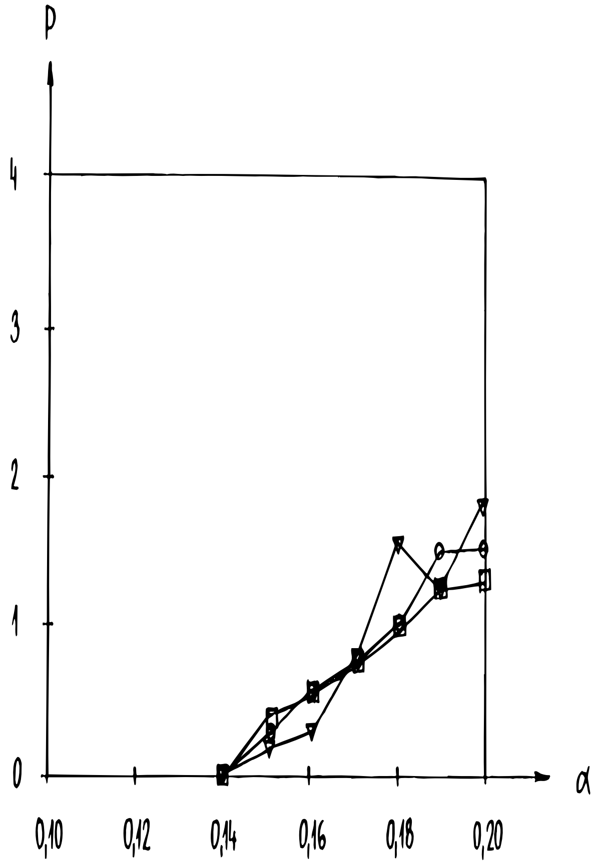


Fig. 8. *Dependence of I' on α for different values of N and $m_{\rm th}$, (triangle) N = 72, (circle) N = 84, (square) N = 96.*