ADAPTATION OF HOPFIELD ASSOCIATIVE MEMORY PARAMETERS IN STATISTIC TRAINING

B.V.Kryzhanovsky, M.V.Kryzhanovsky

Institute of Optical Neural Technologies, Russian Academy of Sciences

44/2 Vavilov Street, 117333 Moscow

A.Fonarev

Department of Engineering Science and Physics, The College of Staten Island

CUNY, SI, New York 10314

Abstract

The paper treats the issue of pattern recognition training in terms of Hopfield associative memory (HAM). The conventional randomization technique is used to determine the exponential extremity of HAM recognition error. The extremity exponent is considered as a function of the training process. In training, the exponent is shown to rise from $(1\text{-}2p)^4$ to $(1\text{-}2p)^2$ where p is the error coefficient at the output of the binary channel of observation ($0 \le p \le 1/2$). The HAM capacity grows in the same way.

1. Introduction

The neural model of HAM [1,2] interests researchers as one of possible approaches to explaining the phenomenon of determined behavior observed in initially chaotic neural structures [4-10].

The paper considers the Hopfield model as a recognition arrangement which can compare external events and place similar patterns to the same memory segment. Assume that M equal-probability events numbered as $m=1,2,...,M$ may occur in the external world. As the number of observations grows, a standard pattern (i.e. a particular fixed point in the M-dimensional attribute space corresponding to the m-th event) begins to form in the m-th memory segment. Correspondingly, the input signal begins asymptotically to look as a random departure from one of M fixed points, the recognition process boiling down to the bringing of the input signal to the original event m. In fact, the Hopfield network is a decoder based on the knowledge of all M fixed reference points of the attribute space. In the next paragraph we will briefly describe the decoder's construction and the method of deriving the upper limit of decoding (recognition) error probability. However, the way reference patterns develop and its mathematical essence are the questions to be answered. At present, adaptation of the Hopfield network to the external world is rather hard to describe in full. Here we constrict ourselves to simple consideration of a single input event, which corresponds to the first step in the formation of neural net interconnections. This allows us to compare the initial working parameters of Hopfield network to its characteristics in the asymptotically stable mode of recognition.

The further description is given in terms of the random coding theory, which permits us to connect problems of neural networks with issues of noise-proof coding studied in the field of probabilistic information theory. In the computations that follow we rely mostly on the technique developed in papers [11-13], which show that from the mathematical point of view, the recognition error probability for the randomized Hopfield model can be determined using the method of large deviation probability evaluation, which, in turn, is based on Chebyshev-Chernov exponential estimates [14].

2. Mean probability of recognition error

We consider a device that is designed for recognizing M equally probable events and can be represented as a five-link chain:

$$m \to \boldsymbol{x}_m \to \boldsymbol{y}_{\boldsymbol{m}} \to HAM \to \tilde{\boldsymbol{y}}_{\boldsymbol{m}} \to \tilde{m} \quad (1)$$

where *m* is the event number, $\boldsymbol{x}_m$ is the reference attribute vector of the *m*-th event, $\boldsymbol{y}_m$ is the apparent attribute vector (vector $\boldsymbol{x}_m$ distorted in the observation channel), and $\tilde{\boldsymbol{y}}_{\boldsymbol{m}}$ is the output signal. In (1) we use vectors notations:

$$\boldsymbol{x}_m = (x_{m1}, x_{m2}, \cdots, x_{mN}) \in \{-1,+1)^N \quad (2)$$

$$\boldsymbol{y}_{\boldsymbol{m}} = (x_{m1}\theta_1, \cdots, x_{mN}\theta_N) \in \{-1,+1\}^N \quad (3)$$

$$\tilde{\boldsymbol{y}}_{\boldsymbol{m}} = (x_{m1}\tilde{\theta}_1, \cdots, x_{mN}\tilde{\theta}_N) \in \{-1,+1\}^N \quad (4)$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_N)$ is the vector of multiplicative noise with N independent identically distributed random components:

$$\theta_i = \begin{cases} -1, & p \\ +1, & 1-p \end{cases} \quad , \quad 0 \le p \le \tfrac{1}{2} \quad , \quad i \in \overline{1,N} \quad (5)$$

Parameter *p* sets the area around reference point $\boldsymbol{x}_m$ where observations $\boldsymbol{y}_m$ experience scattering. The Hopfield neural network receives input vector $\boldsymbol{y}_m$ and reduces the noise level in the observation channel so that the output signal takes form (4) where $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \cdots, \tilde{\theta}_N)$ is the residual noise with much lower component intensity $\mathbf{Pr}\{\tilde{\theta}_i = -1\} << p$, $\forall i \in \overline{1,N}$. Here integer $\tilde{m} \in \overline{1,M}$ denotes a final decision the recognition device (1) comes to.

Now let us evaluate the decoding error $P \equiv \Pr(\tilde{m} \ne m)$. Let $\boldsymbol{B}$ be a *M*×*N* matrix whose rows are reference points of the attribute space:

$$\boldsymbol{B} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ x_{M1} & x_{M2} & \cdots & x_{MN} \end{bmatrix} \quad (6)$$

Papers [11, 12] show that averaged over the set of all equal-probability matrixes of form (6), the component intensity of residual noise (i.e. the probability of erroneous recognition of a particular *i*-th component of the input vector $P_i = P_i(N,M,p)$) meets the following estimate:

$$P_i \equiv \overline{\mathbf{Pr}\{\tilde{\theta}_i = -1\}} \le e^{-NS}, \ \forall i \in \overline{1,N} \quad (7)$$

where

$$S = \frac{M+1}{2}\ln K - M\ln\frac{K+1}{2} - \ln(pK+1-p) \tag{8}$$

$$K = \frac{-(M-1)(1-2p) + \sqrt{(M-1)^2 + 16Mp(1-p)}}{2p(M-1)} \tag{9}$$

This way, recognition error probability *P* in the randomized Hopfield model (i.e. the probability of output vector $\tilde{\boldsymbol{y}}_m$ differing from the reference vector $\boldsymbol{x}_m$ by at least one of *N* attributes) is given by expression:

$$P \equiv \overline{\mathbf{Pr}\{\tilde{\boldsymbol{y}}_m \neq \boldsymbol{x}_m\}} = \mathbf{Pr}\left\{\bigcup_{i=1}^{N}\left(\tilde{\theta}_i = -1\right)\right\} \leq \sum_{i=1}^{N} \overline{\mathbf{Pr}\{\tilde{\theta}_i = -1\}} \equiv \sum_{i=1}^{N} P_i \tag{10}$$

which, in view of (7), can be written as

$$P \leq Ne^{-NS} \tag{11}$$

Below we dwell on those points of the derivation of estimation (7) that will help us in the following discourse. We consider a conventional model of Hopfield memory with a symmetric interconnection matrix whose diagonal elements are set equal to zero. Given matrix $\boldsymbol{B}$ the ***i***-th component of the output vector (4) is defined as:

$$\tilde{y}_{mi} = \mathbf{sign} \sum_{j\neq i} T_{ij} y_{mj}$$

$$\mathbf{sign}\,\varepsilon = \begin{cases} -1, & \varepsilon < 0 \\ +1, & \varepsilon \geq 0 \end{cases} \tag{12}$$

$$\hat{\mathbf{T}} = \hat{\mathbf{B}}^T \hat{\mathbf{B}} - M\hat{\mathbf{I}}$$

where $\hat{\mathbf{T}}$ is the interconnection matrix of the neural network, $\hat{\mathbf{I}}$ is the unit matrix, and ***sign*** is the component threshold operation. The error probability for the ***i***-th component of the ***m***-th output vector is

$$P_i = \mathbf{Pr}\left\{\mathbf{sign}\left(\sum_{j\neq i}^{N} T_{ij} y_{mj}\right) \neq x_{mi}\right\} \tag{13}$$

Omitting a few intermediate transformations [12 ,13] and regarding (12), we can show that condition (13) can be rewritten in the form:

$$P_i \leq \mathbf{Pr}\left\{\sum_{j\neq i}^{N} \theta_j + \sum_{j\neq i}^{N}\sum_{n\neq m}^{M} \theta_j x_{nj} x_{ni} x_{mj} x_{mi} \leq 0\right\} \tag{14}$$

Note that (14) could have been expressed as a strict equality with function ***sign*** in (14) defined more carefully (by observing the complete symmetry around zero).

Further considerations are based on assumption that all elements of matrix (6) are independent and takes ±1 with equal probabilities. We also estimate the probability in the right side of equation (14) using the conventional Chebyshev-Chernov technique [14]. For simplicity, we substitute M and N for M-1 and N-1 in all expressions that follows. In addition, we assume that x_{mi} = -1. Since quantities x_{nj} in the double sum of (14) are statistically independent, so are quantities $\xi_{mnj} = \theta_j x_{nj} x_{ni} x_{mi} x_{mj}$. There are MN of such statistically independent quantities ξ_{mnj} in all, each of them taking ±1 with equal probabilities.

Let us replace the triple index of ξ_{mnj} by a single index ($mnj \to k$) $k \in \overline{1, MN}$. Then expression (14) take the form:

$$P_i \le \mathbf{Pr}\left\{\sum_{k=1}^{MN} \xi_k \ge \sum_{j=1}^{N} \theta_j\right\} \tag{15}$$

As shown in [14], for any $z \ge 0$

$$\mathbf{Pr}\left\{\sum_{k=1}^{MN} \xi_k \ge \sum_{j=1}^{N} \theta_j\right\} = \mathbf{Pr}\left\{\exp\left(z\sum_{k=1}^{MN} \xi_k\right) \ge \exp\left(z\sum_{j=1}^{N} \theta_j\right)\right\} \le$$

$$\le \overline{\exp\left(z\sum_{k=1}^{MN} \xi_k - z\sum_{j=1}^{N} \theta_j\right)} = \left[\overline{\exp(z\xi)}\right]^{NM} \left[\overline{\exp(-z\theta)}\right]^{N} = \tag{16}$$

$$= \left(\frac{e^z + e^{-z}}{2}\right)^{NM} \left[pe^z + (1-p)e^{-z}\right]^N$$

where the bar denotes averaging over all statistically independent variables ξ_k and θ_i. Hence we can evaluate the upper limit of component recognition error as:

$$P_i \le \min_{z \ge 0} \left\{\left(\frac{e^z + e^{-z}}{2}\right)^M \left[pe^z + (1-p)e^{-z}\right]\right\}^N \tag{17}$$

It is easy to show that the expression in the curly brackets reaches the minimum when z meets the equation:

$$p(M+1)e^{4z} + (1-2p)(M-1)e^{2z} - (1-p)(M+1) = 0$$

i.e. when

$$e^{2z} = \frac{-(1-2p)(M-1) + \sqrt{(M-1)^2 + 16Mp(1-p)}}{2p(M+1)} \tag{18}$$

After substituting this expression in (16) and making necessary transformations, we can write the upper limit of recognition error in form (7). In particular, when p=0, the recognition error takes the simple form [11]:

$$P_i \le \left[\left(1-\frac{1}{M}\right)^{M-1}\left(1+\frac{1}{M}\right)^{M+1}\right]^{-\frac{N}{2}} \tag{19}$$

Before proceeding to further analysis, note that the above expressions (7) and (11) are true for any $M>1$. This is a principal point in which they differ from other known expressions (see, for example, papers [15-17] and their references). The exact formulae allow us to determine that known asymptotic expressions are applicable only when $N>>1$ and $M >> \sqrt{N}$. Indeed, if $N>>1$ and $M >> \sqrt{N}$, quantity S in the exponent of (11) can be expanded into powers around small parameter $1/M$. This allows much simpler expressions for error probability P and memory capacity M_{max} (given P_{max}):

$$P \le N \exp\left[-\frac{N(1-2p)^2}{2M}\right] \tag{20}$$

where

$$M_{\max} = \frac{N(1-2p)^2}{2\ln(N/P_{\max})}\,. \tag{21}$$

M_{max} should be understood in the following way: $P<P_{max}$ if $M<M_{max}$, i.e. the recognition error does not exceed the given level P_{max} if the number of stored patterns M is not greater than the upper limit M_{max} .

It is also important to consider the case when input vectors suffer regular distortions (noise):

$$\sum \theta_j = (1-\bar{p})N \tag{22}$$

where $\bar{p} = const < 1$ is a (non-random) quantity determined by the data recording/transfer system's flaws. For example, expression (22) can describe circumstances when the sign of the input signal inverts in particular channels (the number of such channels is $\bar{p}N$). In this case, expression (15) for the component error probability is replaced by the similar formula:

$$\begin{aligned} P_i(N+1,M+1,p) &= \mathbf{Pr}\left\{\sum_{k=1}^{MN}\xi_k \ge \sum_{j=1}^{N}\theta_j\right\} \le \\ &\le \mathbf{Pr}\left\{\exp\left(z\sum_{k=1}^{MN}\xi_k\right) \ge \exp\big(z(1-\bar{p})N\big)\right\} \le \\ &\le \overline{\exp\left(z\sum_{k=1}^{MN}\xi_k - z(1-\bar{p})N\right)} = e^{-z(1-\bar{p})N}\left(\overline{\exp(z\xi)}\right)^{NM} = \\ &= e^{-z(1-\bar{p})N}\left(\frac{e^{z}+e^{-z}}{2}\right)^{NM} \end{aligned} \tag{23}$$

with the minimum at

$$e^{2z} = \frac{M + (1 - \bar{p})}{M - (1 - \bar{p})} \tag{24}$$

Substituting this expression into (23), we get for the recognition error probability:

$$P_i \le \left[\left(1 - \frac{1-\bar{p}}{M}\right)^{M-1+\bar{p}} \left(1 + \frac{1+\bar{p}}{M}\right)^{M+1-\bar{p}} \right]^{-\frac{N}{2}} \tag{25}$$

The numerical analysis shows that when the number of patterns stored in HAM is small ($M < \sqrt{N}$), the recognition error grows with the noise component $\bar{p}$ much quicker than it does in the case of stochastic multiplicative noise described by (7)-(9). However, when $M >> \sqrt{N} >> 1$ (asymptotic limit), this difference is not so significant. Indeed, if we expand (25) in series on small parameter M^{-1}, the expression for the recognition error $P = \sum P_i$ (after summation over all components) takes the form

$$P \le N \exp\left[-\frac{N(1-\bar{p})^2}{2M} \right] \tag{26}$$

which differs from that for random multiplicative noise (20) in substitution of *p* for 2*p*. In other words, regular error in the observation channel is two times more effective than random multiplicative noise.

3. Adaptation process

The previous paragraph considered a *N*-dimensional Hopfield neural network determined by a particular predefined *MxN* matrix ***B*** of reference vectors (6). It was shown that the net's resolution power averaged over all such matrixes is governed by (7). The ability of the Hopfield network to associate similar events is due to a special arrangement of its interconnection array. We may suggest that in evolution live neurons acquire the ability to interact with each other through synapses. However, the determination of interconnection conductivity appears a more serious problem: for the network to behave consciously, the conductivity magnitudes and signs should match input signals. Assume again that under external influence a live network acquired the capability to change the conductivity of its interconnections (probably, very slowly). If this kind of evolution is physiologically realizable, the subsequent reasoning suggests the possible quantitative pattern of the process bringing (by means of slow evolutionary changes) the neural network to ever growing agreement with the outside world.

Indeed, let us given *M* abstract concepts represented by vectors $\boldsymbol{x}_m$ (m=1,...,M), each of concepts being defined by its own set of random real events $\boldsymbol{x}_m^{(k)} \in \overline{1, N}$ ($k=1,2,...L \to \infty$) governed by a particular statistic distribution with central point $\boldsymbol{x}_m$ and mean deviation *p*. Then adaptation of the network to the outside world's events may be regarded as step-by-step establishment of interconnection conductivity magnitudes. The first step involves the building of conductivity matrix $\hat{\boldsymbol{T}}_1$ formed according to (12) using a set of *M* vectors $\boldsymbol{x}_m^{(1)} = (e_{m1}^{(1)} x_{m1}, \cdots, e_{mN}^{(1)} x_{mN})$ where all $M \times N$ random quantities $e_{mi}^{(1)}$ are distributed in accord with (5), i.e.

$$e_{mi}^{(k)} = \begin{cases} -1, & p \\ +1, & 1-p \end{cases} \quad , \quad \forall (m,i) \in \overline{1,M} \times \overline{1,N} \,,\, k = 1,2,... \tag{27}$$

We call the network produced by vectors $\boldsymbol{x}_m^{(1)}$ the first-step adaptation network. After observing another external event, we obtain yet another set of M independently distorted vectors $\boldsymbol{x}_m^{(2)} = (e_{m1}^{(2)} x_{m1}, \cdots, e_{mN}^{(2)} x_{mN})$, where all $M{\times}N$ random quantities $e_{mi}^{(2)}$ are independent and distributed similarly to (27). It is apparent that we can unite samplings $\boldsymbol{x}_m^{(1)}$ and $\boldsymbol{x}_m^{(2)}$ and build a joint basis which gives the more reliable second-step adaptation neural network, etc.

Let us analyze the simplest case when k external events leads to the interconnection matrix taking form $\hat{\boldsymbol{T}} \sim \sum \hat{\boldsymbol{T}}_{\boldsymbol{k}}$ (where $\hat{\boldsymbol{T}}_k$ is the matrix built around the set of vectors $\boldsymbol{x}_m^{(k)}$). Simple reasoning suggests that with the growing number of observations this step-by-step adaptation results in formation of the network based on the exact knowledge of reference vectors $\boldsymbol{x}_{\boldsymbol{m}}$, $m \in \overline{1,M}$. Clear that in this succession of nets the first-step adaptation network has the poorest recognition ability. For this reason, by evaluating the recognition error of this network we can get an idea about the way the neural network has to go to become a "perfect" Hopfield network. We emphasize again that this kind of step-by-step training uses observation of real events from the "cloud" of random events $\boldsymbol{x}_m^{(k)} \in \overline{1,N}$ ($k=1,2,...L\to\infty$) governed by a particular statistic distribution with central point $\boldsymbol{x}_{\boldsymbol{m}}$, $m \in \overline{1,M}$ and mean deviation p. However, each time this kind of network has to recognize a particular vector $\boldsymbol{y}_m$ picked from this "cloud" in a completely random way. This means that the input vector does not correspond to any of the vectors used in the network's formation (rather, the probability of such correspondence is negligibly low).

Let us repeat the considerations from the previous paragraph with respect to the network built in the result of single observation of vectors $\boldsymbol{x}_m^{(1)}$ and, therefore, having the interconnection matrix $\boldsymbol{T}_1$. Let the input vector be $\boldsymbol{y}_{\boldsymbol{m}} = (x_{m1}\theta_1, \cdots, x_{mN}\theta_N) \in \{-1,+1\}^N$ where $\theta_i \in \{-1,+1\}$ is the multiplicative noise described by distribution function (27). Our aim is to decode vector $\boldsymbol{y}_m$. Following (14), we can define the corresponding error probability for the firs-step adaptation network. Then the recognition error probability for the i-th component is written in the form similar to (14):

$$\mathbf{Pr}\left\{\theta_i = -1\right\} \le \mathbf{Pr}\left\{\sum_{j\neq i}^{N} \theta_j e_{mj}^{(1)} \le \sum_{j\neq i}^{N}\sum_{n\neq m}^{M} x_{mi}x_{mj}x_{ni}x_{nj}e_{nj}^{(1)}e_{ni}^{(1)}e_{mi}^{(1)}\theta_j\right\} \tag{28}$$

In the randomized scheme where all elements of the reference matrix $\boldsymbol{B}$ are supposed independent and to take values $B_{mn} \in \{-1,+1\}$ with probability ½, set $\{\boldsymbol{B}\}$ absorbs both $\{\theta^N\}$ and $\{e^N\}$. Therefore, the probability given in the right side of (28) differs from the probability given in the right side of (14) only in that the principal random variable $\boldsymbol{e}_{nj}^{(k)}\theta_j$ has a "less sharp" distribution function as compared to (5):

$$e_{nj}\theta_j = \begin{cases} -1, & p_e \\ +1, & 1-p_e \end{cases}, \quad p_e = 2p(1-p) \tag{29}$$

Let us precise the mathematical sense of the last expression. The observations $\{x_m^{(k)}\}$ ($k=1,2,...L\rightarrow\infty$) are distributed around their common center x_m with a deviation rate p. The first step of the learning process proceeds from the real observation $x_m^{(1)} = \boldsymbol{x_m e_m^{(1)}}$. Note that observation $\boldsymbol{\tilde{x}_m^{(1)} = x_m \theta}$ presented for recognition, is randomly taken from the same ensemble independently of $x_m^{(1)} = \boldsymbol{x_m e_m^{(1)}}$. Here components of vector $\boldsymbol{\theta}$ are distributed as shown in (5). Then a little thought shows that $\boldsymbol{\tilde{x}_m^{(1)}}$ is distributed around $x_m^{(1)}$ as $\boldsymbol{\tilde{x}_m^{(1)} = x_m \eta}$ where components of the vector $\boldsymbol{\eta}$ are $\eta_i = e_{mi}^{(1)}\theta_i$, $i \in \overline{1,N}$. Thus

$$\eta_i = \begin{cases} -1, & 2p(1-p) \equiv p_e \\ +1, & (1-p)^2 + p^2 \equiv 1-p_e \end{cases}$$

as shown in (29).

Let us compare (28) and (14). As it was mentioned in the final section of paragraph 2, the randomized Hopfield network that uses the exact knowledge of randomly chosen matrix $\boldsymbol{B}$ has residual noise whose intensity is a function of parameter p meeting inequality (7). Applying the same randomization technique to matrix $\boldsymbol{B}$ with due regard to distortions $\{e\}$, we obtain the same expression for the residual noise intensity as in (7) except that we should take the intensity of effective noise p_e rather than p. Similarly, using the Chebyshev-Chernov technique, we obtain formulae (7)-(21) in which substitution $p\rightarrow p_e$ should be made. In particular, for M>>1 we have the following formula for estimating the recognition error:

$$P \le N \exp\left[-\frac{N(1-2p)^4}{2M}\right] \tag{30}$$

which is true when $M < M_{\max}$, where $M_{\max}$ is the maximal capacity of HAM at which the recognition error does not exceed the predefined limit $P_{\max}$:

$$M_{\max} = \frac{N(1-2p)^4}{2\ln(N/P_{\max})} \tag{31}$$

We see that formulae (30)-(31) differ from (20) and (21) by a stronger dependence on p. In particular, the memory capacity of a poorly learnt Hopfield network falls with p considerably faster than that of a well trained network described by (21).

Parameter $\delta p \equiv p_e - p = p(1-2p)$ tells us to what extent the "observation channel" improves in the Hopfield model with its adaptation to the outside world's conditions. Fig.1 shows that in adaptation parabola p_e=2p(1-p) changes into straight line $p_e = p$. Simple analysis shows that the amplitude of effective noise decreases ($p_e\rightarrow p$) with the growing number of observations. The recognition error also decreases, reaching the value (20) for the net trained by abstract images when $k\rightarrow\infty$. The numerical experiment (N=50÷200, M=10÷20,

p=0÷0.25) shows that just a few training steps (k~5-7) make this difference negligibly small. Dependence of the recognition error on p is given in Fig.2. In the figure the curves for k>10 almost merge with the curve for k=10.

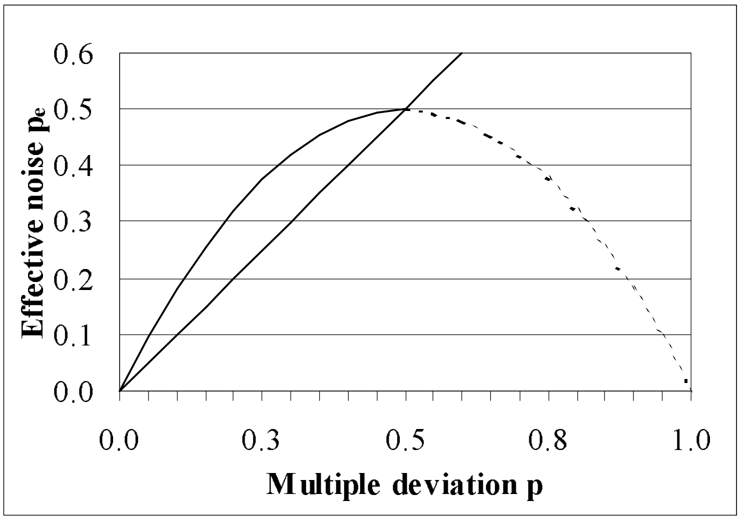

Fig.1 Effective noise p_e versus multiple deviation of traning samle

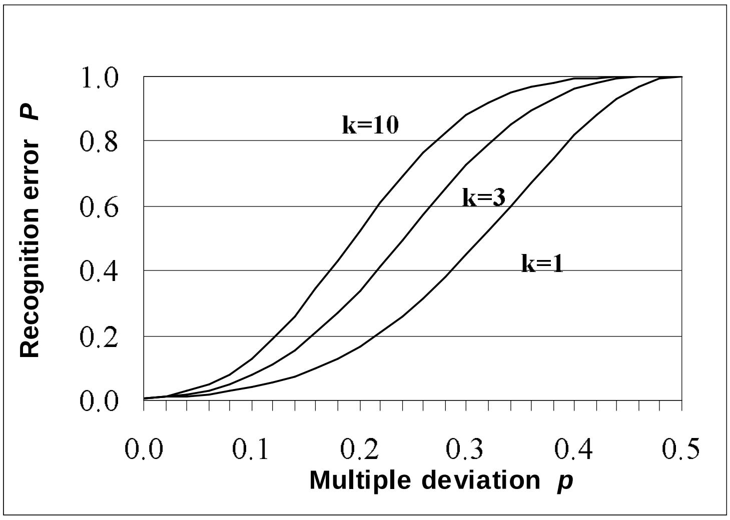

Fig.2. Recognition error probability versus multiple deviation p. The curves are drawn for the number of training steps k=1,3,10 at N=256, M=26.

4. Conclusion

In this paper, we use the standard large-deviation technique to estimate the error probability at the output of the Hopfield associative memory. The exponent of the error probability is considered as a function of the learning process which looks like a process of adaptation to the signals observed through a noisy channel. We managed to show that during the adaptation process, the error probability exponent grows from $(1-2p)^4$ up to $(1-2p)^2$ where $0 \le p \le \frac{1}{2}$ denotes the error rate at the output of the binary observation channel.

The work is partially funded by the RFBR Grants (01-07-90308, 99-01-00325) and by INTAS Grant No.00-738.

References

[1]. W.A.Little, "The ex istatce of persistent states in the brain", Math.Biosci., vol.19, pp.101 -120 (1974).

[2]. J.J.Hopfield, "Neural networks and physical systems with emergent collective computation abilities", Proc.Nat.Acad.Sci.USA, vol.79, pp.2554-2558 (1982).

[3]. C.E.Shannon, "Certain results in coding theory for noisy channels", Information and control, vol.1,pp. 6 -25 (1957).

[4]. H.J.Sussman, "On the number of memories that can be stored in a neural nets with Hebbs weights", IEEE Trans. Inf. Theory, vol.35, pp.174-178 (1989).

[5]. A.Kuh and B.W.Dickson. "Information capacity of Associative Memory", IEEE Trans. Inf. Theory, vol. 35, pp.59-68 (1989).

[6]. B.S.Kiselev, N.V.Kulakov, A.L.Mikaelian, and V.A.Shkitin, "Optical high -order associative memory using neural networks ", Radiotekhnika, Vol.10, pp.54 -62, 1990.

[7]. R.J.McEllise, E.C.Posner et al. "The capacity of Hopfield associative memory", IEEE Trans. Inf. Theory, vol.33, 461(1987).

[8]. B.S.Kiselev, N.V.Kulakov, A.L.Mikaelian, and V.A.Shkitin, "Optical implementat ion of high-order associative memory", Intern.Journ.of Optical Computing, vol.1, pp.89 -92, 1990.

[9]. D.Psaltis and Y.S.Abu-Mustafa, "Computation power of parallelism in optical computer", Dep.Elec.Eng., Calif.Inst.Tech., Pasadena, preprint, 1985.

[10]. B.S.Kiselev, N.V.Kulakov, A.L.Mikaelian, and V.A.Shkitin, "Optical auto - and hetero-associative memory for strongly correlated patterns", Radiotechnika, vol.4, No.3, pp.80 -90, 1992.

[11] B.V.Kryzhanovsky, V.N.Koshelev. "On recognition error bound for associ ative Hopfield memory", Optical Memory&Neural Networks, vol. 9, №4 , pp.187-193 (2000).

[12]. V.N.Koshelev, B.V.Kryzhanovsky, A.Fonarev. "Estimating efficiency of randomized Hopfield memory". Problemi Peredachi Informatsii (to be published), 2001.

[13]. V.N.Koshelev, B.V.Kryzhanovsky. "On Recognition Capability of Hopfield Networks". Pattern Recognition and Image Analysis, Vol.11, №1, pp.47-49 (2001). Proc. of 5th Int. Conf. on Pattern Recognition and Information Analysis PRIA-5-2000, vol.1, pp.81-84, Samara, Russia, 2000.

[14]. H. Chernov, "A measure of asymptotic efficiency for tests of hypothesis based on the sum of observations", Ann. Math. Statistics, vol.23, pp.493-507 (1952).

[15]. D.Burshtein. "Long -Term Attraction in Higher Order Neural Networks", IEEE Trans. of Neur. Netw., vol. 9, N4, pp.42-50 (1998).

[16]. L.Wang, "Effects of noise in training patterns on the memory capacity of the fully connected binary Hopfield neural network", IEEE Trans. Neur. Netw., vol. 9, N4, pp.697 -704 (1998).

[17]. N.Farhat. "Corticonics: the way to designing machines with brain -like intelligence", SPIE'2000 , pp. 158-170, San-Diego, 2000.