A memory which tentatively forgets

1993 J. Phys. A: Math. Gen. 26 2099

(http://iopscience.iop.org/0305-4470/26/9/009)

J. Phys. A: Math. Gen. 26 (1993) 2099–2112. Printed in the UK

A memory which tentatively forgets

Michał Żochowski†, Maciej Lewenstein† and Andrzej Nowak‡

† Centrum Fizyki Teoretycznej, Polska Akademia Nauk, Al. Lotników 32/46, 02-668 Warsaw, Poland

‡ Institute for Social Studies, Warsaw University, Stawki 5/7, 00-183 Warsaw, Poland

Received 14 July 1992

Abstract. We present a model of neural network that consists of several subnetworks or categories. Each of the subnetworks may serve to store a family of correlated or uncorrelated data. Recognition of a pattern from a given family (subnetwork) consists in identification of that family first, and then in recognition within the corresponding subnetwork. Family identification is based on appropriate novelty checks. Results of those checks are used for self-control of the recognition process. The model works as a memory which tentatively forgets and therefore allows for a significant increase in its storage capacity.

1. Introduction

One of the most fundamental properties of neural network memories [1] is their limited storage capacity. Hopfield [2,3] in his pioneering paper has shown that a fully connected network of N neurons with Hebbian learning rule can store $p = 0.14N$ unbiased random patterns, provided small errors are allowed. Gardner [4,5] demonstrated that with optimal choice of synaptic connections the maximal storage capacity for such patterns is $\alpha = p/N = 2$. There have been several attempts to increase the storage capacity of Hopfield-type networks. Derrida *et al* [6] have shown that strongly diluted networks can store $\alpha = 2/\pi$ random unbiased patterns per connection.

Willshaw *et al* [7] observed that it is possible to store in a perceptron memory [8] $p \propto N^2/\ln^2 N$ sparsely coded patterns (i.e. patterns consisting mostly of equal sign entries). Tsodyks and Feigelman [9] generalized this result to Hopfield-like nets. However, there is a price for increased capacity, because sparsely coded patterns contain much less information than random ones. Nevertheless, information capacity per neuron remains non-zero in the case of sparsely coded data. This result also holds for correlated data in the limit of increasing correlation length [10].

The main aim of the present paper is to construct a model that allows infinite storage capacity α without the requirement of sparse coding or strong correlations. Our model can store y families of p patterns. Different families may be uncorrelated, whereas patterns within a family are biased on a prototype pattern of the family. As we discuss in section 5, in the limit of strong bias the storage capacity indeed tends to infinity. Even in the case of families of uncorrelated patterns our model allows for an increase in the capacity for the Hebbian rule from 0.14 to $2/\pi$.

A similar problem was considered by Franz *et al* [11]. In their paper, they constructed synaptic metrics that can store uncorrelated classes of patterns. The drawback of the method is that their prescription makes the synaptic metrics non-local. Our approach is different. We use standard or slightly modified learning algorithms and achieve a similar goal in

0305-4470/93/092099+14$07.50

a dynamical instead of a static way. During the dynamics, our model recognizes the subnetwork (category) first, and then the pattern within it.

In a sense our model is a generalization of Parisi's 'memory which forgets'. In Parisi's model [12] the newest patterns are learned with biggest weight. In this way older patterns are consecutively forgotten in the course of learning. In our model, the memory tentatively forgets information about all families of patterns except the one to which the recognized pattern belongs. This aim is achieved using a neural network with self-control mechanisms, recently proposed by us and called 'nervous' neural networks [13, 14].

'Nervous' neural networks are able to recognize novelty or a category of the pattern presented to them in the very early stages of the recognition process. In order to construct a network, which would self-consistently check novelty or a category of input patterns and which would incorporate a control mechanism based on this novelty check, one has to solve two crucial problems. First of all, one has to identify the quantity that may be used to detect the novelty of the input patterns; one then has to determine how this quantity can control the dynamics. Fortunately, an indication of how these problems may be solved follows from an observation made by Hopfield [2].

In the Monte Carlo (MC) simulation of Ising spin dynamics one randomly chooses spins and then checks whether or not they should undergo flips. The relative number of spins or neurons, which flip in the very initial stage of the evolution, may be identified with the desired global parameter that provides a check of the novelty. The same parameter characterizes very precisely how well the current pattern is known to the system. The relative flip frequency of the tossed neurons also provides a powerful means of measuring the distance of the current state to some of the equilibrium states, due to the flatness of an energy landscape in the vicinity of a local minimum (see the discussion of relevant simulations in [13, 14]). In the present paper we generalize this method of pre-recognition of novelty to the case of pre-recognition of several distinct categories. The results of the category check are then used to modify and control standard MC dynamics.

The plan of the paper is as follows. In section 2 we decribe our model in detail. In section 3 we present the results of numerical simulations. In section 4 we formulate the analytic theory of the model for the case of unbiased random patterns. To this end we generalize the theory of Derrida *et al* [6]. In section 5 we formulate the theory for the case of uncorrelated families of strongly biased patterns. As a result we obtain infinite storage capacity as in Willshaw's paper *without losing the diversity of stored patterns*. Section 6 contains a conclusion and a discussion of the possible applications of our model to cognitive science and artificial intelligence.

2. Description of the model

We shall now present a mathematical description of 'nervous' neural networks, leaving a detailed discussion of their possible applications to section 6.

The models in question are supposed to store y categories of patterns. The patterns within each category may but do not have to be correlated. By different categories we simply mean here groups of patterns which correspond to an assigned quality. In psychological applications (see section 6), these may be, for instance, groups of patterns related to different qualities of emotions, such as fear, joy etc.

In numerical simulations and in the first part of an analytic treatment we have stored in the memory random uncorrelated and unbiased patterns $\xi^{\mu\nu}$, $\nu = 1, \ldots, y$, $\mu = 1, \ldots, p$, where p was the number of patterns in the category. Later, we have also considered y

uncorrelated families of patterns strongly biased on family prototypes. The model was constructed so that it was able to:

(i) recognize the category to which an input pattern belongs, using an appropriate novelty check; and

(ii) recognize the input pattern within its category only.

Our aim here was to model categories of patterns with different qualities in such a way that initial recognition of input data as 'known' for a given category would facilitate final recognition within the same category.

In order to construct y appropriate novelty checks we consider a fully connected network that consists of y diluted Hopfield networks [1]. We have divided all existing bonds (i, j) into y disjoint sets of the same size S_ν, $\nu = 1, \ldots, y$. The sum of these sets is the set of all bonds. The number of sets y might remain finite in the thermodynamical limit—that is the case of moderate dilution [1, 15, 16]. Alternatively, we may study the case of strong dilution [6, 17], for which the number of bonds in each of the sets S_ν, K, behaves as $K/N \to 0$ for $N \to \infty$.

In the case of random unbiased patterns the couplings in these models take the standard Hebbian form

$$J_{ij} = \frac{1}{K} w_\nu \sum_{\mu=1}^{p} \xi_i^{\mu\nu} \xi_j^{\mu\nu} \tag{1}$$

for (i, j) belonging to the set S_ν. In the case of uncorrelated families of biased patterns we use a modified Hebb rule (see section 5).

In the above expression we have introduced weights of the category w_ν. Note that w_ν may be absorbed in the definition of temperature and in such a case $\beta_\nu = \beta w_\nu$ would correspond to the temperature (noise level) of a given category of patterns. We expect that in stationary states one and only one of the weights w_ν remains relevant. Although the bonds are diluted in this limit, the system recognizes nearly perfectly stored patterns $\{\xi^{\mu\nu}\}$.

We may now decompose the local field acting on neuron i, $h_i = \sum_{j \neq i} J_{ij}\sigma_j$ into the contributions coming from different families of patterns,

$$h_i = \sum_{i \neq j} J_{ij}\sigma_j = \sum_\nu h_{\nu i} \tag{2}$$

where

$$h_{\nu i} = \frac{w_\nu}{N} \sum_{j \neq i}^{(i,j) \in S_\nu} \sum_{\mu=1}^{p} \xi_i^{\mu\nu} \xi_j^{\mu\nu} \sigma_j. \tag{3}$$

The MC updating is performed as usual by demanding that a spin σ_i would flip with probability 1, if it is aligned antiparallel to its local field h_i, i.e. $\sigma_i h_i < 0$, or with probability $p = 1 - \exp(-2\beta\sigma_i h_i)$, if $\sigma_i h_i \geqslant 0$.

The novelty checks can now be constructed for each category by monitoring the frequency of spin flips, which *would take place if only a given category was present.* We introduce y parameters s_ν at each MC step, which are defined as follows: $s_\nu = 1$ if $\sigma_i h_{\nu i} < 0$ or if $\sigma_i h_{\nu i} \geqslant 0$ with conditional probability $p = 1 - \exp(-2\beta h_{\nu i}\sigma_i)$; otherwise $s_\nu = 0$. Instead of variation in the noise level which was used in [13, 14], we introduce here random variations of the weights w_ν (or, in another words, variations of the noise levels of the categories). For the $(n+1)$th MC step we use the formula

$$w_\nu^{n+1} = w_\nu^n - \Delta_- s_\nu^n + \Delta_+(1 - s_\nu^n) \tag{4}$$

with $\Delta_- = 0.04$, $\Delta_+ = 0.03$ Additionally we let $w_\nu^{n+1} = 0.1$, if w_ν^n becomes less than 0.8. If w_ν^m becomes negative, we set it and keep it equal to zero.

In simulations one may start with $w_\nu = 1$ for all ν at relatively low temperatures, so that all retrieval states are stable. The response of the system is then typically of two kinds:

(i) If the input pattern is 'unknown' to the system (i.e. the input pattern is only weakly correlated to any of the stored patterns $\xi^{\mu\nu}$), the initial frequency of potential flips s_ν is large enough, so that each of the w_νs will drift towards zero.

(ii) If the input pattern is macroscopically correlated to any of the patterns $\xi^{\mu\nu}$, the corresponding w_ν increase. All other $w_{\nu'}$ for $\nu' \neq \nu$, again drift towards $w_{\nu'} = 0$; the ancestor state will, in this case, be recognized perfectly, but recognition will take place within the category ν.

Clearly these models may be generalized to the case of correlated patterns which we discuss in section 5. Dynamical correlation among different categories may be included in expression (1.2) for the variations of w_ν. For example, we may model a situation in which 'known' patterns belonging to one category facilitate recognition of patterns belonging to some other category or a group of categories.

3. Numerical results

It seems possible that the models of the class discussed above may be used to increase the storage capacity of the Hopfield model. As already observed by Hopfield and studied in our simulations, novelty checks based on measurement of the spin flip frequency do work for overloaded Hopfield models (i.e. for the case when the number of patterns p exceeds the critical value $\alpha_c N$). Then, the model may first recognize the category to which the input pattern belongs, and then perform recognition within a subnetwork of the corresponding category. The storage capacity of the model would then be a sum of the capacities of subnetworks which, in principle, may exceed the standard result for fully connected networks. Such an increase in the storage capacity is, in a sense, analogous to that discussed by Parisi [12], in the paper on memory, which can learn arbitrarily many new patterns, forgetting the old ones in order to avoid overloading. Our model presents a 'memory which tentatively forgets', due to the introduced self-control mechanism, and which remains underloaded in the course of recognition.

In this section we will discuss our numerical results. The system is built from a regular fully connected Hopfield model where memories are stored using the Hebbian rule. First, we divide our network into y subnetworks by randomly selecting synaptic connections to different subnets. To every subnet there is an assigned weight w_ν, which at the beginning of the simulation is the same for every subnet and set to one. During the simulation the w_ν are updated in every MC step, using formula (4) shown in the previous section. After dividing the network into subnetworks we teach it configurations which it should recall. This is done, as we said earlier, by using the Hebbian rule separately for every subnet. In the simulation all subnetworks have the same number p of uncorrelated patterns. The starting configuration is built by randomly distorting one of the stored memories.

The dynamics of the system is performed by modified MC updating. The difference is that in every step we calculate the local fields $h_{\nu i}$ from each subnet separately, then check whether its sign agrees with the total field of system and whether the tossed spin would change its sign if it experienced the local field $h_{\nu i}$ only. We then change the weights accordingly. The total field is calculated by summing all local fields. The actual dynamical behaviour of tossed spins results from the standard MC rule applied for the total field.

During simulations we checked whether our system converges toward the initially distorted state. If yes, we considered that the pattern was recalled. The procedure was repeated for every stored pattern. We encountered three classes of behaviour of the system.

(i) *Identification* (correct recognition). The system converges towards an ancestor pattern, first choosing the subnet to which it belongs.

(ii) *Classification* (recognition of category). The system fluctuates around the pattern, but cannot acquire the required precision of recognition. The category of pattern was classified correctly.

(iii) *Failure* The system randomly walks over configurations, rapidly escaping from the ancestor pattern.

The subnet in which the pattern was stored, i.e. the category to which it belonged, was usually (except for the case of *failure*) recognized much earlier than the pattern itself. The weight of that subnetwork steadily increases while all other weights rapidly converged towards zero and stayed there for the rest of the simulation.

The simulations were performed for $N = 200$ spins. The results averaged over few realizations are shown in the figures at the end of this section. The figures show the ratios of pattern identified, classified and failed to identify to the number of patterns stored in the whole network as a function of imposed capacity on the network. As one can see the network shows a tendency to increase its capacity. Note that our numerical results still contain some fluctuations due to the finite number of neurons. In addition, these results correspond to finite y, i.e. moderate dilution within each subnetwork. For these reasons the increase in capacity is not as large as expected in the case of strong dilution. We discuss this point in detail in the next section.

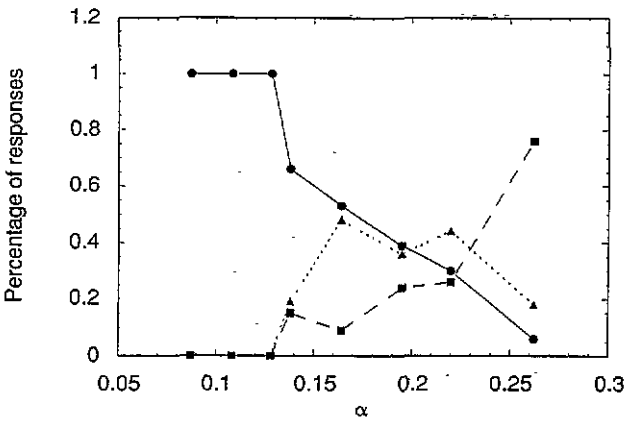

Figure 1. The percentage of various types of responses as a function of the total capacity for a network consisting of one subnet: full line with circles, percentage of *identification*; dotted line with triangles, percentage of *classification*; and broken line with squares, percentage of *failure*.

4. Network with strongly diluted subnetworks

When the number of networks y is finite, the number of connections per neuron in each of the networks $K = N/y$ is comparable to N. This is the case of moderate dilution which

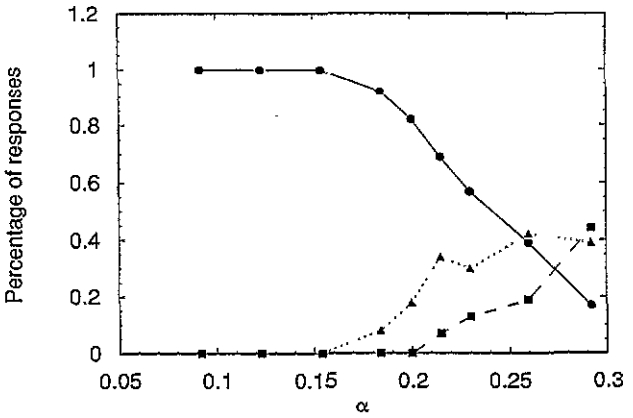

Figure 2. The same as figure 1 but for a network devided into three subnetworks.

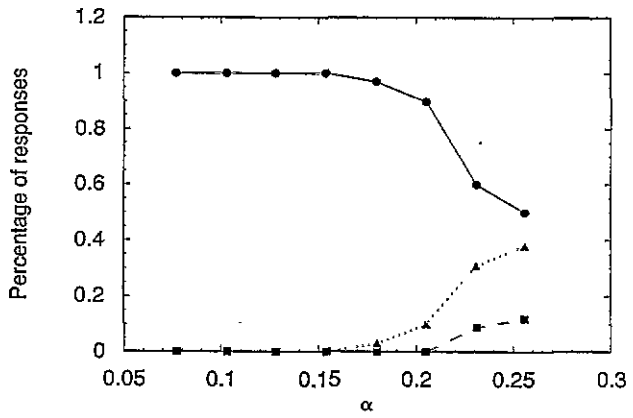

Figure 3. The same as figure 1 but for a network devided into five subnetworks.

is very difficult to analyse. We can, however, get some analytic insight into the discussed model in the strongly diluted case, when $y \to \infty$ in such a way that $K/N \to 0$. In this situation we expect that the method developed by Derrida *et al* [6] will turn out to be useful.

In the present section we will consider the case $y \to \infty$ and perform self-consistent signal-to-noise analysis. Let us then consider a network of N neurons, undergoing dynamics

$$\sigma_i(t+1) = \text{sign}(h_i(t)). \tag{5}$$

The local field contains Hebbian contributions from all subnetworks,

$$h_i = \frac{1}{K} \sum_j \left(\sum_\nu^y w_\nu \sum_\mu^p \xi_i^{\mu\nu} \xi_j^{\mu\nu} \right) \sigma_j \tag{6}$$

where $\xi_i^{\mu\nu}$ are independently distributed unbiased random variables.

To complete the description of the model we have to specify dynamics of weights w_ν. In the case of asynchronous dynamics one could adopt the definition used in numerical

simulations (4). Here, however, we shall consider synchronous dynamics, which usually allows for a much simpler description in the thermodynamical limit. The dynamics of w_ν should, in this case, be determined by appropriately defined mean frequencies of neuron flips. We follow here the approach used in [18].

First, we define quantities that measure mean frequencies of spin flips, which would take place if only one subnetwork was present. For each ν we define

$$\alpha_\nu(t) = \frac{1}{4N}\sum_{i=1}^{N}(1 - \mathrm{sign}(\sigma_i(t)h_{\nu i}))^2. \tag{7}$$

Note that if the system configuration is completely unrelated to memories stored in the νth subnetwork, $\alpha_\nu \simeq 1/2$. Conversly, if the system configuration is close to one of the states stored in the νth subnetwork, say $\xi^{\mu\nu}$, then α_ν tends to zero.

It is thus useful to introduce exponential dynamics for w_ν

$$w_\nu(t+1) = \exp(\gamma(1 - 4\alpha_\nu(t)))w_\nu(t). \tag{8}$$

Equation (8) ensures that w_ν will exponentially decay to infinity if the pattern is known (i.e. α_ν is small) or to zero if the pattern is unknown (i.e. α_ν greater than $\frac{1}{4}$).

We will now assume that, in the thermodynamical limit and for increasing time, only one subnetwork survives and all other w_ν tend rapidly to zero. We will treat the contribution to the local field from these subnets as vanishing noise. On the other hand, the dynamics within one remaining subnet corresponds to the dynamics of a strongly diluted network. Here, we cannot neglect the noise coming from other patterns of the category. In the limit of large number of connections per neuron K we may, however, treat this noise as Gaussian, according to the theory of Derrida *et al* [6].

Let us assume that initial state corresponds to randomly distorted pattern ξ_i^{11}. For sufficiently long times, all w_ν for $\nu \neq 1$ tend rapidly to zero and we may treat the system as a single strongly diluted network. According to the theory of Derrida *et al*, probability distribution of the neuron configuration σ_i, averaged over initial state distribution and ancestors of the ith neuron in the subnetwork is a function of ξ_i^{11} only and is given by

$$p(\sigma_i) = \frac{1 + \sigma_i m(\xi_i^{11})}{2}. \tag{9}$$

The mean signal (i.e. averaged total local field $h_i(t)$) is then given by

$$\langle h_i \rangle = w_1 \xi_i^{11}(m(+) - m(-)). \tag{10}$$

On the other hand, the total noise has two contributions: one coming from subnet 1 and the other coming from other subnets. The square of the noise is defined as $\langle h_i^2 \rangle - \langle h_i \rangle^2$ and is equal to

$$R^2 = \frac{1}{K}\left(w_1^2(p-1) + \sum_{\nu=2}^{y} w_\nu^2 p\right). \tag{11}$$

Our theory will be correct provided the second term in equation (11) is negligible in comparison to the first one. This will happen provided

$$\sum_{\nu=2}^{y} w_\nu^2/w_1^2 \simeq (y-1)\langle w_\nu^2/w_1^2 \rangle \ll 1 \tag{12}$$

where $\langle w_\nu^2/w_1^2\rangle$ denotes a typical value of w_μ^2/w_1^2. According to equation (8) we obtain

$$\langle w_\nu^2/w_1^2\rangle \simeq \exp(-4\gamma t). \tag{13}$$

From equation (12) follows the lower bound for time

$$t \gg (1/4\gamma)\ln y. \tag{14}$$

On the other hand, the Derrida theory is valid provided the time is not too long so that there are no closed loops in the ancestor tree of the neurons within a diluted network. The condition is

$$K^t \ll N. \tag{15}$$

Let us assume that $y = N^x$, with $0 \leqslant x \leqslant 1$. From the above two equations it follows that $K = N/y = N^{1-x}$ and

$$\frac{x}{4\gamma}\ln N \ll t \ll \frac{1}{1-x}. \tag{16}$$

This condition is hard to fulfil in the limit of $N \to \infty$, but may be fulfilled for any large but finite N provided x is sufficiently close to 1.

In the critical case $x = 1$, and we may take $y = N/a$ where a is some positive constant. Then $K = a$ and we obtain

$$\frac{\ln N}{\ln K} \gg t \gg \frac{1}{4\gamma}(\ln N - \ln K). \tag{17}$$

Equation (17) is easy to fulfil provided

$$4\gamma > \ln K. \tag{18}$$

In conclusion, we have shown that there is a well defined region of time for which the dynamics of our model reduces to that of the strongly diluted models of Derrida *et al.* Denoting $m(t) = m(+) - m(-)$ the dynamics takes the form

$$m(t+1) = \langle \mathrm{sign}(m(t) + R(t)z)\rangle_z \tag{19}$$

where $\langle\ldots\rangle_z$ denotes averaging over a Gaussian noise with mean zero and variance 1. The noise strength is

$$R^2(t) = \frac{p}{K}(1 + y\exp(-4\gamma t)). \tag{20}$$

Of course, our asymptotic analysis indicates that retrieval states for which all w_ν except one are equal to zero and for which the configuration of the network has non-vanishing overlap with one and only one stored pattern are locally stable. We cannot say anything about the domain of attraction of retrieval states and about our approach to them in the initial phase of the dynamics. In this respect we rely on numerical simulations from the previous section. They clearly indicate that typically one subnet survives in the course of the dynamics, provided the evolution of w_ν is fast enough. In the notation of the analytic theory of this section this means that γ should be large enough.

What is the maximal storage capacity of our network? Since the maximal storage capacity of strongly diluted network is $p = 2K/\pi$ [6], so is the capacity of each subnetwork. The total capacity is thus $P = py = 2N/\pi$. Since we used a standard Hebbian rule, this result should be compared to the Hopfield result $P \simeq 0.14N$. As we see, using our dynamical pre-recognition of categories we were able to increase the capacity by a factor greater than four!

5. Storage of uncorrelated families of strongly biased patterns

There have been many attempts to increase the storage capacity of neural network memories. One that seems to be very successful is that of Willshaw *et al* [7], who showed that a perceptron may store $p \simeq N^2/\ln^2 N$ strongly biased patterns. Tsodyks and Feigelman generalized this method to attractor networks. The method allows the informational capacity to be kept finite with a growing number of stored patterns. Its weaker points are that the stored configurations carry less information and their variety sharply decreases in the sparse coding limit. As will be shown below, our model maintains a similar storage capacity allowing the storage of a practically unlimited variety of patterns.

Let us consider, as in the previous section, a network whose synaptic connections are randomly divided into y subnetworks (i.e. categories). The number of those subnetworks at that point is arbitrary. Every category has a stored prototype, which is uncorrelated with other prototypes of different categories. Prototypes may be stored in the memory using, for example, the standard Hebb rule for each category. In the next step we construct a family of patterns which are biased on the corresponding prototype for every subnetwork. The probability distribution of these configurations is

$$P(\xi^{\mu\nu}, \xi^{\nu}) = \left(\frac{(1+a\xi^{\nu})}{2} \frac{(1+\xi^{\mu\nu})}{2} + \frac{(1-a\xi^{\nu})}{2} \frac{(1-\xi^{\mu\nu})}{2} \right) \tag{21}$$

where ξ_i^{ν} is the value of the prototype of subnetwork ν at site i and a is a constant taking values between values 0 and 1, describing the bias of patterns on the prototypes. It is kept the same for all patterns in all subnetworks for simplicity.

In this way the model has stored families of biased pattern which are uncorrelated with one another. To construct a matrix of connections J_{ij} we use, similar to Tsodyks and Feigelman [9] and Amit [1], a generalized and modified Hebb rule. The difference is that, in addition, every neuron corresponds to a different threshold value. Then, the local field at site i is given by

$$h_i = \sum_{\nu} w_{\nu} \left(\sum_{j \in S_{\nu}} \frac{1}{K} \left(\sum_{\mu} (\xi_i^{\mu\nu} - a\xi_i^{\nu})(\xi_j^{\mu\nu} - a\xi_j^{\nu})(\sigma_j - a\xi_i^{\nu}) + (1-a)\xi_i^{\nu}\xi_j^{\nu}\sigma_j\theta \right) \right) \tag{22}$$

where $\xi_i^{\mu\nu}$ is the value of the μth pattern biased on the νth prototype (and stored in the νth subnetwork) at site i, θ is the constant that describes the threshold and takes a value between 0 and 4, w_{ν} is the weight for the νth subnetwork, defined in the same way as in the previous section, K is a normalizing constant, equal to the number of connections per neuron within a subnetwork.

The sum over j runs for a given ν over those connections that belong to the νth subnet. As one can see, by taking $a = 0$ the equation for the local field, except for the threshold value, looks the same as equation (6) in the previous section, where we used Hebb's rule to store patterns. If, on the other hand, one takes the limit of $a \to 1$, one gets $\xi_i^{\mu\nu} \to \xi_i^{\nu}$ which means that the stored patterns tend to be the same as the prototypes in corresponding subnetworks. We will consider this limit in later calculations because in this case (i.e. strongly biased patterns) the noise of the network will tend to zero in comparison with the signal.

The dynamics of the system will be analogous to that from the previous section and can be described by the equation

$$\sigma_i(t+1) = \operatorname{sign}(h_i(t)). \tag{23}$$

The local field is described in equation (22). The weights of the subnetworks and their dynamics are introduced in the same way as in the previous section, so they will not be repeated here. Again, the behaviour of the system is that all weights except for one, which is assigned to the subnetwork in which the stored pattern is recognized, tend to zero (i.e. all subnetworks (categories) are 'turned off' except for the one where the correct pattern is stored). Then the pattern within that subnetwork is recognized.

Let us now consider the problem more carefully. As was said earlier in the case of strongly biased patterns ($a \to 1$) the noise will tend to zero in comparison with the signal. In this case we can expect our system to converge to one of the stored patterns. To focus attention let us consider that it is the first pattern from the first subnet. Thus we can assume that the state of the ith neuron is given by

$$\langle \sigma_j \rangle = \xi^{11} + \delta\sigma = \xi^{11} + D(1-a)^x \tag{24}$$

where D is an independent constant. The above expression means that $\langle \sigma_j \rangle$ differs from ξ_j^{11} by a small correction $\delta\sigma$ of the order of $D(1-a)^x$. We combine the above formula with equation (22) to calculate averaged local field and noise. After simple calculations we get formulae for the local field:

$$\langle h_i \rangle = w_1(\xi^{11} - a\xi_i^1)(1-a^2) + a(1-a)\theta\xi_i^1 \tag{25}$$

and for the noise

$$\begin{aligned} R^2 &= \langle h_i^2 \rangle - \langle h_i \rangle^2 \\ &= \frac{w_1^2(1-a)^2(1-a^2)\theta^2}{K} + \frac{1}{K}\sum_{\nu=2}^{y} w_\nu^2(1-a)^2\theta^2 + \frac{1}{K}w_1^2(\xi_i - a\xi_i)^2 4a^2(1-a^2) \\ &\quad + \frac{p}{K}w_1^2(1-a^2)^3 + \frac{p}{K}\sum_{\nu=2}^{y} w_\nu^2(1+a^2)(1-a^2)(1-a)^2 \\ &\quad - \frac{4}{K}w_1^2a^2(1-a)^2\xi_i^1(\xi_i^{11} - a\xi^1)\theta \end{aligned} \tag{26}$$

where p is the capacity of the subnetwork.

We anticipate that in the limit $a \to 1$ the capacity of subnetworks p tends to infinity and we assume that

$$K = \frac{A|\ln(1-a)|}{(1-a)} \tag{27}$$

$$p = \frac{B}{(a-1)^2}. \tag{28}$$

Inserting these two formulae into equation (26), finally, taking the limit $a \to 1$ we get the following equation for the signal:

$$S = (1-a)(2(\xi_i^{11} - a\xi_i^1) + \theta\xi_i^1) \tag{29}$$

and for the noise, considering only the relevant parts,

$$R^2 = \frac{32w_1^2(\xi_i^{11} - a\xi_i^1)^2(1-a)^2}{A|\ln(1-a)|} + \frac{8Bw_1^2(1-a)^2}{A|\ln(1-a)|} + \frac{8B}{A}\sum_{\nu=2}^{y}\frac{w_\nu^2(1-a)^2}{(1-a)|\ln(1-a)|}. \tag{30}$$

Now we can rewrite our dynamics from equation (23) to the form

$$\langle \sigma_i(t+1) \rangle = \langle \operatorname{sign}(S + Rz) \rangle_z. \tag{31}$$

As in [18], we want to derive equation (24) from this equation. After easy calculations we see that when $\xi_i^\nu = 1$ or $\xi_i^\nu = -1$ we obtain a correction term of order $\delta\sigma \simeq (1-a)^x$ where x is given by

$$x = \frac{w_1^2(2(\xi_i^{11} - a\xi_i^1) + \theta\xi_i^1)^2}{2((w_1^2/A)(\xi_i^{11} - a\xi_i^1)^2 + (8Bw_1^2/A) + (8B/A)\sum_{\nu=2}^{y}(w_\nu^2/(1-a)))}. \tag{32}$$

From the condition that the exponent has to be greater than or equal to 1, one concludes that the third term in the denominator, which describes the noise coming from subnets whose weights tend to zero, must not be greater than 1:

$$\frac{8B}{A}\sum_{\nu=2}^{y}\frac{w_\nu^2}{(1-a)} \leqslant 1. \tag{33}$$

Now, assuming that the dynamics of the weights w_ν, $\nu \geqslant 1$, are defined as in an earlier section and given by

$$\langle w_\nu^2/w_1^2 \rangle \simeq \exp(-4\gamma t) \tag{34}$$

one can derive the lower bound for time which can be written as

$$4\gamma t > \ln(y/(1-a)). \tag{35}$$

As one can see from the above equation if the time is long enough the noise resulting from the existence of many subnetworks will tend to zero and one is then left with a model with one diluted network with strongly biased patterns.

Let us now consider the capacity of our model. In the case of strongly biased patterns we do not have any limitations on the number of subnetworks which was required as the applicability condition of Derrida's theory in a previous section. First let us consider the number of subnets y to be finite (i.e. we have only a few subnetworks). The lower bound for time is given by

$$t \geqslant (1/\gamma)|\ln(1-a)| \tag{36}$$

and the capacity of the network is given by

$$\frac{P}{N} = \frac{B}{A(1-a)|\ln(1-a)|} \simeq \frac{N}{\ln^2 N} \tag{37}$$

so that we recover the Willshaw result. In the case where y tends to infinity we can assume that it is given by

$$y = C/(1-a)^n. \tag{38}$$

Then our lower bound on time takes the form

$$t \geqslant ((n+1)/\gamma)|\ln(1-a)| \tag{39}$$

and the capacity is given by

$$\frac{P}{N} \sim \frac{(n+1)N^{1/(n+1)}}{(\ln N)^{(n+2)/(n+1)}}. \tag{40}$$

It is interesting to see that it is largest when $n = 0$, i.e. y is constant Then we get the same result as Tsodyks and Feigelman [9].

6. Conclusions

In this paper we have presented a model with a self-control mechanism. The model, on one hand, leads to increased storage capacity for a network storing uncorrelated patterns. On the other, for strongly biased patterns, we have been able to obtain the same storage capacity as the Willshaw model, without having to decrease the variety of stored patterns. The mechanism is based on novelty check, consisting of the measurement of appropriately defined spin-flip frequencies. The novelty check allows us to distinguish the subnetwork (i.e. category) to which the input pattern belongs in the initial stage of the recognition process, thus enabling us to turn off networks with irrelevant information.

In this paper we have combined numerical and analytical results for the case of uncorrelated patterns. The numerical results agree fairly well with the analytical ones.

The model described here can be used to construct more efficient networks and thus will have practical applications in artificial intelligence. It can also explain results from cognitive psychology showing that classification on the category level is faster than identification of an exemplar [19]. Effects of this sort can be demonstrated even when such rapid recognition cannot be made by using a feature distinctive for the category [20]. In our model categories correspond to subnetworks, whereas exemplars are represented by states of the network. The model would thus predict that a faster response on the category level would only be possible for well established stable categories. Empirical results in cognitive psychology show that this is exactly the case [21]. Our model also explains why it is beneficial for the cognitive system first to classify and then to identify the exemplar. Such automatic classification occurs even when it interferes with the task [22].

The model can also be used to explain the influence of emotional state on the recognition process [23]. Each emotion would constitute one category, and would thus correspond to one subnetwork in our model. Memories could then be classified, according to this perspective, into happy, angry, sad, etc. Being in an emotional state is modelled by increasing the weights of a corresponding subnetwork. Emotions would then be modelled as a control factor of a cognitive processing which is congruent with the current understanding of the role of emotions in psychology [24]. The rapid recognition of an emotional content by the novelty detection mechanism could also lead to activation of one of the subnetworks in a self-control mechanism. It has been shown that the emotional content of a stimulus is classified before the cognitive content is identified [25]. Our model shows how such dual emotional and cognitive coding may lead to increased memory capacity.

Acknowledgment

This paper has been financed by Polish Government Grant KBN 1-1113-91-02.

References

[1] Amit D J 1989 *Modeling Brain Function: The World of Attractor Neural Networks* (Cambridge: Cambridge University Press)

[2] Hopfield J J 1982 Neural networks and physical systems with emergent collective computational capabilities *Proc. Natl Acad. Sci. USA* **79** 2554–8

[3] Amit D J, Gutfreund H and Sompolinsky H 1985 Storing infinite number of patterns in a spin glass model of neural networks *Phys. Rev. Lett.* **55** 1530–3

[4] Gardner E 1988 The space of interactions in neural network models *J. Phys. A: Math. Gen.* **21** 257–70

[5] Cover T M 1965 Geometrical and statistical properties of systems of linear inequalities with application in pattern recognition *IEEE Trans. Electron. Comput.* **EC-14** 326–34
[6] Derrida B, Gardner E and Zippelius A 1987 An exactly soluble asymmetric neural network model *Europhys. Lett.* **4** 167
[7] Willshaw D J, Buneman O P and Longuet-Higgins H C 1969 Non-holographic associative memory *Nature* **222** 960
[8] Minsky M and Papert S 1969 *Perceptrons* (Cambridge: MIT) new edition 1989
[9] Tsodyks M V and Feigelman M V 1988 The enhanced storage capacity in neural networks with low activity level *Europhys. Lett.* **6** 101
[10] Lewenstein M and Tarkowski W 1992 Optimal storage of correlated patterns in neural network memories *Phys. Rev.* A in print
[11] Franz S, Amit D J and Virasoro M A 1990 Prosopagnosia in high capacity neural networks storing uncorrelated classses *J. Physique* **52** 387
[12] Parisi G 1986 A memory which forgets *J. Phys. A: Math. Gen.* **19** L617
[13] Lewenstein M and Nowak A 1989 Fully connected neural networks with self-control of noise levels *Phys. Rev. Lett.* **62** 225–8
[14] Lewenstein M and Nowak A 1989 Recognition with self-control in neural networks *Phys. Rev.* A **40** 4652–64
[15] Sompolinsky H 1987 The theory of neural networks: the Hebb rule and beyond *Heidelberg Colloquium on Glassy Dynamics* ed J L van Hemmen and I Morgenstern (Berlin: Springer)
[16] Sompolinsky H 1986 Neural networks with non-linear synapses and static noise *Phys. Rev.* A **34** 2571
[17] Kree R and Zippelius A 1987 Continuous-time dynamics of asymmetrically diluted neural network *Phys. Rev.* A **36** 4412
[18] Lewenstein M and Olko M 1992 Storage capacity of 'quantum' neural networks *Phys. Rev.* A in print
[19] Schneider W and Shiffrinm M R 1977 Controlled and automatic human information processing: I. detection, search and attention *Psychol. Rev.* **84** 1
[20] Jonides J and Gleitman H 1972 A conceptual category effect in visual search: O as a letter or a digit *Perception and Psychophysics* **12** 457
[21] Shiffrin M R and Schneider W 1977 Controlled and automatic human information processing: I. perceptual learning, automatic attending and a general theory *Psychol. Rev.* **84** 127
[22] Stroop R J 1935 Studies of inferences in serial verbal reactions *J. Experiment. Psychol.* **18** 643
[23] Ucros G C 1989 Mood state dependent memory: a meta-analysis *Cognition and Emotion* **3** 139
[24] Oatley K and Johnson-Laird P N 1987 Toward a cognitive theory of emotions *Cognition and Emotion* **1** 29
[25] Posner M I 1978 *Chronometric Explorations of Mind* (Hillsdale, NJ: Lawrence Erlbaum)