Complex Systems **2** (1988) 387–408

Basins of Attraction in a Perceptron-like Neural Network

Werner Krauth
Marc Mézard
Jean-Pierre Nadal
Laboratoire de Physique Statistique,
*Laboratoire de Physique Théorique de l'E.N.S.,**
24 rue Lhomond, 75231 Paris Cedex 05, France

Abstract. We study the performance of a neural network of the perceptron type. We isolate two important sets of parameters which render the network fault tolerant (existence of large basins of attraction) in both hetero-associative and auto-associative systems and study the size of the basins of attraction (the maximal allowable noise level still ensuring recognition) for sets of random patterns. The relevance of our results to the perceptron's ability to generalize are pointed out, as is the role of diagonal couplings in the fully connected Hopfield model.

1. Introduction

An important aspect of the physicists' approach to the study of neural networks has been to concentrate on some standard situations which can be described as probability distributions of instances. For these one can then obtain quantitative comparison of the performances of different networks for large numbers of neurons and connections. A typical example is Hopfield's model [1] of associative memory. In order to quantify its performance, it has been calculated how many independent randomly chosen patterns can be stored with such an architecture, in the "thermodynamic limit" where the number N of neurons is large. For unbiased patterns the original Hebb rule allows to store $0.14N$ patterns [2,3], and more sophisticated, but still perceptron-type, rules [3–5] can reach the upper storage limit [7,8] of $2N$ patterns.

While Hopfield's model and variants of it have been studied thoroughly from a statistical physics point of view (for recent reviews see [9,10]),other widely used models such as layered networks [11] have not been analyzed in this way so far.

*Laboratoire Propre du Centre National de la Recherche Scientifique, associé à l'Ecole Normale Supérieure et à l'Université de Paris Sud.

In this paper we shall deal with the simplest such network, namely the perceptron, which consists of two layers (the usual description of a perceptron [12] contains an initial layer which insures some frozen precoding; in this paper we will not consider this first stage). In particular, we study its associative properties, which are interesting, even though the limitations of the perceptron are well known [13]. A recent review of previous studies of associative properties in other two layers networks can be found in [14].

Associativity is an important feature of neural networks as it allows for the correction of errors: even noisy input configurations can be mapped close to the desired output in the sense of Hamming distance. This is a linearly separable problem, and therefore it can be solved by a perceptron, in contrast to, e.g., the parity and the connectivity problems, which fall into a different class of computational problems, where the correlations between input configurations are not naturally related to the Hamming distance, and where the definition of noise would not be appropriate.

Hereafter we shall study the storage capacity of the perceptron, concentrating on the size of the basins of attraction. The basic result is that the size of the basin of attraction of a pattern depends primarily on its stability. (The precise definition of "stability" is given in the next section. For the pattern to be recognizable by the network in the absence of noise, its stability has to be positive.) For independent random patterns (which may be biased or not) we then calculate the typical stabilities of the patterns achieved by two learning rules, the pseudoinverse rule [15,24] and the minimal overlap rule [6] which can reach optimal stability.

Besides fully determining the associative power, knowledge about the stability achieved in the network gives us information about its capacity; an interesting outcome of our analysis is that the optimal capacity (defined as the ratio of the number of stored patterns to the number of neurons in the input layer) tends to infinity when all the output patterns coincide provided the input patterns are correlated. This result can be interpreted as reflecting the perceptron's ability to generalize: it is able to infer a simple rule from a large enough number of examples.

When studying the auto-association in a perceptron (mapping the patterns — and their nearby configurations — onto themselves) we shall see that a second parameter becomes important in order to obtain large basins of attraction: the values of the diagonal elements in the matrix of couplings, which link the neurons to themselves and tend to freeze the configurations. As the problem of auto-association can be regarded as one single parallel update of a Hopfield network, we then emphasize the relevance of these results to the fully connected Hopfield model. We show by numerical simulations that the stability and the strength of the diagonal couplings are indeed two important parameters for the dynamics of the Hopfield net. There exists an optimal value of the diagonal couplings which maximizes the radius of the basins of attraction.

The evolving simple picture — the stability of a perceptron governs its static properties (the storage capacity) as well as its dynamics (associativity)

— becomes considerably more complicated as soon as one allows several iterations of the perceptron's mapping. The correlations of the synaptic strengths start to play an important role, especially the degree of symmetry of the matrix, and it is no longer possible to make as general statements as for the perceptron. These questions have been stressed in another article [16] which is complementary to the present one. Related work on the role of the stability can be found in [17,18].

The plan of this article is as follows: In section 2 we define the network, its dynamics, the notion of attraction basins and the probability distribution of the patterns to be used for quantitative analysis. In section 3 we compute the quality of retrieval for a noisy input for two general classes of coupling matrices. Section 4 contains a detailed comparison of the associative properties of two specific learning rules: the pseudoinverse and the minimum overlap rules. In section 5 the relevance of the results to auto-association in fully connected networks is discussed. Section 6 shows how some of the results can be interpreted as the ability of generalization of the perceptron. Lastly some concluding remarks are given in section 7.

2. Dynamics of a two-layer network

We study a network of the perceptron type which consists of two layers of neurons. The neurons are Boolean units which we write as (Ising-) spins taking values ± 1. The input layer consists of N spins $\vec{\sigma} = \{\sigma_j = \pm 1, j = 1, \ldots, N\}$ and the output layer contains N' spins $\vec{\sigma}' = \{\sigma'_i = \pm 1, i = 1, \ldots, N'\}$. We shall concentrate on the limiting case where the numbers of neurons N and N' both go to infinity.

The coupling (synapse) between neuron σ_j of the input layer and the neuron σ'_i of the output layer is denoted by J_{ij} so that the coupling matrix (J_{ij}) is of size $(N' \times N)$. The output corresponding to a given input configuration is given by a (zero-)threshold automaton rule

$$\sigma'_i = \text{Sign}\,(\sum_{j=1,N} J_{ij}\sigma_i), \qquad i = 1, \ldots, N' \tag{2.1}$$

The network is taught (through the determination of the J_{ij}) to map each of the $p = \alpha N$ input patterns $\vec{\xi}^{\mu} = \{\xi_j^{\mu} = \pm 1, j = 1, \ldots, N\}$ onto a certain output pattern $\vec{\xi}'^{\mu} = \{\xi_i'^{\mu} = \pm 1, i + 1, \ldots, N\}$. We shall distinguish between two different cases: hetero-association, in which input and output patterns differ and auto-association, in which they are identical. In the latter case we have $N' = N$, and the coupling matrix is square. In this case a special role will be played by the diagonal coupling matrix elements J_{ii} which connect corresponding neurons (i) on the input and on the output layer.

Whenever we need to specialize to a specific distribution of patterns (mostly in section 4), we shall consider the case where the patterns are chosen randomly following the prescription

$$\xi_j'^{\mu} = \begin{cases} +1 \text{ with probability } & (1+m)/2 \\ -1 \text{ with probability } & (1-m)/2 \end{cases} \quad j = 1, \ldots, N \tag{2.2}$$

The probabilities are adjusted so that the patterns carry a mean magnetization $m \equiv 1/N \sum_j \xi_j^{\mu}$ (the parameter m is related to the activity of the neuron). In the case of hetero-association the output patterns are similarly chosen randomly with magnetization m'. This type of bias — and its generalization to more structured hierarchically correlated patterns — has been studied in the case of the Hopfield model [19–21].

For associativity we need that configurations close to $\vec{\xi}^{\mu}$ also be mapped close to $\vec{\xi}'^{\mu}$. To give this notion a precise meaning we shall suppose that the input configuration $\vec{\sigma}$ is chosen randomly, but with a fixed overlap q:

$$q \equiv 1/N \sum_j \xi_j^{\mu} \sigma_j \tag{2.3}$$

with the pattern $\vec{\xi}^{\mu}$ under study. This is achieved by the following choice:

$$\sigma_j = \begin{cases} +\xi_j^{\mu} \text{ with probability } & (1+q)/2 \\ -\xi_j^{\mu} \text{ with probability } & (1-q)/2 \end{cases} \quad j = 1, \ldots, N \tag{2.4}$$

i.e. we assume the noise on different neurons in the input layer to be uncorrelated and of equal strength. The average over the realizations of the noise (2.4) will be denoted by $\langle\,\rangle$.

The perceptron works as an associator, which means that configurations $\vec{\sigma}$ having a large overlap q with $\vec{\xi}^{\mu}$ should also be mapped onto $\vec{\xi}'^{\mu}$. However in the cases we consider this will be exactly true only if the input overlap q is of the order $q = 1 - O(1/\sqrt{N})$. In contrast, the noise will be reduced for a much larger number of configurations with an input overlap of the order of $q = 1 - O(1)$. This means that the output overlap obtained from equation (2.1),

$$q' \equiv 1/N' \sum_{i=1,N'} \xi_i'^{\mu} \sigma_i' \tag{2.5}$$

will be greater than q. In order to characterize this noise reduction by a number r (the "radius" of the basin of attraction), we will therefore choose a cutoff q_c' on the retrieval quality. A noisy input configuration will be said to lie inside the basin of attraction of $\vec{\xi}^{\mu}$ if $q' \geq q_c'$. As we will see in the next section, this will happen with probability 1 when q is larger than a critical value q_c. The radius of the basin of attraction is then defined as $r = 1 - q_c$.

3. Basins of attraction and stability

In order to calculate the basins of attraction of one pattern it turns out that we need rather little information on the elements of the synaptic matrix (J_{ij}). We distinguish two typical situations.

3.1 Equilibrated matrix of synaptic connections

A rather general case is that all the J_{ij} are of the same order of magnitude (i.e. $1/\sqrt{N}$). As we shall see explicitly in the study of the two learning rules, this is the typical situation for random patterns in hetero-association, or in auto-association when the diagonal couplings are set to zero. If this condition is fulfilled, the calculation of moments:

$$\begin{aligned} \langle \xi_i'^{\mu} h_i \rangle &= q\, \xi_i'^{\mu} \sum_j J_{ij} \xi_j \mu \qquad (3.1) \\ \langle (\xi_i'^{\mu} h_i - \langle \xi_i'^{\mu} h_i \rangle)^2 \rangle &= (1-q^2) \sum_j J_{ij}^2 \\ \langle (\xi_i'^{\mu} h_i - \langle \xi_i'^{\mu} h_i \rangle)^4 \rangle &= 3[(1-q^2) \sum_j J_{ij}^2]^2 (1 + O(1/N)) \\ &\cdots \end{aligned}$$

shows explicitly that $\xi_i'^{\mu} h_i$, and therefore $\xi_i'^{\mu} h_i / \sqrt{\sum_j J_{ij}^2}$, are Gaussian random variables with respect to the realization of the input noise (2.4). It is this latter quantity which we refer to as the stability of pattern μ on site i:

$$\Delta_i^{\mu} = \sum_j J_{ij}\, \xi_i'^{\mu} \xi_j^{\mu} / \sqrt{\sum_j J_{ij}^2} \qquad (3.2)$$

It follows from equation (3.1) that this Gaussian random variable has a width equal to $\sqrt{(1-q^2)}$. From equations (2.1) and (2.5) we now find that the average output overlap q' on pattern μ is related to the input overlap by

$$q' = \int P(\Delta) d\Delta \int_0^{\Delta q/\sqrt{1-q^2}} dz \sqrt{\frac{2}{\pi}}\, e^{-z^2/2} \qquad (3.3)$$

where $P(\Delta)$ reflects the site to site fluctuations of the stability of the pattern under study:

$$P(\Delta) = 1/N' \sum_i \delta(\Delta - \Delta_i^{\mu}) \qquad (3.4)$$

Equation (3.3) is the basic result of this section. It shows that the quality of retrieval of one pattern, as measured by the output overlap q', is the better the larger the stability parameters Δ_i^{μ}. The condition of perfect retrieval ($q' = 1$ when q goes to 1) is that almost all stabilities be non-negative.

Now, depending on the learning rule and the choice of patterns, the values of Δ_i^{μ} may fluctuate from site to site, and this will affect the final result for q'. Let us first suppose that the stability parameters are all equal in the thermodynamic limit:

$$\Delta_i^{\mu} = \Delta \quad i = 1, \ldots, N' \qquad (3.5)$$

In this case the radius of the basin of attraction $r = 1 - q_c$ is a function of one single parameter, the stability Δ, given implicitly by

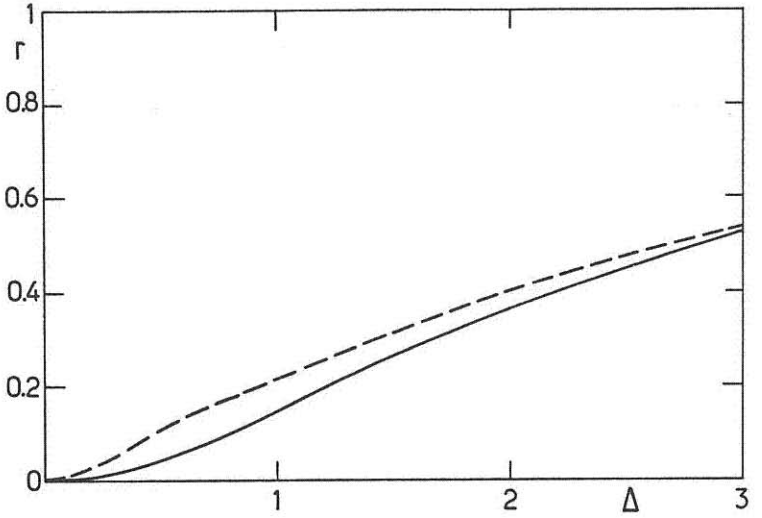

Figure 1: The radius $r = 1 - q_c$ of the basins of attraction as a function of the stability Δ. The cutoff on the output overlap is $q_c' = 0.9$ (see (3.6)). Full curve: zero diagonal couplings $J_{ii} = 0$; dashed curve: optimal diagonal couplings ($J_{ii} = J_{\mathrm{opt}}$), in the case of autoassociation.

$$q_c' = \mathrm{erf}\left(\frac{\Delta(1-r)}{\sqrt{2r(2-1)}}\right) \tag{3.6}$$

This function is plotted in figure 1 for $q_c' = 0.9$ (i.e. less than 5% wrong bits in the output).

As a simple example showing how site-to-site fluctuations of Δ can ruin this result, let us consider the case of Hebb's rule:

$$J_{ij} = 1/N \sum_{\mu} \xi_i'^{\mu} \xi_j{}^{\mu} \tag{3.7}$$

for unbiased *random* patterns ($m = m' = 0$). Then a simple calculation shows that the distribution of stabilities is Gaussian with mean $1/\sqrt{\alpha}$ and width 1. Thus the fluctuations and the mean value of the stability are of the same order. In this special case of Hebb's rule, the fluctuations over configurations and fluctuations from site to site combine in such a way that $\sum_j J_{ij}\xi_i'^{\mu}\xi_j{}^{\mu}$ is a Gaussian with unit width (independent of q). The final resulting noise on the output pattern is independent of the input overlap:

$$q' = \int_0^{q/\sqrt{\alpha}} dz \sqrt{\frac{2}{\pi}}\, e^{-z^2/2} \tag{3.8}$$

The most important qualitative difference between equation (3.8) and equation (3.3) is that for q approaching 1 the output overlap q' does not tend to 1: the memorized pattern differs slightly from the correct output pattern, as in Hopfield's model [1].

3.2 Unequilibrated matrix of synaptic connections

In this section we shall explore the case which is of importance for the perceptron in auto-associative mode and for the Hopfield model. There the diagonal elements of the matrix of couplings play a special role since ξ_i^μ and $\xi_i'^\mu$ are identical. It often happens that these diagonal elements are much larger than the off-diagonal elements. This special role of J_{ii} has been recognized also by Kanter and Sompolinsky [22] (see also [14] and reference therein).

Let us therefore assume that $J_{ii} = J_0(\sqrt{\sum_{j(\neq i)} J_{ij}^2})$, while $J_{ij} = O(1/\sqrt{N})$ $(j \neq i)$. Then the terms J_{ii} in equation (2.1) must be treated separately, and the formula generalizing equation (3.3) is

$$q' = \int P(\Delta) d\Delta \sum_{\tau=\pm 1} \frac{1+q\tau}{2} \int_0^{\frac{\Delta q + J_0 \tau}{\sqrt{1-q^2}}} dz \sqrt{\frac{2}{\pi}}\, e^{-z^2/2} \tag{3.9}$$

The quality of retrieval now depends both on the stability and the diagonal coupling. A well chosen value of J_0 can increase the basin of attraction. Supposing again that all the stabilities are equal to Δ, it is easy to see that the slope evaluated at $q = 1$ is zero if $J_0 < \Delta$, but it is equal to one if $J_0 > \Delta$; if the diagonal coupling is too large, the network cannot flow towards the correct output pattern, even when started from a configuration very close to the input pattern, and its noise will not be reduced. For fixed Δ there exists an optimal value of the diagonal coupling, between 0 and Δ, which maximizes the basin of attraction. The plot of the optimal value as a function of Δ is given in figure 2, and the corresponding new value of the radius of the basin of attraction (evaluated for any Δ, and for J_0 taken at its optimal value) is given in figure 1.

4. Comparison of learning rules: pseudoinverse and minimum overlap

4.1 Definition of the learning rules

Several learning rules have been proposed to choose the J_{ij}'s which allow for the memorization of a given set of patterns $\{\vec{\xi}^\mu, \vec{\xi}'^\mu, \mu = 1, p\}$. A necessary condition for perfect memorization is

$$\xi_i'^\mu = \mathrm{Sign}\Big(\sum_{j=1,N} J_{ij}\xi_j^\mu\Big), \quad i = 1, \ldots, N' \tag{4.1}$$

which is equivalent to having $\Delta_i^\mu > 0$, $i = 1, \ldots, N'$.

One efficient learning rule, the pseudoinverse (P.I.), has been proposed by Kohonen [15] in the context of linear networks. The idea is to look for a matrix (J_{ij}) which is the solution of the equation

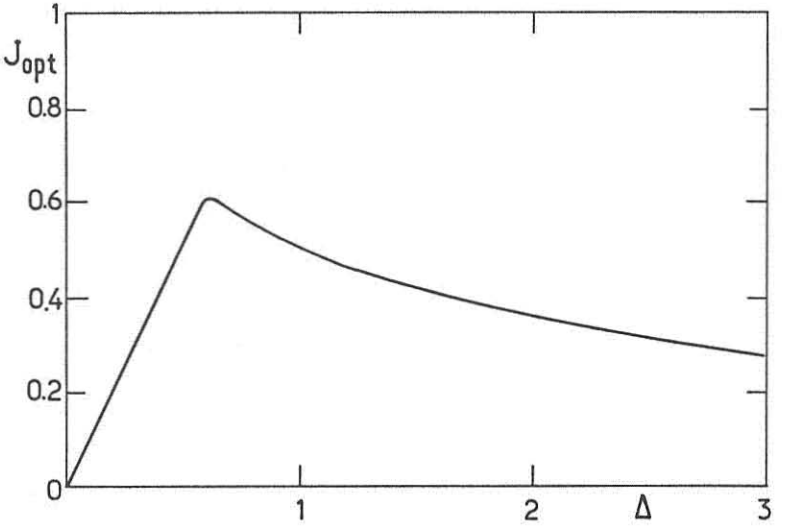

Figure 2: Value of the optimal diagonal couplings $J_{ii} = J_{opt}$ (for autoassociation) as a function of the stabilities Δ.

$$t_i^\mu \equiv \xi_i'^\mu \sum_{j=1,N} J_{ij}\xi_j{}^\mu = 1,\ \mu = 1,\ldots,p,\ i = 1,\ldots,N' \tag{4.2}$$

(or more generally the (J_{ij}) which minimizes $\sum_\mu (1-t_i^\mu)^2$, for each i). When the input patterns are linearly independent, which will always be the case in the situations we study below, an explicit form of the matrix (J_{ij}) is

$$J_{ij} = 1/N \sum_{\mu,\nu} (Q^{-1})_{\mu\nu}\xi_i'^\mu \xi_j{}^\mu \tag{4.3}$$

where $Q_{\mu\nu}$ is the matrix of overlaps of the input patterns:

$$Q_{\mu\nu} = 1/N \sum_{j=1,N} \xi_j^\mu \xi_j^\nu \tag{4.4}$$

(The most general solution of (4.2) contains an additional arbitrary projector onto the subspace orthogonal to the input patterns. For definiteness we keep to the case where this term vanishes).

Another important family of learning rules uses error correcting algorithms to find iteratively a matrix (J_{ij}) such that all the stabilities are positive, the convergence of these algorithms being assured (if a solution exists) by the famous perceptron convergence theorem [12]. Recently this has been refined in order to obtain optimal stability parameters [6]. The corresponding "minimal overlap" (M.O.) algorithm finds a matrix of couplings (J_{ij}) which guarantees that the smallest stability parameter is maximized:

$$K_i = \inf_\mu \Delta_i^\mu \text{ is maximized}, i = 1,\ldots,N' \tag{4.5}$$

We will now calculate the values of the stability which can be reached with these algorithms for sets of random patterns introduced in section 2 (see formula (2.2)).

4.2 Hetero-association

We begin with the case of hetero-association, and with the minimal overlap algorithm. There we follow [8]. For each output neuron i we calculate the fraction of the volume of the N-dimensional space Ω of couplings such that $\Delta_i^\mu \geq K$ for all the patterns $\mu = 1,\ldots p = \alpha N$. Given α there is a critical value of K above which this fraction of Ω has zero volume for $N \to \infty$: this is the maximal value of K, K_{opt}, which can possibly be reached by any algorithm, and which *is* reached by the M.O. algorithm. The calculation of K_{opt} is a mild generalization of the work of Gardner [8] which we shall not reproduce here. The result is that K_{opt} is related to α by the equation:

$$\frac{1}{\alpha} = \sum_{\tau=\pm 1} \frac{1+m'\tau}{2} \int_{\frac{-K_{\text{opt}}+mM\tau}{\sqrt{1-m^2}}}^{\infty} dz \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \left(\frac{K_{\text{opt}} - mM\tau}{\sqrt{1-m^2}} + z \right)^2 \tag{4.6}$$

where M is an auxiliary parameter fixed by the condition that it should make the above expression stationary:

$$\sum_{\tau=\pm 1}(1+m'\tau)\tau \int_{\frac{-K_{\rm opt}+mM\tau}{\sqrt{1-m^2}}}^{\infty} dz \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\left(\frac{K_{\rm opt}-mM\tau}{\sqrt{1-m^2}}+z\right)=0 \quad (4.7)$$

The resulting dependence of $K_{\rm opt}$ as function of α, for various values of m and m', is plotted in figure 3. The maximal storage capacity for this network is $p=\alpha_c N$, where α_c is the critical value for which one gets $K_{\rm opt}=0$. The M.O. algorithm can in fact produce a coupling matrix for which Δ_i^μ is strictly larger than $K_{\rm opt}$ for a few sites i and a few patterns μ. So the basins of attraction will always be larger than (and almost equal to) the values obtained under the assumption that $\mathcal{P}(\Delta)=\delta(\Delta-K_{\rm opt})$. These values can be taken from figures 1 and 3 and the result is plotted in figure 4.

If one wants an exact measure of the radius one must calculate the distribution of stabilities $\mathcal{P}(\Delta)$ reached by the M.O. algorithm. As has been noted by Kepler and Abbott [18] this can be done using the same kind of replica formalism which has been used to determine the value of $K_{\rm opt}$: of the space of couplings such that $\Delta_i^\mu \geq K$ for all the patterns there is a subspace of volume Ω' such that the stability of pattern 1 is $\Delta_i^1=\Delta$. Then

$$\mathcal{P}(\Delta)=\lim_{K\to K_{\rm opt}} \overline{(\Omega'/\Omega)} \quad (4.8)$$

where the $\overline{()}$ means an average over the realizations of random patterns. This is in turn calculated as

$$\mathcal{P}(\Delta)=\lim_{K\to K_{\rm opt}} \lim_{n\to 0} \overline{(\Omega'\,\Omega^{n-1})} \quad (4.9)$$

which allows a replica calculation analogous to the one of Gardner [8]. We shall not reproduce the details of this calculation here, but just quote the results: for the M.O. algorithm, the distribution of stabilities is

$$\begin{aligned}\mathcal{P}(\Delta) &= \sum_{\tau=\pm 1}\frac{1+m'\tau}{2}\{\theta(\Delta-K_{\rm opt})\frac{e^{\frac{-(\Delta-mM\tau)^2}{2(1-m^2)}}}{\sqrt{2\pi(1-m^2)}} \\ &+ \delta(\Delta-K_{\rm opt})\int_{\frac{-K_{\rm opt}+mM\tau}{\sqrt{1-m^2}}}^{\infty}\frac{dz}{\sqrt{2\pi}}e^{-z^2/2}\}\end{aligned} \quad (4.10)$$

where $K_{\rm opt}$ and M are related to the capacity α by equations (4.6) and (4.7). This formula has been derived independently in reference [25]. Using equations (4.10) and (3.3) one can determine the radius of the basins of attraction, which is plotted in figure 4 and differs very little from the one obtained by putting $\mathcal{P}(\Delta)=\delta(\Delta-K_{\rm opt})$.

Let us now turn to the pseudoinverse rule. As we have seen before this leads to a stability which is equal to one before normalization, so that, in order to calculate the parameters Δ_i^μ it is necessary to calculate the normalization of the couplings. From equation (4.3) we have:

$$\sum_j J_{ij}^2=\frac{1}{N}\sum_{\mu,\nu=1,\alpha N}\xi_i^{l\mu}\xi_i^{l\nu}(Q^{-1})_{\mu\nu}\equiv A \quad (4.11)$$

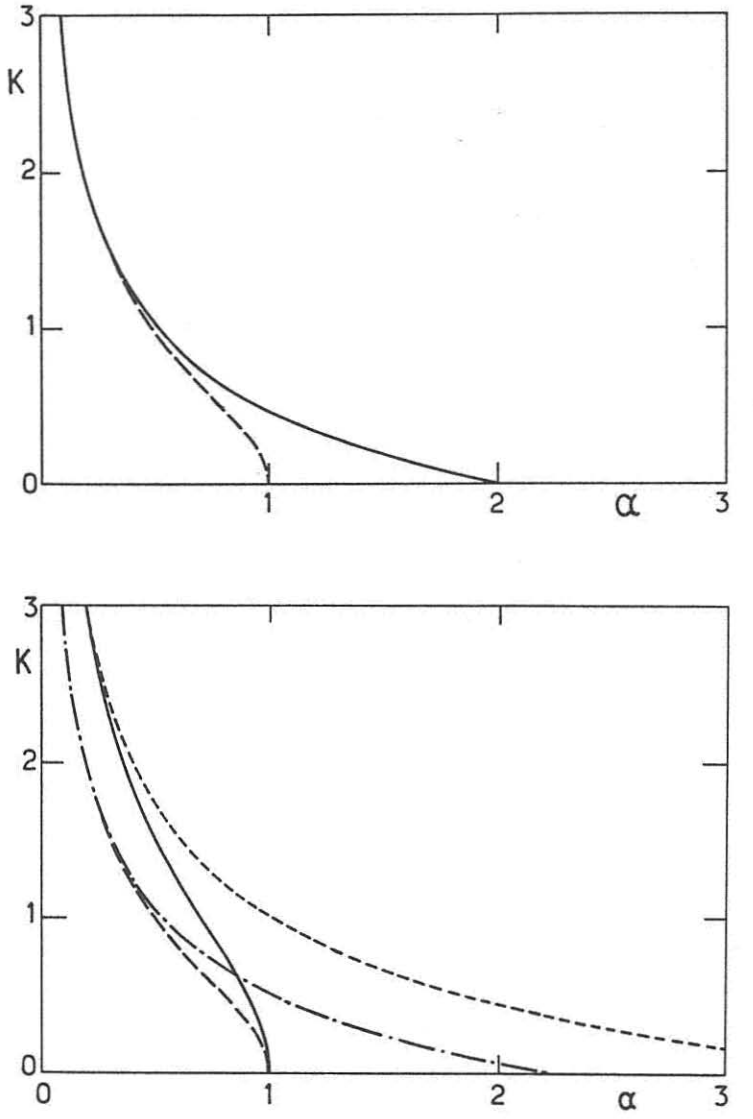

Figure 3: Lower bound on the stabilities, K, as a function of the number of stored patterns per input neuron, $\alpha = p/N$. a) Unbiased input and output patterns ($m = m' = 0$). Full curve: $K_{\rm opt}$, optimal K (M.O. algorithm); dashed curve: K reached by the P.I. algorithm. b) Biased input ($m = 0.4$) and output ($m' = 0.4$ and 0.8) patterns. Full curve: P.I. algorithm ($m' = 0.8$); dashed curve: P.I. algorithm ($m' = 0.4$); dashed-dotted curve: M.O. algorithm ($m' = 0.4$); dotted curve: M.O. algorithm ($m' = 0.8$).

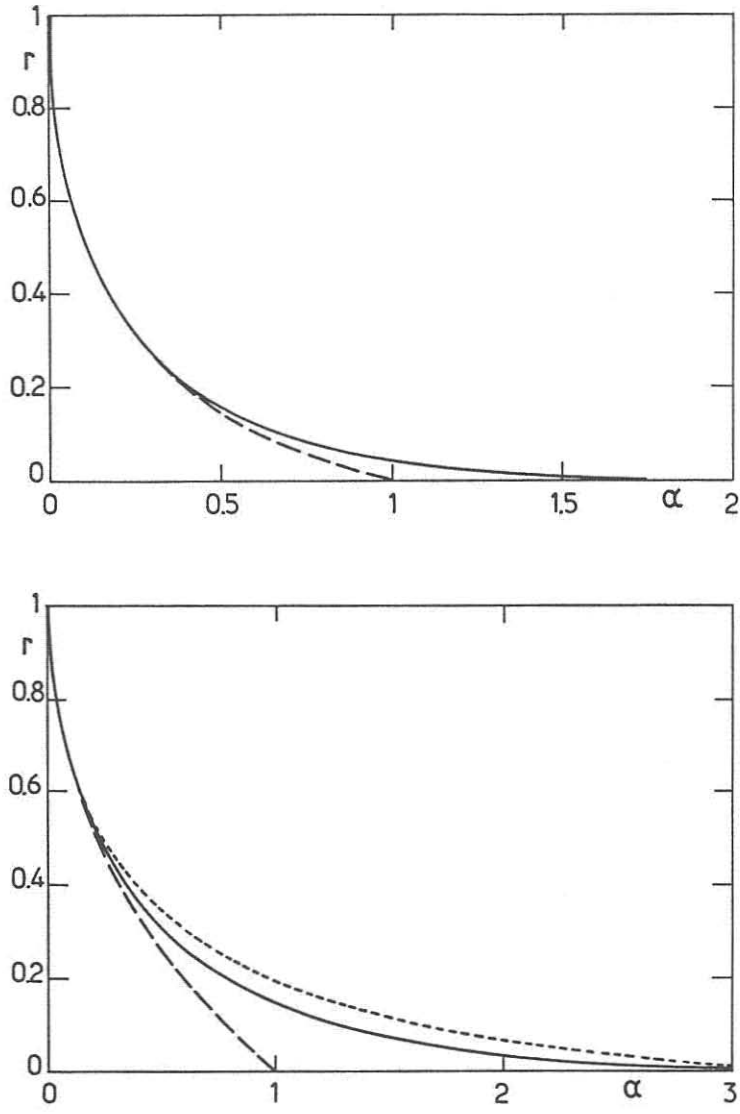

Figure 4: Radius of the basins of attraction r as a function of the number of stored patterns per input neuron, $\alpha = p/N$, in hetero-association. Dashed curve: P.I. algorithm; full curve: theoretical result which would be obtained in a network with fixed stabilities ($\mathcal{P}(\Delta) = \delta(\Delta - K_{\rm opt})$); dotted curve: result for the M.O. algorithm, taking into account the fluctuations of the stabilities above $K_{\rm opt}$. a) Unbiased input and output patterns ($m = m' = 0$). The full curve and dotted curve are essentially indistinguishable on this scale. b) Biased input ($m = 0.4$) and output ($m' = 0.8$) patterns.

where as before Q is the overlap matrix of the input patterns. As the input and output patterns are mutually uncorrelated, it will not be surprising to find that A self-averages to:

$$\bar{A} = \frac{m'^2}{N} \sum_{\mu \neq \nu} \overline{(Q^{-1})_{\mu\nu}} + \frac{1}{N} \sum_{\mu} \overline{(Q^{-1})_{\mu\mu}} \tag{4.12}$$

In order to prove this self averageness and to compute A we write

$$e^{\frac{\lambda^2}{2}A} = \frac{Z(\lambda)}{Z(0)} \tag{4.13}$$

where we have introduced the partition function

$$Z(\lambda) = \int \Pi_{\mu=1}^{P} \frac{dx_\mu}{\sqrt{2\pi}} e^{-\frac{1}{2}\sum_{\mu,\nu} x_\mu Q_{\mu\nu} x_\nu + \frac{\lambda}{\sqrt{N}} \sum_\mu x_\mu \xi_i'^\mu} \tag{4.14}$$

The calculation of $\overline{Z(\lambda)}$ and $\overline{(Z(\lambda)/Z(0))}$ can be done using standard techniques from the statistical physics of disordered systems. It is sketched in appendix A. One finds that A is self averaging and

$$\begin{aligned} \overline{(Q^{-1})_{\mu\mu}} &= \frac{\alpha}{(1-\alpha)(1-m^2)} \\ \mu \neq \nu : \overline{(Q^{-1})_{\mu\nu}} &= \begin{cases} 0 & \text{if } m = 0 \\ \frac{-\alpha}{(1-\alpha)(1-m^2)} & \text{if } m \neq 0 \end{cases} \end{aligned} \tag{4.15}$$

So, finally, the stability parameter for the P.I. rule is

$$\begin{aligned} \Delta_{\text{P.I.}} &= \sqrt{\tfrac{1-\alpha}{\alpha}} && \text{if } m = 0 \\ \Delta_{\text{P.I.}} &= \sqrt{\tfrac{1-\alpha}{\alpha}} \sqrt{\tfrac{1-m^2}{1-m^2}} && \text{if } m \neq 0 \end{aligned} \tag{4.16}$$

(the crossover region which decides if m is close enough to zero is $m < 1/\sqrt{N}$). The P.I. rule realizes the case (3.5), where all stabilities are equal in the thermodynamic limit.

$\Delta_{\text{P.I.}}$ is plotted as a function of α in figure 3. Formulas (3.6) and (4.16) (or figures 1 and 3) allow one to obtain the radius of the basins of attraction for a given number of stored patterns. The result is plotted in figure 4.

4.3 Auto-association

The case of auto-association does not call for changes in the case of the M.O. algorithm, the optimal stability is given exactly by equation (4.6), with $m = m'$.

For the pseudoinverse rule one proceeds as in section 4.2. This requires now the calculation of

$$\sum_{j \neq i} J_{ij}^2 = B - B^2 \tag{4.17}$$

where

$$B = \frac{1}{N} \sum_{\mu,\nu} \xi_i^\mu (Q^{-1})_{\mu\nu} \xi_i^\nu \tag{4.18}$$

Using the same method as before (see appendix B), we find that $\sum_{j(\neq i)} J_{ij}^2$ is self averaging and that for any i it tends toward the limit $\alpha(1-\alpha)$ when $N \to \infty$. On the other hand the stability (3.2) (without the contribution of the diagonal coupling) is

$$\Delta_{\text{P.I.}} = \sqrt{\frac{1-\alpha}{\alpha}} \tag{4.19}$$

This formula coincides with (4.16) for $m' = m$. Therefore the dependence of Δ as a function of α can be read from figure 3.

It is interesting to notice that the M.O. algorithm (as any perceptron-type algorithm) uses the correlations between patterns to increase the storage capacity, while the pseudoinverse method in some way orthogonalizes the patterns, so that its capacity remains the same whatever the correlations between patterns.

5. Relevance to Hopfield-type models

The dynamics of the perceptron in auto-association can be considered as the evolution after one time-step of a Hopfield-type network for associative memory (using parallel updating). So the value of q' is the overlap on the pattern at time $t = 1$. Unfortunately, the reasoning which led to equation (3.3) cannot be reapplied to calculate q_{t+1} as function of $q_t (t \geq 1)$, because in general the noise of the configuration will no longer be Gaussian, the spins on various sites become correlated. Derrida et al. [23] have invented a special type of strongly diluted lattice, on which these correlations can be neglected, to which our results on auto-association with $J_{ii} = 0$ can be applied. Formula (3.3) gives, for parallel updating, the evolution of $q(t+1) = f(q(t))$:

$$q(t+1) = \int P(\Delta) d\Delta \int_0^{\Delta q(t)/\sqrt{1-q(t)^2}} dz \sqrt{\frac{2}{\pi}} e^{-z^2/2} \tag{5.1}$$

and the generalizations to thermal noise and asynchronous updating are straightforward. The radius of the basin of attraction is given by the unstable fixed point $q^\star = f(q^\star)$. We have therefore found that the dynamics at all times on this strongly diluted lattice is governed only by the stability, irrespective of the special learning rule used. The introduction of diagonal coupling ($J_{ii} \neq 0$) reintroduces correlations even on this lattice.

Nevertheless, some of the conclusions of the previous section can be used as hints for what happens in Hopfield's model. There the importance of the stability has been investigated in previous papers [16–18] even though the connection between stability and the basins of attraction there depends also on the correlations of the synaptic matrix (J_{ij}), especially on the symmetry of the matrix [16]. To test the role of diagonal couplings, we have performed numerical simulations on the fully connected Hopfield net with $100 \leq N \leq 400$

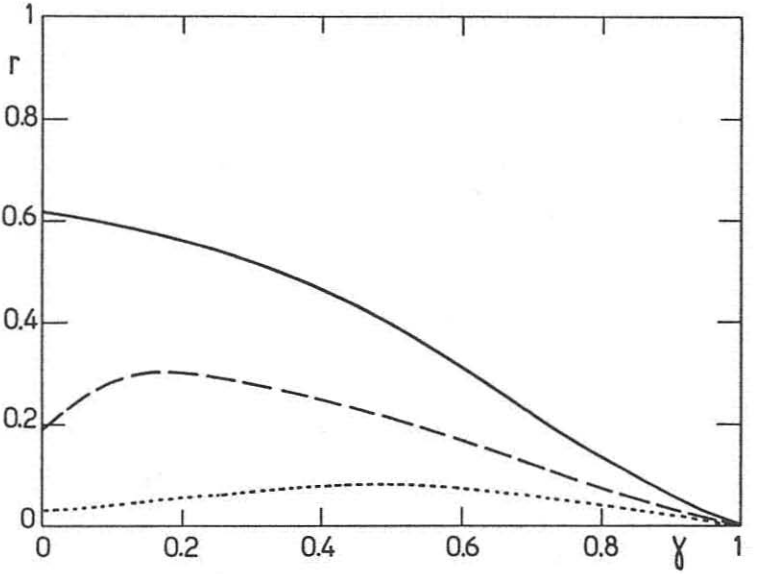

Figure 5: Radius of the basins of attraction r as a function of $\gamma = J_{ii}/\sum_{j\neq i} J_{ij}\xi_i^\mu\xi_j^\mu$, in a fully connected Hopfield model. The non-diagonal couplings are obtained by the P.I. algorithm. Full curve: $\alpha = 1/4$; dashed curve: $\alpha = 1/2$; dotted curve: $\alpha = 3/4$.

and $p = \alpha N (1/4 \leq \alpha \leq 3/4)$ unbiased random patterns. The non-diagonal couplings were chosen with the pseudo-inverse rule while the diagonal ones were all taken to be equal $J_{ii} = \gamma \sum_{j\neq i} J_{ij}\xi_i^\mu\xi_j^\mu$. The results (radius r as a function of γ) are presented in figure 5. They show clearly that varying the strength of the diagonal couplings has a strong effect on the convergence properties of the fully connected net (cf. Kanter and Sompolinsky [22]) and that the best choice of J_{ii} may not be $J_{ii} = 0$. We find e.g that choosing $\gamma \cong 0.15$ instead of $\gamma = 0$ can increase the radius of the basins of attraction by about 50%, for $\alpha = 1/2$.)

6. Using a perceptron to generalize: A simple case

The formulas we have found in section 4.2 show that, for the two algorithms, the size of the basins of attraction increases with the correlations in the output state, i.e. with the value of m', but *only* if the input correlation m is nonzero. Any perceptron-type algorithm can even achieve an infinite capacity in the limit $m' \to 1 (\lim_{m'\to 1} \alpha = \infty)$, provided $m \neq 0$. This is an interesting result which sheds some light on the perceptron's ability to generalize. If $m' \to 1$ there is an unique output state, whatever the input

state. Let us suppose, e. g., that the input correlation is $m > 0$. Then the memorization of p input patterns can be considered from a totally different point of view: the network must send all inputs with positive magnetization towards the unique output state which has been given to it (and by symmetry the input configuration with negative magnetization must be sent to the reversed output state). It is taught to learn this task through the presentation of p examples (the patterns). In this context the fact that the capacity is infinite simply means that the M.O. algorithm is able to learn this task [25]: in fact it is clear that for $p \gg N$ the set of coupling constants reached by the system is stable: each (magnetized) new pattern is automatically memorized, so that it does not lead to a change in the couplings.

This ability to generalize is a nice property. We have pointed it out here as a simple consequence of the results of the previous section, its detailed study is, however, beyond the scope of the present paper.

Let us however point out the relationship with the usual language of data analysis. The problem considered here is a simple case of classification: all patterns are distributed into two classes, depending on the sign of the magnetization. In classification tasks one usually looks for some distance criterium such that patterns belonging to a same class are grouped into a cluster of nearby elements, and patterns of different classes are as far apart as possible (see for example [26]). In the commonly used Discriminant Analysis method, one looks for axes (in state space) such that the projection on these axes optimally distributes the patterns into clusters. In the neural network language, the directions of the N' axes are given by the lines of the coupling matrix of a Perceptron with N' output units [27].

On one given axis i, that is for one given output unit i, the distances between the patterns are in fact directly given by the stabilities as defined in (3.2). Indeed, on axis i defined by the vector $\vec{J}_i = (J_{ij})_{j=1,N}$, the abscissa X_i^μ of the projection of a pattern μ is plus or minus its stability Δ_i^μ — depending on the class it belongs to: in the above example, $X_i^\mu = \Delta_i^\mu$ for a positive magnetization and $X_i^\mu = -\Delta_i^\mu$ for a negative magnetization. The choice of the Pseudo-Inverse rule corresponds to a typical choice in Discriminant Analysis method, which result in the minimization of the dispersion within classes [27]. Indeed, we have seen that all the patterns have the same stability $\Delta_{\text{P.I.}}$. On the axis, the two clusters are reduced to two points, distant of $2\,\Delta_{\text{P.I.}}$. The choice of the M.O. algorithm corresponds to the maximization of the distance between the two clusters: the distance between any two patterns belonging to different classes is at least equal to $2K_i$ (see (4.5)). But now the elements within a cluster are distributed according to the distribution (4.10).

7. Conclusion

We have calculated the storage capacity and the size of the basins of attraction for a perceptron-type network storing random pattern. In the case of hetero-association the important parameter which determines this size is

the stability which is maximized by the M.O. algorithm. When one considers auto-association, another parameter allows to improve the performance of the network: the diagonal couplings. This is also a useful parameter in Hopfield's network.

The dynamics which has been studied in this article is one-step (as the net is feed-forward and consists of two layers). It would be very useful to understand how this may be extended to the dynamics at several time steps, either in fully connected models, or in a multilayered feed forward architecture.

Acknowledgments

We acknowledge helpful discussions with L. F. Abbott, H. Gutfreund, O. Lefèvre, and M. Virasoro. This work has been partially supported by the European program "BRAIN", contract number ST2J-0422-C (EDB), and the numerical work has been performed in part on Sun workstations provided by DRET. WK acknowledges financial support by Studienstiftung des deutschen Volkes.

Appendix A.

We first calculate the average over the choices of random patterns of the partition function $Z(\lambda)$ defined in equation (4.14). Using the definition (4.4) of the matrix Q, we have

$$Z(\lambda) = \int \Pi_{\mu=1}^{P} \frac{dx_\mu}{\sqrt{2\pi}} \Pi_{j=1}^{N} \frac{dt_j}{\sqrt{2\pi}}$$

$$\exp\left(-\frac{1}{2}\sum_j t_j^2 + \frac{i}{\sqrt{N}}\sum_{j,\mu} t_j \xi_j^\mu x_\mu + \frac{\lambda}{\sqrt{N}}\sum_\mu \xi_i'^\mu x_\mu\right) \tag{A.1}$$

The average with respect to the input pattern is

$$\overline{Z(\lambda)} = \int \Pi_{\mu=1}^{P} \frac{dx_\mu}{\sqrt{2\pi}} \Pi_{j=1}^{N} \frac{dt_j}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\sum_j t_j^2 + \frac{i}{\sqrt{N}} m \sum_j t_j \sum_\mu x_\mu\right.$$

$$\left. -\frac{1}{2N}(1-m^2)\sum_j t_j^2 \sum_\mu x_\mu^2 + \frac{\lambda}{\sqrt{N}}\sum_\mu \xi_i'^\mu x_\mu\right) \tag{A.2}$$

Writing $X = 1/N \sum_\mu x_\mu^2$ and enforcing this constraint through an auxiliary parameter $\hat{X}$, we have after integrating over the t_j:

$$\overline{Z(\lambda)} = \int dX \int_{-i\infty}^{i\infty} \frac{d\hat{X}}{2\pi} \int \Pi_{\mu=1}^{P} dx_\mu$$

$$\exp\left\{\frac{1}{2}\hat{X}(NX - \sum_\mu x_\mu^2) - \frac{N}{2}\log(1 + X(1-m^2))\right.$$

$$-\frac{m^2}{2}\frac{1}{1+X(1-m^2)}(\sum_\mu x_\mu)^2+\frac{\lambda}{\sqrt{N}}\sum_\mu \xi^{\prime\mu}{}_i x_\mu\Bigg\} \tag{A.3}$$

The integral over the x_μ is Gaussian with a quadratic form

$$e^{-\frac{1}{2}\sum_{\mu,\nu} x_\mu M_{\mu\nu} x_\nu} \tag{A.4}$$

where

$$M_{\mu,\nu}=\hat{X}\delta_{\mu,\nu}+\frac{m^2}{1+X(1-m^2)} \tag{A.5}$$

Performing the integrations over x_μ' s we find

$$\overline{Z(\lambda)}=\int dX\int_{-i\infty}^{i\infty}\frac{d\hat{X}}{2\pi}e^{N\left\{\frac{X\hat{X}}{2}-\frac{1}{2N}\log\det M-\frac{1}{2}\log(1+X(1-m^2))\right\}}$$

$$\times e^{\frac{\lambda^2}{2N}\sum_{\mu,\nu}\xi_i^{\prime\mu}(M^{-1})_{\mu\nu}\xi_i^{\prime\nu}} \tag{A.6}$$

The determinant and the inverse of M are

$$\det M=\hat{X}^{p-1}\left[\hat{X}+p\frac{m^2}{1+X(1-m^2)}\right] \tag{A.7}$$

$$(M^{-1})_{\mu\nu}=\frac{1}{\hat{X}}\left[\delta_{\mu\nu}-\frac{m^2}{pm^2+\hat{X}(1+X(1-m^2))}\right] \tag{A.8}$$

For p and N going to infinity with fixed capacity $\alpha=p/N$, one can perform the integrals over X and $\hat{X}$ by saddle point. It is easy to see from equation (A.8) that the last term in equation (A.6) does not contribute to the saddle point. The solution to the saddle point equations are

$$X=\frac{\alpha}{\hat{X}}=\frac{\alpha}{(1-\alpha)(1-m^2)} \tag{A.9}$$

so that finally

$$\begin{aligned}\overline{Z(\lambda)}&=\overline{Z(0)}e^{\frac{\lambda^2}{2}\frac{\alpha}{1-\alpha}\frac{1-m'^2}{1-m^2}} && \text{if } m\neq 0\\ \overline{Z(\lambda)}&=\overline{Z(0)}e^{\frac{\lambda^2}{2}\frac{\alpha}{1-\alpha}} && \text{if } m=0\end{aligned} \tag{A.10}$$

This is not yet enough to prove the announced result since we have computed so far $\overline{Z(\lambda)}/\overline{Z(0)}$ instead of $\overline{Z(\lambda)/Z(0)}$. In order to compute this last quantity we could use the replica method (see e.g. Mézard et al. [10])

$$\overline{\left(\frac{Z(\lambda)}{Z(0)}\right)}=\lim_{n\to 0}\overline{Z(\lambda)Z(0)^{n-1}} \tag{A.11}$$

$Z(\lambda)Z(0)^{n-1}$ being calculated by the introduction of n copies of the variables x_μ.

However, in the present case, this is not even necessary, because $Z(\lambda)$ turns out to be self averaging: we can directly calculate $\overline{Z^2(\lambda)}$ using the same techniques as before. We introduce two types of x-variables x_μ and x'_μ and write

$$X = 1/N \sum_\mu x_\mu^2, X' = 1/N \sum_\mu x'^2{}_\mu, Q = 1/N \sum_\mu x_\mu x'_\mu,$$

together with the auxiliary parameters $\hat{X}$, $\hat{X}'$, $\hat{Q}$. Then

$$\begin{aligned}
\overline{Z^2(\lambda)} &= \int dX \frac{d\hat{X}}{2\pi} dX' \frac{d\hat{X}'}{2\pi} dQ \frac{d\hat{Q}}{2\pi} \int \Pi_{\mu=1}^P \frac{dx_\mu}{\sqrt{2\pi}} \frac{dx'_\mu}{\sqrt{2\pi}} && \text{(A.12)} \\
&\times \exp\left\{ \frac{\hat{X}}{2}(NX - \sum_\mu x_\mu^2) + \frac{\hat{X}'}{2}(NX' - \sum_\mu x'^2_\mu) \right. \\
&\left. + \hat{Q}(NQ - \sum_\mu x_\mu x'_\mu) \right\} && \text{(A.13)} \\
&\times \exp\left\{ -\frac{N}{2} \log D - \frac{m^2}{2} \frac{1 + X'(1-m^2)}{D} (\sum_\mu x_\mu)^2 \right. \\
&\left. - \frac{m^2}{2} \frac{1 + X(1-m^2)}{D} (\sum_\mu x'_\mu)^2 \right\} && \text{(A.14)} \\
&\times \exp\left\{ m^2 Q \frac{1-m^2}{D} (\sum_\mu x_\mu)(\sum_\mu x'_\mu) + \frac{\lambda}{\sqrt{N}} \sum_\mu \xi_i^{l\mu} (x_\mu + x'_\mu) \right\}
\end{aligned}$$

where D is a notation for the determinant

$$D = \{1 + X(1-m^2)\}\{1 + X'(1-m^2)\} - Q^2(1-m^2)^2 \qquad \text{(A.15)}$$

The Gaussian integrals over x_μ and x'_μ lead as before to a determinant of the corresponding quadratic form. Let us begin with $\lambda = 0$. We obtain

$$\begin{aligned}
\overline{Z^2(0)} &= \int dX \frac{d\hat{X}}{2\pi} dX \frac{d\hat{X}'}{2\pi} dQ \frac{d\hat{Q}}{2\pi} \\
&\times \exp \frac{N}{2} \left\{ X\hat{X} + X'\hat{X}' + 2Q\hat{Q} - \log D - \alpha \log(\hat{X}\hat{X}' - \hat{Q}^2) \right\} && \text{(A.16)}
\end{aligned}$$

A careful examination of the saddle point equations of this integral shows that the dominant saddle point is always at

$$\begin{aligned}
Q &= \hat{Q} = 0 && \text{(A.17)} \\
X &= \frac{\alpha}{\hat{X}} = \frac{\alpha}{(1-\alpha)(1-m^2)}; \qquad X' = \frac{\alpha}{\hat{X}'} = \frac{\alpha}{(1-\alpha)(1-m^2)}
\end{aligned}$$

so that

$$\overline{Z^2(0)} \overset{N\to\infty}{\sim} \overline{Z(0)}^2 \tag{A.18}$$

As the term in λ in equation (A.12) is not of the leading order in N, it cannot change the saddle point and one finds

$$\overline{Z^2(\lambda)} \overset{N\to\infty}{\sim} \overline{Z(\lambda)}^2 \tag{A.19}$$

This shows that, for any λ, the fluctuations of $Z(\lambda)$ can be neglected; therefore we obtain from equation (A.10)

$$\overline{\left(\frac{Z(\lambda)}{Z(0)}\right)} = \overline{\left(e^{\frac{\lambda^2}{2}A}\right)} \overset{N\to\infty}{\sim} \begin{cases} e^{\frac{\lambda^2}{2}\frac{\alpha}{1-\alpha}\frac{1-m'^2}{1-m^2}} & \text{if } m' \neq 0 \\ e^{\frac{\lambda^2}{2}\frac{\alpha}{1-\alpha}} & \text{if } m' = 0 \end{cases} \tag{A.20}$$

which is the announced result.

Appendix B.

We calculate the stability parameters for the P.I. rule in the case of auto-association. From equation (4.18) we need to calculate

$$B = \frac{1}{N}\sum_{\mu,\nu=1}^{P} \xi_i^\mu \xi_i^\nu (Q^{-1})_{\mu\nu} \tag{B.1}$$

As in appendix A we write $e^{\frac{\lambda^2}{2}B} = \overline{\left(\frac{Z(\lambda)}{Z(0)}\right)}$, with

$$\begin{aligned} Z(\lambda) &= \int \Pi_{\mu=1}^{P} \frac{dx_\mu}{\sqrt{2\pi}} \Pi_{j=1}^{N} \frac{dt_j}{\sqrt{2\pi}} \exp(-\frac{1}{2}\sum_j t_j^2) \times \\ &\times \left[\exp\left(\frac{i}{\sqrt{N}}\sum_{j,\mu} t_j \xi_j^\mu x_\mu + \frac{\lambda}{\sqrt{N}}\sum_\mu \xi_i^\mu x_\mu\right)\right] \end{aligned} \tag{B.2}$$

We proceed as in appendix A: we first average over the ξ_j^μ $(j \neq i)$, then integrate over the t_j $(j \neq i)$ and finally integrate over the x_μ, with $X = 1/N \sum_\mu x_\mu^2$, fixed by an auxiliary parameter $\hat{X}$. The result is

$$\begin{aligned} Z(\lambda) &= \int dX d\hat{X} \exp N \left(\frac{\hat{X}X}{2} - \frac{\alpha}{2} Ln\hat{X} - \frac{1}{2} Ln[1 + X(1-m^2)]\right) \times \\ &\times \left[\int \frac{dt_i}{\sqrt{2\pi}} \exp\left(-\frac{t_i^2}{2} + \frac{1}{2N}\sum_{\mu,\nu}(M^{-1})_{\mu\nu}(\lambda + it_i)^2 \xi_i^\mu \xi_i^\nu\right)\right] \end{aligned} \tag{B.3}$$

where M and M^{-1} are given in equations (A.5–8). As before it is easily seen that the last term does not affect the saddle point (A.9) on X and $\hat{X}$, so that

$$\frac{\overline{Z(\lambda)}}{\overline{Z(0)}} = \frac{\int \frac{dt}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} e^{\frac{1}{2}\frac{\alpha}{1-\alpha}(\lambda+it)^2}}{\int \frac{dt}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} e^{\frac{1}{2}\frac{\alpha}{1-\alpha}(it)^2}} = e^{\frac{\lambda^2}{2}\alpha} \tag{B.4}$$

The self averageness of $Z(\lambda)$ derived in appendix A also applies here, so that

$$\overline{\left(\frac{Z(\lambda)}{Z(0)}\right)} \underset{\sim}{N\to\infty} e^{\frac{\lambda^2}{2}\alpha} \tag{B.5}$$

Hence B self averages to

$$\lim_{N\to\infty} B = \alpha \tag{B.6}$$

and

$$\sum_{j(\neq i)} J_{ij}^2 = \alpha(1-\alpha) \tag{B.7}$$

References

[1] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities", *Proc Nat Acad Sci USA* , **79** (1982) 2554.

[2] D. J. Amit, H. Gutfreund, and H. Sompolinsky, "Storing infinite number of patterns in a spin-glass model of neural network", *Phys Rev Lett*, **55** (1985) 1530.

[3] A. Crisanti, D. J. Amit, and H. Gutfreund, "Saturation level of the Hopfield model for neural network", *Europhys Lett*, **2** (1986) 337.

[4] E. Gardner, N. Stroud, and D. J. Wallace, "Training with noise and the storage of correlated patterns in a neural network model", Preprint Edinburgh 87/394 (1987).

[5] S. Diederich and M. Opper, "Learning of correlated patterns in spin-glass networks by local learning rules", *Phys Rev Lett*, **58** (1987) 949.

[6] W. Krauth and M. Mézard, "Learning algorithms with optimal stability in neural networks", *J Phys A: Math Gen*, **20** (1987) L745.

[7] S. Venkatesh, in *Neural Networks for Computing, AIP Conference Proceedings*, **151**, ed. J. S. Denker (Am. Inst. Phys., New York, 1986) 440.

[8] E. Gardner, "The space of interactions in neural networks models", *J. Phys. A*, **21** (1988) 257; "Maximum storage capacity in neural networks", *Europhys Lett*, **4** (1987) 481.

[9] D. J. Amit, "The properties of models of simple neurals networks", in *Heidelberg Colloquium on Glassy Dynamics*, eds. J. L. van Hemmen and I. Morgenstern (Springer, Berlin, 1987).

[10] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).

[11] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing*, vol.1 and 2 (Bradford Books, Cambridge MA, 1986).

[12] F. Rosenblatt, *Principles of Neurodynamics* (Spartan Books, New York, 1962).

[13] M. Minsky and S. Papert, *Perceptrons* (MIT Press, Cambridge, MA, 1969).

[14] F. Fogelman Soulié, P. Gallinari, Y. Le Cun, and S. Thiria, "Automata networks and artificial intelligence" in *Automata Networks in Computer Science*, eds. F. Fogelman Soulié, Y. Robert, and M. Tchuente (Manchester Univ. Press, 1987) 133.

[15] T. Kohonen, *Self Organization and Associative Memory* (Springer, Berlin, 1984).

[16] W. Krauth, J. P. Nadal, and M. Mézard, "The roles of stability and symmetry on the dynamics of neural networks", to appear in *J Phys A: Math Gen*, **21** (1988) 2995.

[17] B. M. Forrest, "Content addressability and learning in neural networks", *J. Phys. A: Math Gen.*, **21** (1988) 245.

[18] T. B. Kepler and L. F. Abbott, "Domains of attraction in neural networks", Brandeis University preprint (1988).

[19] N. Parga and M. A. Virasoro, "The ultrametric organization of memories in a neural network", *J. Physique* (Paris) **47** (1986) 1857.

[20] D. J. Amit, H. Gutfreund, and H. Sompolinsky, "Neural networks with correlated patterns: towards pattern recognition", *Phys. Rev.*, **A35** (1987) 2293; H. Gutfreund, "Neural Networks with Hierarchically Correlated Patterns", *Phys. Rev. A*, **37** (1988) 570.

[21] M. V. Feigelman and L. B. Ioffe, "The augmented models of associative memory, asymmetric interaction and hierarchy of patterns", *Int. J. of Mod. Phys.*, **B1** (1987) 51.

[22] I. Kanter and H. Sompolinsky, "Associative recall of memory without errors", *Phys. Rev.*, **A35** (1987) 380.

[23] B. Derrida, E. Gardner, and A. Zippelius, "An exactly soluble asymmetric neural network model", *Europhys Lett*, **4** (1987) 167.

[24] L. Personnaz, I. Guyon, and G. Dreyfus, "Information storage and retrieval in spin-glass like neural networks", *J de Physique*, **L16** (1985) 359.

[25] P. Delgiudice, S. Franz, and M. A. Virasoro, Preprint 605, Rome University.

[26] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (Wiley, 1973).

[27] P. Gallinari, S. Thiria, and F. Fogelman Soulié, "Multilayer Perceptrons and Data Analysis," to appear in ICNN 88, IEEE Annual International Conference on Neural Networks (San Diego, CA, July 24–27, 1988).