Properties of neural networks storing spatially correlated patterns

1992 J. Phys. A: Math. Gen. 25 3701

(http://iopscience.iop.org/0305-4470/25/13/019)

J. Phys. A: Math. Gen. 25 (1992) 3701–3720. Printed in the UK

Properties of neural networks storing spatially correlated patterns

Rémi Monasson

Laboratoire de Physique Théorique de l'Ecole Normale Supérieure†, 24 Rue Lhomond, 75231 Paris Cedex 05, France

Received 26 February 1992, in final form 21 April 1992

Abstract. We study the behaviour of a feedforward neural network supplied with spatially organized data. This inner structure is taken into account by a matrix C_{ij}, whose coefficients equal the average correlation between two pixels i and j of the input patterns. The storage capacity α is computed as a function of the required stability and of the eigenvalues of C. We propose a geometrical transformation allowing an intuitive interpretation of these results. Numerical simulations using real and binary patterns show a very good agreement with the theory. Finally, we analyse the synaptic couplings correlations resulting from the training of the network with structured patterns. Focusing on exponentially decreasing correlations in one and two dimensions, we find that they exhibit a 'Mexican hat' profile, the excitatory centre size of which depends on α.

1. Introduction

For the last few years, storage capacities of neural networks have been intensively studied. Thanks to statistical mechanics tools developed by Gardner [1], many different networks have been investigated, ranging from the simple perceptron [2] to multilayered neural nets [3]. Besides the network architecture, the nature of their couplings (discrete or continuous) [4] and the number of activity levels of the units [5] have been shown to influence their storage properties.

From another point of view, we may ask the following question. Considering a particular network, how will its capacity and its synaptic weights distribution depend on the internal structure of the data? Such internal correlations would for example be present if we try to store binary images and their corresponding outputs in a fixed neural network. Until now, the statistical physics analysis has been restricted to patterns without internal correlations between the different pixels. In other words, the notion of distance separating the input neurons was meaningless. This is the problem which we solve in the present paper.

Here we consider a perceptron-like neural network. We introduce internal correlations inside the input patterns and compute its storage capacity. It is natural that the internal correlations in patterns in turn induce correlations in the coupling matrix. We are also able to determine these induced correlations exactly. We consider

† Unité Propre de Recherche du CNRS, associée à l'Ecole Normale Supérieure et à l'Université de Paris-Sud.

0305-4470/92/133701+20$04.50

the hetero-associative case, in which the output is chosen independently of the input pattern.

In this paper, two main results will be derived. Firstly, the storage capacity remains unchanged whatever the inner structure of the input patterns. Secondly, this spatial structure induces interesting correlations between the synaptic weights: two couplings are positively correlated when they are close enough and become anti-correlated when the distance separating them increases. One important finding of the calculations concerns the duality between correlations in the patterns and those in the couplings. In fact, the calculations which we will present solve, at the same time, the problem of correlated patterns and the case of storage with correlated couplings (for uncorrelated patterns).

2. Presentation of the problem

2.1. *General definitions*

We consider a single-layer network including N binary neurons S_i, $i = 1, \ldots, N$, and an ouput σ. The couplings J_i between units S_i and σ are continuous. The output value is computed following the classical rule

$$\sigma = \operatorname{sign}\left(\frac{1}{\sqrt{N}} \sum_i J_i S_i\right). \tag{2.1}$$

Let us now choose an N binary components vector ξ and a corresponding output σ. We say that the pair (ξ, σ) is stored by the network if ξ is mapped onto σ:

$$\frac{\sigma}{\sqrt{N}} \sum_i J_i \xi_i > 0. \tag{2.2}$$

We also require that noisy versions η of ξ (i.e. having a large overlap with ξ) should be correctly classified (i.e. mapped onto σ) as often as possible (see section 3.1). A common way to increase this robustness is to impose a stability $\kappa > 0$ [1, 7]. The formula (2.2) becomes

$$\frac{\sigma}{\sqrt{N}} \sum_i J_i \xi_i > \kappa. \tag{2.3}$$

For a training set (ξ^μ, σ^μ), $\mu = 1, \ldots, P$, we define its size as $\alpha = P/N$. In the large-N limit, the critical capacity $\alpha_c(\kappa)$ is the largest size below which there exist with probability one, couplings $\{J\}$ fulfilling the condition (2.3) for each μ. The storage capacity $\alpha_c(\kappa)$ is a self-averaging quantity, and thus does not depend on the particular choice of the training set but only on its statistical distribution.

2.2. *The auto-correlation matrix*

We impose a probability distribution on the input patterns as follows:

$$\overline{\xi_i^\mu} = 0 \qquad \forall i, \forall \mu$$
$$\overline{\xi_i^\mu \xi_j^\nu} = \delta^{\mu\nu} C_{ij} \qquad \forall (i,j), \forall (\mu,\nu) \tag{2.4}$$

where the bar denotes an average over this distribution. The first equation means there is no external bias, which occurs, for example, when a pattern and its opposite are drawn with the same probability. The Kronecker symbol in (2.4) implies that patterns are chosen independently of each other. The matrix C contains informations about correlations inside one pattern ξ and we choose it to obey the following requirements:

$C_{ii} = \overline{\xi_i^2} = 1 \qquad \forall i.$

C_{ij} depends only on the distance $|i-j|$ between the neurons i and j because of translational and rotational invariance of the pattern ξ. No restriction is imposed on the dimension of the input space, and subscripts may be considered as vectors.

C must be a positive matrix: $\sum_{ij} x_i C_{ij} x_j = \overline{[\sum_i x_i \xi_i]^2} \geqslant 0 \qquad \forall \boldsymbol{x}.$

The common case with unbiased random patterns is given by $C_{ij} = \delta_{ij}$.

In the following we will call $\lambda_1, \lambda_2, \ldots, \lambda_N$ the eigenvalues of C. When N tends to infinity, we note that

$$[f(\lambda)]_\lambda = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} f(\lambda_i) . \tag{2.5}$$

The conditions (2.4) do not actually define the whole distribution of probabilities $\mathcal{P}(\xi)$ from which the patterns ξ are drawn. Further calculations based on the replica method show however that one needs only the first two moments of $\mathcal{P}$ in order to compute the critical capacity α_c, provided that the higher connected moments satisfy clustering conditions

$$\sum_{i_1, i_2, \ldots, i_k} \overline{\xi_{i_1} \xi_{i_2} \cdots \xi_{i_k}}^{\text{connected}} \sim \Gamma_k N \qquad (N \to \infty) \tag{2.6}$$

where the Γ's are some constants.

Any matrix C may be thus obtained with Gaussian patterns following the law

$$\mathcal{P}(\xi) = \frac{1}{\sqrt{\det\ 2\pi C}} \exp\left(-\frac{1}{2} \sum_{ij} \xi_i (C^{-1})_{ij} \xi_j \right) . \tag{2.7}$$

Alternatively one may choose binary patterns. For instance, a one-dimensional Ising model at temperature T will lead to equilibrium configurations distributed as in (2.4), (2.6) with

$$C_{ij} = \left[\tanh\left(\frac{1}{T}\right) \right]^{|i-j|} . \tag{2.8}$$

2.3. The case of exponentially decreasing correlations

Let us first remark that inner correlations may result from a preprocessing stage. Previous studies have shown that neurons connected to a fully random input layer but with overlapping receptive fields will be synchronized [6]. Such induced correlations are positive and decrease with the distance separating the processing units. They may be taken into account as shown in section 2.2.

Moreover, the above description of the auto-correlation matrix is valid whatever the input space dimension, and general results for any C will be derived in section 2. Then we focus on patterns presenting patches of typical sizes L. Defining $x = \exp(-1/L)$, we consider the matrix

$$C_{ij} = x^{|i-j|}. \tag{2.9}$$

The parameter x ($0 \leqslant x < 1$) is the correlation strength inside one input pattern. When x equals zero, we recover the classic case of independent patterns and $\mathcal{P}(\xi_i) = \frac{1}{2}\delta(\xi_i + 1) + \frac{1}{2}\delta(\xi_i - 1)$. As x tends to 1, only two patterns remain present: the one with all parts equal to one, and its opposite. The patterns are indeed drawn on a circle (or a torus) so as to ensure periodic boundaries conditions.

In section 4.1, we apply the general theory to one-dimensional patterns verifying (2.9). The more realistic case of two-dimensional inputs presenting the same covariance matrix (but for continuous patterns) is investigated in section 4.2.

3. General theory for any correlation matrix

3.1. Analytical calculation of the critical capacity

We follow the now classic method introduced by Gardner [1]. For a training set (ξ^μ, σ^μ), we define the fraction of the space of the couplings which store the patterns with the stability κ:

$$V(\{\xi^\mu, \sigma^\mu\}) = \frac{\int \mathrm{d}J\, \delta(J^2 - N) \prod_{\mu=1}^{P} \theta\left[\kappa - \sigma^\mu (J \cdot \xi^\mu / \sqrt{N})\right]}{\int \mathrm{d}J\, \delta(J^2 - N)} \tag{3.1}$$

where we have imposed spherical constraints on the synaptic weights. As $\ln V$ is an extensive quantity, one assumes it becomes self-averaging in the large-N limit. We then compute its average $\overline{\ln V}$ over the patterns distribution using the replica-symmetric approximation

$$\frac{1}{N}\overline{\ln V} = -\frac{1}{2} q\hat{q} + s\hat{s} + \hat{u} + \frac{1}{2}\ln(2\pi) - \frac{\hat{q}}{2}\left[\frac{\lambda}{2\hat{u} + (2\hat{s} - \hat{q})\lambda}\right]_\lambda$$
$$- \frac{1}{2}\left[\ln(2\hat{u} + (2\hat{s} - \hat{q})\lambda)\right]_\lambda + \alpha \int \mathrm{D}z \ln H\left(\frac{z\sqrt{q} + \kappa}{\sqrt{s - q}}\right) \tag{3.2}$$

where

$$H(z) = \int_z^{+\infty} \mathrm{D}t$$
$$\mathrm{D}z = \frac{\mathrm{d}z}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2} \tag{3.3}$$

and the two order parameters q, s are defined as follow. Let $\langle\langle \cdots \rangle\rangle$ be the average over the coupling J storing perfectly the P patterns. The order parameters are typical overlaps between these solutions:

$$q = \frac{1}{N}\sum_{ij} C_{ij}\ \overline{\langle\langle J_i\rangle\rangle\langle\langle J_j\rangle\rangle}$$
$$s = \frac{1}{N}\sum_{ij} C_{ij}\ \overline{\langle\langle J_i J_j\rangle\rangle}\,. \tag{3.4}$$

$\hat{q}, \hat{s}$ are Lagrange parameters enforcing constraints (3.4) and $\hat{u}$ ensures the normalization of J. These five parameters are found by solving the stationarity equations associated with (3.2). In the usual uncorrelated case, $C_{ij} = \delta_{ij}$ and $s = 1$. For more complicated correlation matrices, the critical capacity will be reached in the limit of q equals s, i.e. when the space of suitable couplings shrinks to a 'single point'.

The analysis of the saddle-points equations is not easy below the critical line $\kappa(\alpha_c)$ when the eigenvalues distribution is not explicitly given. In section 4, we give some results for the particular choice $C_{ij} = x^{|i-j|}$ concerning the evolution of the order parameters q and s towards their critical common value s_c.

We therefore restrict our analysis to the critical line $\kappa(\alpha_c)$ where the three Lagrange parameters diverge and the stationarity equations simplify to

$$s_c\left[\frac{\lambda}{(\lambda+\nu_c)^2}\right]_\lambda = \left[\frac{\lambda^2}{(\lambda+\nu_c)^2}\right]_\lambda$$
$$\alpha_c = \frac{1}{H(-\kappa/\sqrt{s_c})}\left[\frac{\lambda}{\lambda+\nu_c}\right]_\lambda \tag{3.5}$$

where ν_c is a function of κ and s_c defined by

$$\nu_c = \kappa^2 + \kappa\sqrt{\frac{s_c}{2\pi}}\,\frac{e^{-\kappa^2/2s_c}}{H(-\kappa/\sqrt{s_c})} \tag{3.6}$$

and monotonously increases from 0 ($\kappa = 0$) to ∞ ($\kappa = \infty$).

For zero stability, the critical capacity is always equal to $\alpha_c = 2$, whichever correlation matrix we choose. This is not surprising since correlations do not destroy the linear independance of any subset of N patterns taken among the P patterns. They are thus in general positions and the argument of Cover [8] is still valid. The critical value of q and s is

$$s_c = \frac{1}{(1/\lambda)_\lambda} \tag{3.7}$$

which is always less than 1. A simple geometric derivation of this result will be given in the next paragraph.

The stability appears in (3.5) only through $\kappa/\sqrt{s_c}$. This may be explained by calculating the probability that a random noisy version η of a given stored pattern (for instance ξ^1) is mapped onto the right output (σ^1). This probability depends of course on the overlap between η and its reference and also on the pure pattern

stability Δ ($=\frac{\sigma^1}{\sqrt{N}} J \cdot \xi^1$) [7]. One can compute in the limit $\alpha \to \alpha_c(\kappa)$ the distribution of Δ [7]:

$$\mathcal{P}_\kappa(\Delta) = H\left(-\frac{\kappa}{\sqrt{s_c}}\right)\delta(\Delta-\kappa) + \theta(\Delta-\kappa)\frac{e^{-\Delta^2/2s_c}}{\sqrt{2\pi s_c}}. \tag{3.8}$$

As compared with the usual uncorrelated case $C_{ij} = \delta_{ij}$ and $s_c = 1$, this formula indicates that the effective stability of the training set is $\kappa/\sqrt{s_c}$ rather than κ [7]. One can verify that near $\alpha_c \simeq 2$

$$\frac{\kappa}{\sqrt{s_c}} \simeq (2-\alpha_c)\frac{\sqrt{2\pi}}{8} \tag{3.9}$$

is independent of C. Thus, in the critical capacity limit, the presence of correlations inside the training patterns should not modify the ability of the network to classify correctly noisy versions of the stored data.

All results have been derived here under the replica-symmetric assumption. It is physically justified by the convexity of the subset of vectors storing the P patterns on the N-dimensional sphere. Moreover, we have analytically checked the stability of the mean-field solution with respect to small deviations of the order parameters $\hat{u}^a, \hat{s}^a, \hat{q}^{ab}, s^a, q^{ab}$ around the symmetric saddle point ($a < b$ are replica indices running from 1 to n and the limit $n \to 0$ is taken afterwards) [9].

Extending the analysis of the case $C_{ij} = \delta_{ij}$ [1], we have found two types of eigenvectors for the second-derivatives matrix of $g = \frac{1}{N}\overline{\ln V}$. It turns out that there are 5 eigenvalues related to the longitudinal fluctuations and that the $5(n-1)$ eigenvectors symmetric under interchange of all but one of the indices give five more eigenvalues which are indeed degenerate with the longitudinal ones in the limit $n \to 0$.

An instability may in fact arise from the transverse fluctuations of g. They give two eigenvalues (each one being $\frac{1}{2}n(n-3)$-fold degenerate) whose product equals

$$p(\alpha,\kappa) = \alpha\left(\frac{\partial^2 g}{(\partial q^{12})^2} - \frac{2\,\partial^2 g}{\partial q^{12}\partial q^{13}} + \frac{\partial^2 g}{\partial q^{12}\partial q^{34}}\right) \times \left(\frac{\partial^2 g}{(\partial \hat{q}^{12})^2} - \frac{2\,\partial^2 g}{\partial \hat{q}^{12}\partial \hat{q}^{13}} + \frac{\partial^2 g}{\partial \hat{q}^{12}\partial \hat{q}^{34}}\right) - 1\,. \tag{3.10}$$

The stability of the replica-symmetric solution requires that no sign change of the eigenvalues occurs between $\alpha = 0$ and $\alpha = \alpha_c(\kappa)$. Using definition (3.6), we find $p(0,\kappa) = -1$ and

$$p(\alpha_c(\kappa),\kappa) = -\frac{\nu_c}{s_c+\nu_c} < 0 \tag{3.11}$$

proving that the critical line $\alpha_c(\kappa)$ lies in the domain of stability of the symmetric saddle point.

3.2. Geometrical interpretation

Reaching a geometrical understanding of the previous analytical calculations is not obvious, since correlations involved here cannot be interpreted as easily as in the uniform-bias case. To cope with this difficulty, we eliminate the self-correlations inside each pattern by diagonalizing the C matrix and apply the inverse transformation to the couplings vector so as to leave the storing prescriptions (2.3) unchanged. We are thus led to a strictly equivalent problem where we have $P = \alpha N$ unbiased random patterns (ξ^μ, σ^μ) in a perceptron whose synaptic vector J is constrained to lie on the N-dimensional ellipsoid $\mathcal{E}$:

$$\sum_\gamma \frac{J_\gamma^2}{\lambda_\gamma} = N . \tag{3.12}$$

We seek the critical capacity of this network and compute the fraction of couplings storing all the (ξ^μ, σ^μ) with a stability κ defined as

$$\kappa = \max_{J \in \mathcal{E}} \left[\min_{1 \leqslant \mu \leqslant P} \left(\sigma^\mu \frac{J \cdot \xi^\mu}{\sqrt{N}} \right) \right] . \tag{3.13}$$

Calculations with the replica method on this model can be done easily and lead of course to the same result as before. Let us now relate this problem to the usual Gardner analysis.

To each vector J belonging to $\mathcal{E}$, we associate the vector J' pointing in the same direction and lying on the sphere $\mathcal{S}$ with radius $\sqrt{N}$ by

$$J' = \frac{1}{\sqrt{\frac{1}{N}\sum_\gamma J_\gamma^2}} J . \tag{3.14}$$

Reciprocally, any J' on $\mathcal{S}$ defines one single vector J on $\mathcal{E}$ by

$$J = \frac{1}{\sqrt{\frac{1}{N}\sum_\gamma \frac{(J')_\gamma^2}{\lambda_\gamma}}} J' . \tag{3.15}$$

The zero stability case is easily solved thanks to this very simple bijection between $\mathcal{S}$ and $\mathcal{E}$. Let us indeed consider $P = \alpha N$ random unbiased patterns $\eta^\mu = \sigma^\mu \xi^\mu$ on $\mathcal{S}$. Since proportionality factors in (3.14) and (3.15) are always strictly positive, finding J on $\mathcal{E}$ storing all the patterns is equivalent to finding J' on $\mathcal{S}$ satisfying the same property. Therefore the storage capacity is always $\alpha_c = 2$ [1, 8], regardless of the matrix C, i.e. the shape of the ellipsoid. When α reaches α_c, there remains, to leading order in N, only one vector J_0' verifying $J_0' \cdot \eta^\mu > 0 \ \forall \mu$. Then (3.15) defines 'one single' J_0 on $\mathcal{E}$, whose squared norm is the critical parameter s_c. When averaging over the training set, J_0' is an isotropic distribution on $\mathcal{S}$ and we recover (3.7).

We now analyse the positive stability case. Let us assume as before that we have drawn lots for $P = \alpha N$ unbiased patterns $\eta^\mu = \sigma^\mu \xi^\mu$ on the sphere $\mathcal{S}$ with $\alpha < 2$. There exists J_0' on $\mathcal{S}$ maximizing the stability $\kappa_G(\alpha)$ of the patterns:

$$\kappa_G(\alpha) = \max_{J' \in \mathcal{S}} \left[\min_{1 \leqslant \mu \leqslant P} \left(\frac{J' \cdot \eta^\mu}{\sqrt{N}} \right) \right] = \min_{1 \leqslant \mu \leqslant P} \left(\frac{J_0' \cdot \eta^\mu}{\sqrt{N}} \right) \tag{3.16}$$

which is given by [1]. The vector J_0 on $\mathcal{E}$ obtained from (3.15) leads to a stability

$$\kappa_0 = \min_{1 \leqslant \mu \leqslant P} \left(\frac{J_0 \cdot \eta^\mu}{\sqrt{N}} \right) = \frac{\kappa_G(\alpha)}{\sqrt{\frac{1}{N} \sum_\gamma (J_0')_\gamma^2 / \lambda_\gamma}} . \tag{3.17}$$

Unfortunately, κ_0 is lower than the optimal stability κ defined in (3.13). Let us indeed consider a small shifting $\mathrm{d}J$ on the surface of $\mathcal{E}$ from J_0 (see figure 1). Constraint (3.12) is described by

$$\sum_\gamma \frac{(J_0)_\gamma}{\lambda_\gamma} \mathrm{d}J_\gamma = 0 . \tag{3.18}$$

During this slight movement, the stability has increased by $\mathrm{d}\kappa$ from its initial value κ_0 with

$$\mathrm{d}\kappa = \frac{\kappa_0}{(J_0)^2} \sum_\gamma (J_0)_\gamma \mathrm{d}J_\gamma . \tag{3.19}$$

We see that $\mathrm{d}\kappa$ may be strictly positive provided that the auto-correlation matrix is not proportional to the identity. So knowning $\kappa_G(\alpha)$ provides us only with a lower bound for κ when averaging (3.17) over training data:

$$\kappa > \frac{1}{\sqrt{\langle 1/\lambda \rangle}} \kappa_G(\alpha) . \tag{3.20}$$

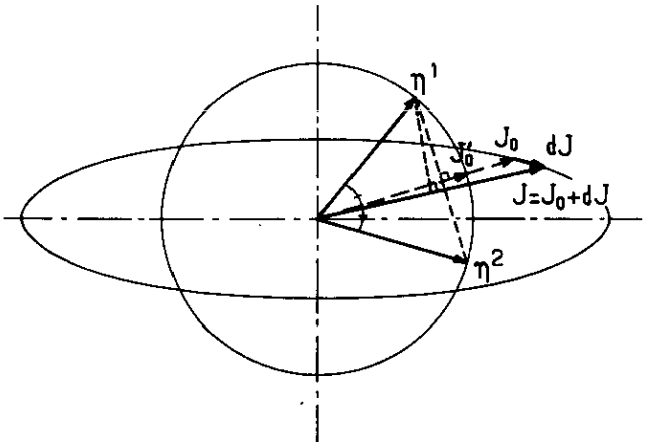

Figure 1. The sphere $\mathcal{S}$ and the ellipsoid $\mathcal{E}$ are drawn in the case $N = 2$. Patterns η^1 and η^2 rely on the cone of stability κ whose axis is the optimal vector J_0'. A small shifting $\mathrm{d}J$ from the prolonged vector J_0 may improve the stability of β^1 though the angle between $J_0 + \mathrm{d}J$ and η^1 increases.

These arguments may be extended to the case $\overline{\xi}_i \neq 0$. If the inputs are biased with a uniform bias m, the connected self-correlation matrix becomes $C_{ij} = (1 - m^2)\delta_{ij}$. $\mathcal{E}$ is now a sphere, obtained from $\mathcal{S}$ thanks to a homothetic reduction of ratio equal to $\sqrt{\lambda} = \sqrt{1 - m^2}$. Condition (3.18) implies that $\mathrm{d}\kappa = 0$ and we find [1]

$$\kappa = \frac{1}{\sqrt{\langle 1/\lambda \rangle}} \kappa_G(\alpha) = \sqrt{1 - m^2}\, \kappa_G(\alpha) . \tag{3.21}$$

3.3. Study of the induced synaptic structure

We analyse here the influence of the data spatial structure on the synaptic vectors storing perfectly $P = \alpha_c N$ inputs patterns ($\kappa > 0$). We restrict ourselves to the critical stability line since calculations become simpler: all quantities no longer fluctuate around their mean values $\langle\langle \cdots \rangle\rangle$. As we deal in this paper with hetero-associative storage, we have obviously

$$\overline{\langle\langle J_i \rangle\rangle} = 0 \,. \tag{3.22}$$

However, inside one pattern, the pixels ξ_i and ξ_j are not independent and lead to correlations between the couplings J_i and J_j. One easily gets (for instance by inserting an infinitesimal external field h acting on J in (3.1) and differentiating $\overline{\ln V}$ with respect to h_i and h_j)

$$\overline{\langle\langle J_i J_j \rangle\rangle} = \frac{[C(C+\nu_c I)^{-2}]_{ij}}{[C(C+\nu_c I)^{-2}]_{ii}} \tag{3.23}$$

where the coefficient ν_c is given by (3.6).

Thus, for a small number of patterns, the correlations between synaptic weights reflect the internal structure of the patterns

$$\overline{\langle\langle J_i J_j \rangle\rangle} = C_{ij} \qquad (\alpha_c \to 0) \,. \tag{3.24}$$

On the other hand, in the saturation limit, the structure of the couplings is given by the inverse matrix

$$\overline{\langle\langle J_i J_j \rangle\rangle} = s_c \, (C^{-1})_{ij} \qquad (\alpha_c \to 2) \tag{3.25}$$

whereas (3.23) interpolates continuously between these two behaviours in the finite stability range.

These results are applied to exponentially decreasing correlations in one- and two-dimensional patterns in the following parts.

3.4. Modified Hebb rule for self-correlated patterns

For uncorrelated patterns and small enough training sets (σ^μ, ξ^μ), an efficient storage rule is the Hebb rule, corresponding to the synaptic vector

$$L = \frac{1}{\sqrt{P}} \sum_{\mu=1}^{P} \sigma^\mu \xi^\mu \,. \tag{3.26}$$

L_i is nothing but the average over the training set of $\sigma^\mu \xi_i^\mu$, namely the ith component of pattern μ. However, in our present case, correlations between different components of the input data suggest us that a given coupling J_i should take into account not only the ith pixel of each input pattern but also its neighbours. We are thus led to change rule (3.26) into

$$J_i = \sum_j K_{ij} \, L_j \tag{3.27}$$

where K is a linear kernel assumed to be translationally invariant. As a result, this simple transformation also modifies the correlations among the components of the synaptic vector

$$\overline{J_i J_j} = \frac{(KCK)_{ij}}{(KCK)_{ii}} \tag{3.28}$$

ensuring the correct normalization $\overline{J_i^2} = 1$.

We choose K optimally thanks to a signal-over-noise analysis [11]. With Gaussian inputs, the distribution of the stability $\sigma^1 J.\ \xi^1/\sqrt{J^2}$ of a training pattern follows a normal law whose mean m and variance v are

$$\begin{aligned} m &= \frac{(KC)_{ii}}{\sqrt{\alpha\,(KCK)_{ii}}} \\ v &= \frac{(KCKC)_{ii}}{(KCK)_{ii}} \end{aligned} \tag{3.29}$$

The probability $f(K)$ that this pattern is stored (i.e. has a positive stability) equals $m/\sqrt{v}$ and is of course all the more important when this signal-over-noise ratio is large [11]. It is maximal for $K = C^{-1}$ and $f(C^{-1}) = H(-1/\sqrt{\alpha})$ is independent of C as the critical capacity $\alpha_c = 2$ obtained in section 3.1.

Instead of optimizing the above quantity $f(K)$, let us restrict our attention now to matrices K ensuring only

$$f(K) = g \tag{3.30}$$

where g is a given positive number lower than $H(-1/\sqrt{\alpha})$. But we seek the largest possible mean stability m for our given pattern. Maximizing m under the constraint (3.30) gives

$$K = (C + \nu I)^{-1} \tag{3.31}$$

where ν is obtained when inserting the above result into condition (3.30). From formulae (3.28) and (3.31), we recover the behaviour contained in (3.23). It is quite remarkable that the simple storage prescription described here agrees qualitatively with the much more complicated replica calculations done in section 3.3.

4. Application to exponentially decreasing correlations

We now apply the previous theory to some particular cases of correlation matrices (see section 2.3). The simple choice of exponentially decreasing correlations is also in good agreement with some experimental fits of natural images [13]. We consider one-dimensional patterns for which all calculations can be done analytically and lead to interesting features, mainly concerning the synaptic correlations. The bidimensional case is more realistic and gives roughly the same results.

4.1. One-dimensional patterns

According to section 2.3, we consider the correlation matrix (2.9). When N grows to infinity, the distribution of its eigenvalues become a continuous function of $\phi \in [0, 2\pi]$:

$$\lambda(\phi) = \frac{1 - x^2}{1 - 2x\cos\phi + x^2}. \tag{4.1}$$

Following notation of (2.5), the average of any function f over the eigenvalues is

$$[f(\lambda)]_\lambda = \frac{1}{\pi}\int_0^\pi \mathrm{d}\phi\ f[\lambda(\phi)] \tag{4.2}$$

and we have for instance

$$[\lambda]_\lambda = 1 \qquad [\lambda^2]_\lambda = \left[\frac{1}{\lambda}\right]_\lambda = \frac{1 + x^2}{1 - x^2}. \tag{4.3}$$

At the saturation limit ($\alpha_c = 2$), the couplings correlations are related to the inverse matrix

$$(C^{-1})_{ij} = \frac{1 + x^2}{1 - x^2}\delta_{i,j} - \frac{x}{1 - x^2}(\delta_{i,j-1} + \delta_{i,j+1}) \tag{4.4}$$

which exhibits one positive centre ($i = j$) surrounded by two negative peaks ($i = j \pm 1$). The surprising existence of anti-correlations between the synaptic weights appears clearly in the finite-stability range ($0 < \alpha_c < 2$). From (3.23), we get

$$\overline{\langle\langle J_i J_j\rangle\rangle} = \chi^{|i-j|}\left(1 - \frac{|i-j|}{\zeta}\right) \tag{4.5}$$

where

$$\begin{aligned} \chi &= \frac{1 + x^2}{2x} + \frac{1 - x^2}{2x\nu_c}\left[1 - \sqrt{1 + 2\nu_c\left(\frac{1 + x^2}{1 - x^2}\right) + \nu_c^2}\right] \\ \zeta &= \sqrt{1 + 2\nu_c\left(\frac{1 + x^2}{1 - x^2}\right) + \nu_c^2} - \frac{1 + \chi^2}{1 - \chi^2}. \end{aligned} \tag{4.6}$$

Thus, two couplings are positively correlated if the distance separating them is lower than a characteristic length ζ and negatively correlated otherwise (see figure 2). Synaptic correlations look like the well known 'Mexican-hat' profile, including a positive excitatory centre and negative inhibitory flanks. We will comment on this point in section 4.3.

Using the eigenvalue function (4.1), the saddle-point equations (3.5) may be solved numerically. The evolution of the optimal stability κ as a function of α is shown on figure 3. We see it is a decreasing function of x for fixed α. It vanishes for $x = 1$ since the two remaining patterns $(+\cdots+)$ and $(-\cdots-)$ are opposite and thus cannot be mapped independently onto random outputs.

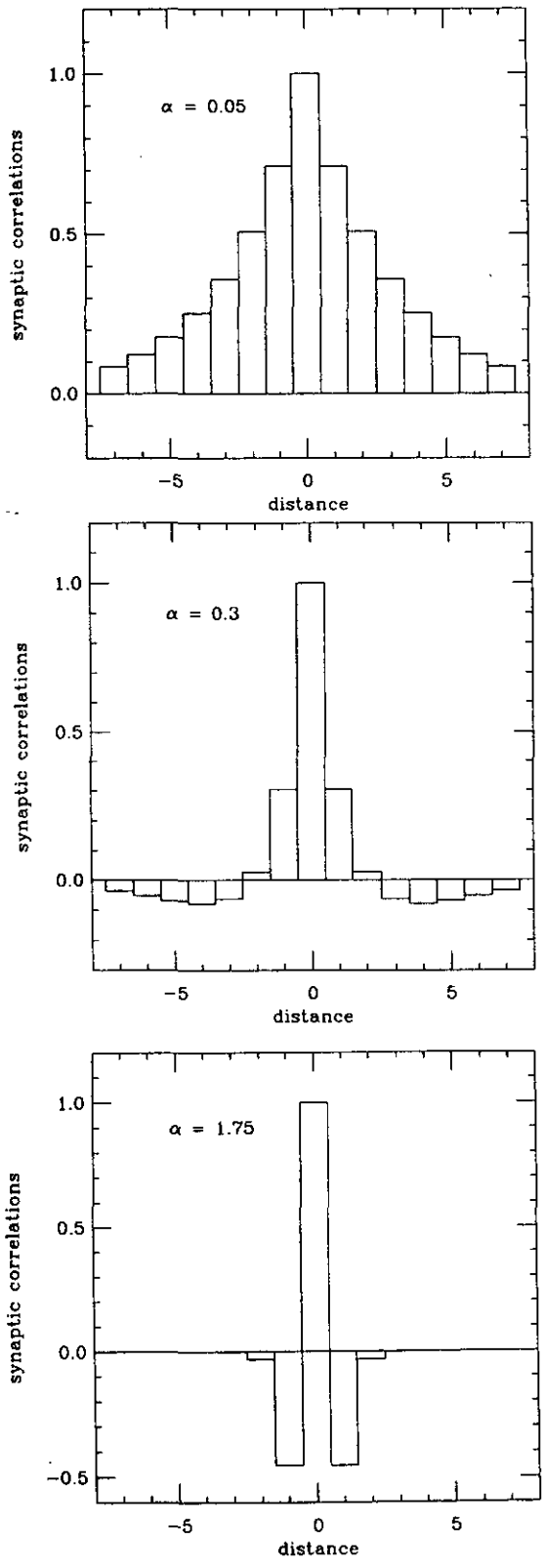

Figure 2. One-dimensional case: synaptic profiles $\overline{\langle J_i J_j \rangle}$ as a function of the distance $i - j$ for $x = 0.8$. The sizes α of the training set are equal to 0.05, 0.3 and 1.75 and the corresponding values of ν_c are 16.64, 1.49 and 0.016 respectively (see (3.24)). As α increases, the correlations curves interpolate smoothly between C_{ij} ($\alpha_c = 0$) and $(C^{-1})_{ij}$ ($\alpha_c = 2$).

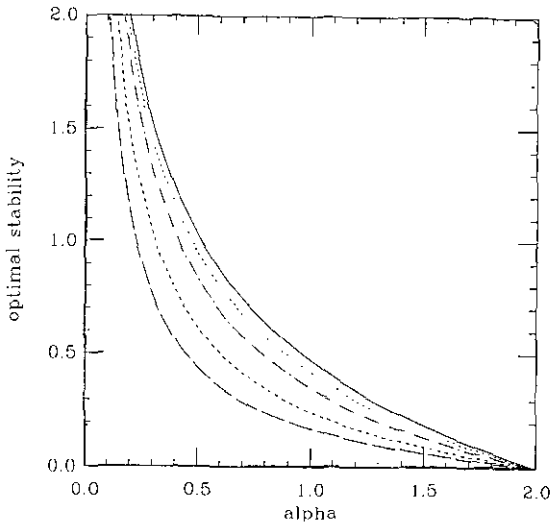

Figure 3. The optimal stability κ as a function of the size α for $x = 0$ (——), 0.4 (· · · · · ·), 0.6 (— · —), 0.8 (- - - -) and 0.9 (— — —). Note that all curves end in $\alpha_c = 2$ for zero stability and that κ is a decreasing function of x for given α (see figure 6).

We have also studied the evolution of q and s as functions of the size α for different x and zero stability. In this case, the three Lagrange parameters $\hat{q}, \hat{s}, \hat{u}$ appearing in (3.2) may be eliminated and we find two implicit equations for q and s

$$\frac{2\left(\frac{q}{s-q}\right)^{3/2}\left(\frac{1-x^2}{1+x^2}\right)}{1+\sqrt{1+\frac{4q}{s(s-q)^2}\left(\frac{1-x^2}{1+x^2}\right)}} = \frac{\alpha s}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \mathrm{D}z\; z\; \frac{\mathrm{e}^{-\frac{z^2}{2}\left(\frac{q}{s-q}\right)}}{H\left(z\sqrt{\frac{q}{s-q}}\right)} \tag{4.7}$$

and

$$\frac{1}{2}\left[1+\sqrt{1+\frac{4q}{s(s-q)^2}\left(\frac{1-x^2}{1+x^2}\right)}\right](1+x^2) = \sqrt{4x^2+\left(\frac{1-x^2}{s-q}\right)^2}. \tag{4.8}$$

The critical parameter $s_c = (1-x^2)/(1+x^2)$ is obtained when $\alpha \to 2$. For $x = 0$, (4.8) gives $s = 1$ and we recover the classical saddle point equation for q in (4.7).

Figure 4 displays the curves $s(\alpha), q(\alpha)$ for several values of x. The parameter $q(\alpha)$ may have a maximum at $\alpha < 2$ but $s - q$ always decreases with α. This may be explained thanks to geometrical arguments exposed in section 3.2. Formula (3.4) tells us that q is nothing but the mean overlap between two synaptic vectors storing a given training set. When increasing α, the volume of suitable couplings shrinks and the typical angle between the two solutions decreases. However, the variations of s prove that their norms decrease too. Thus, if the ellipsoid is flat enough, i.e. if x is sufficiently large, one can get non-monotonous q-functions as shown on figure 4.

4.1.1. Numerical simulations. We have checked our analytical predictions in one dimension with Gaussian patterns following the distribution law (2.7). Since we are

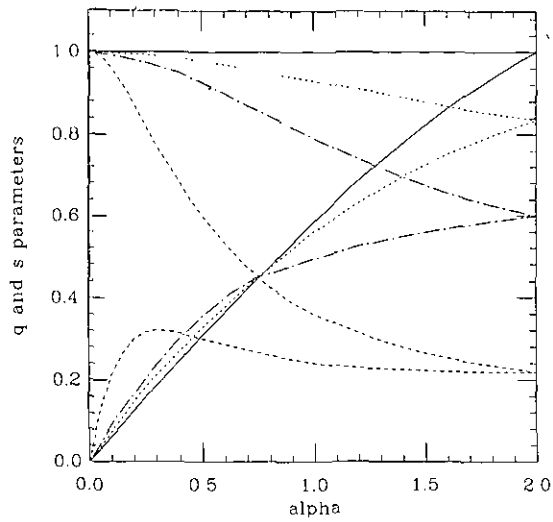

Figure 4. The curves $q(\alpha)$ and $s(\alpha)$ for several values of the correlation strength $x = 0$ (——), 0.3 (· · · · · ·), 0.5 (— · —) and 0.8 (– – –). Note the presence of a maximum at $\alpha < 2$ for q at sufficiently large x. The quantity $s - q$ is a decreasing function of α. At critical capacity $\alpha_c = 2$, s and q are both equal to $(1 - x^2)/(1 + x^2)$.

interested in evaluating the optimal stability $\kappa(\alpha, x)$ of a single-layered perceptron, we have resorted to the so-called Minover algorithm. According to the notations of reference [10], there are three relevant parameters:

(i) The lower stability c: when updating J, the algorithm does not care about the norm of the synaptic vector. It stops as soon as the scalar product between J and each pattern becomes greater than an arbitrary bound c. Thus the optimal stability is reached in the limit $c \to \infty$ and the errors due to finite c are larger since α approaches 2. A good estimate of the accuracy of finite c results is given by the performance guarantee factor A (see [10]) which satisfies $\kappa(c) \leqslant \kappa(\infty) \leqslant A\kappa(c)$ for each sample. A is easily computed from the norm of the synaptic vector and the number of running steps.

(ii) The number T of samples: the algorithm runs for one given training set and we must average the resulting stabilities over the statistical distribution of the self-correlated patterns. All the data we present here have been obtained with $100 < T < 1000$. The error bars take into account these statistical errors and the uncertainty related to A.

(iii) The size of the input layer N: this parameter influences upon the mean value $\overline{\kappa}$ and the width of the distribution of the stabilities $\kappa(\alpha, x)$ for different training sets. Attempts with different sizes ($N = 100, 200, 300$) show a weak dependence of $\overline{\kappa}$ with respect to N (less than the error bars due to finite c) for Gaussian patterns. We present here results obtained with $N = 300$.

In figure 5, we show the result of the simulation $\kappa(\alpha)$ for $x = 0.8$ and compare it with the usual $x = 0$ case. In figure 6, we display the optimal stability κ at $\alpha = 0.5$ (full curve). Simulations have been performed for different correlation strengths $x = 0.2$, 0.4, 0.6, 0.8 and 0.9. There is an excellent agreement with the theoretical predictions.

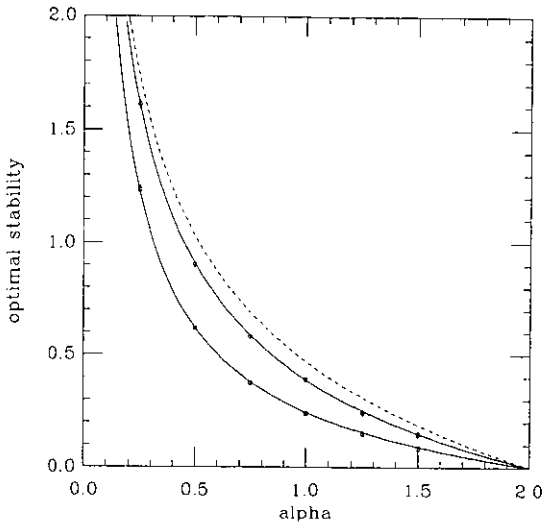

Figure 5. The optimal stability as obtained from numerical simulations with the Minover algorithm. The broken curve is the usual stability obtained with uncorrelated patterns ($x = 0$). The full curves are theoretical predictions for $x = 0.5$ and $x = 0.8$. The points associated with the two curves indicate the results found with Ising and Gausssian patterns respectively.

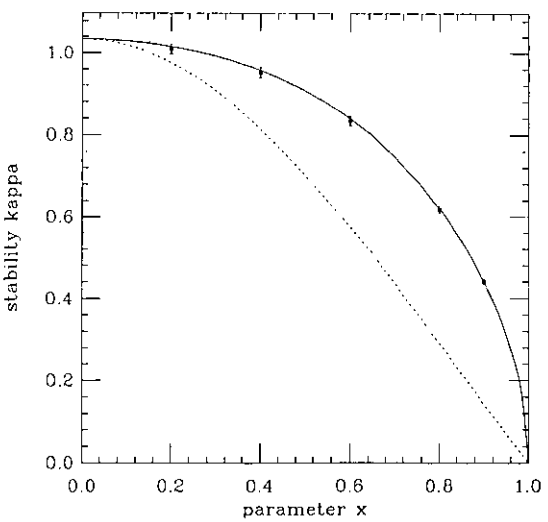

Figure 6. The optimal stability at fixed $\alpha = 0.5$ as a function of the correlation strength and the results of the numerical simulations with Gaussian patterns (see text for details). The full curve shows theoretical predictions. The broken curve is the lower bound obtained from (3.20). The points indicate the results of the numerical simulations.

Another theoretical provision, the importance of which has been already stressed in section 2.2 is that the storage properties depend only on the first two moments of the patterns distribution. This remark stems from replica method calculations and is a generalization of the well known equivalence between Gaussian and binary patterns,

for instance in the Hopfield model [12].

To verify this statement, we have simulated the storage of non-Gaussian patterns. From a practical point of view, we have considered a one-dimensional Ising lattice with N spins described by the Hamiltonian

$$H(\xi) = -\sum_{i=1}^{N} \xi_i \xi_{i+1} \qquad (\xi_{N+1} = \xi_1)\,. \tag{4.9}$$

Equilibrium configurations at temperature T will satisfy the law (2.8). We generate the patterns through a Monte Carlo simulation of this Ising model. We have carried out simulations with the Minover algorithm for $x = 0.5$. Figure 5 displays the numerical results and their very good agreement with the theoretical curve. Moreover, we have seen that important finite-size effects are present up to sizes equal to $N = 500$ as soon as x increases. They are due to the corrective terms (2.6) which diverge quickly when $x \to 1$.

4.2. Bidimensional patterns

We now consider bidimensional images ξ_i where the vector i is taken continuous for simplicity of calculation. Considering two pixels i and j, they are spatially correlated with

$$C_{ij} = \frac{1}{2\pi L^2} e^{-|i-j|/L}\,. \tag{4.10}$$

L is the characteristic length of the inner correlations and the right-hand side coefficient ensures that C_{ij} tends to $\delta(i-j)$ when L vanishes. However, as soon as L is strictly positive, it only defines a unit of distance for the continuous input patterns. Without loss of generality, we choose $L = 1$ in the following.

The matrix (4.10) is easily diagonalized by means of a Fourier transform, and in the momentum space q the corresponding eigenvalues are

$$\lambda(q) = \frac{1}{(q^2+1)^{3/2}}\,. \tag{4.11}$$

Inserting this eigenvalue distribution in (3.23), we obtain the correlation function between two synaptic weights separated by a vector x:

$$\overline{\langle J_0 J_x \rangle} = \frac{\int_0^\infty \mathrm{d}q\; q(1+q^2)^{3/2}[1+\nu_c(1+q^2)^{3/2}]^{-2} J_0(q|x|)}{\int_0^\infty \mathrm{d}q\; q(1+q^2)^{3/2}[1+\nu_c(1+q^2)^{3/2}]^{-2}} \tag{4.12}$$

where J_0 is the first-kind Bessel function of order zero. We show in figure 7 some typical curves given by (4.12).

Our analytical results applied to other types of spatial correlations may lead to the same 'Mexican-hat' profile. For instance, let us replace (4.11) by

$$\lambda(q) = \frac{1}{q^2 + 1/L^2} \tag{4.13}$$

where L is a cut-off length. The corresponding spatial correlations are given by

$$C_{ij} = C_L(|i-j|) = K_0(|i-j|/L) \tag{4.14}$$

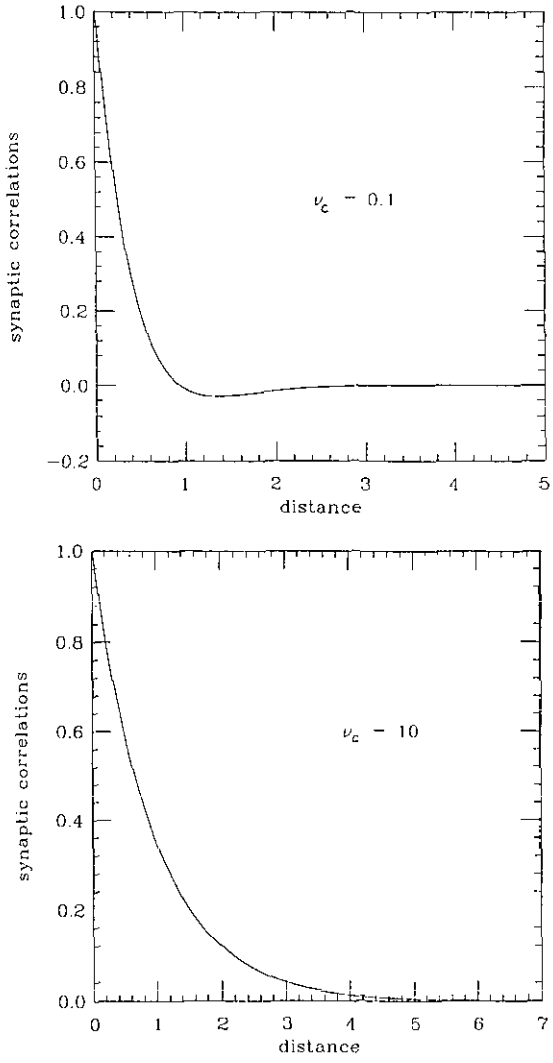

Figure 7. Bidimensional case: correlations between two couplings for two sizes α plotted as a function of the distance separating them. The corresponding values of ν_c are 0.1 and 10 (see (3.6)). The characteristic length of the spatial correlations is taken equal to one.

where $K_0(r)$ is a Hankel function decreasing as $r^{-1/2}e^{-r}$ at large distance r. For short distances, the eigenvalues (4.13) scale approximatively as $1/q^2$ which agrees with recent measures of correlations inside natural images [14, 15]. As before, we compute the synaptic correlation function using formula (3.23). We need to define the characteristic length L_s of these correlations which is related to L by

$$\frac{1}{L_s^2} = \frac{1}{L^2} + \frac{1}{\nu_c}. \tag{4.15}$$

We remark that $L_s < L$, which is similar to the unidimensional case where χ defined in (4.6) is lower than x: the synaptic correlations always decrease faster than the

input ones. Thus the correlations between the couplings are given by

$$\overline{\langle\langle J_0 J_x \rangle\rangle} = \left(C_{L_s} - \frac{1}{\nu_c^2} C_{L_s} * C_{L_s} \right)(|x|) \tag{4.16}$$

where the symbol $*$ denotes the convolution product. The right-hand side includes two terms. The first one is responsible for the positive centre of the synaptic profile, whereas the second one ensures that the flanks will be negative.

4.3. Discussion

Such a curve including negative surrounding flanks reminds us strikingly of experimental measures of synaptic sensitivities *'in vivo'*. Ganglion cells of mammals' retinas may indeed present this centre-on organization with lateral inhibition (see [13,15] and references therein). Recent studies based on information theory have recovered those synaptic profiles [15]. Two assumptions were made: first, neurons were linear units, computing a weighted average of their inputs. Secondly, the synaptic structure was derived following a redundancy-reduction principle in presence of intrinsic noise inside the input signal. The optimal couplings matrix depends on the ratio $\mathcal{L}$ signal/noise which is related to the incident luminosity. For large $\mathcal{L}$, the system essentially uses prediction techniques, estimating the activity of each input cell from its neighbours. On the contrary, when $\mathcal{L}$ is small, the processing neurons perform a smoothing stage aiming at reducing noise by means of the spatial correlations of the input signal.

We remark that there is a strong qualitative analogy between these synaptic profiles varying with $\mathcal{L}$ and the couplings correlations plotted on figures 2 and 7 as a function of the size of the training set α. It is all the more interesting that this feature apparently does not depend on the input space dimension and on the particular choice of (reasonable) spatial correlations.

From a mathematical standpoint, (3.23) is strictly equivalent to the so-called SPI hypothesis (smoothing-prediction interpolation [15]) but we must emphasize several differences concerning the conditions under which they have been derived. First, we have computed here correlations inside the synaptic couplings, whereas biological experiments focus on the synaptic weights themselves. Secondly, one must not forget that the problem we have concentrated upon is not related to some optimization principle. We are only interested here in storage properties of neural networks. Nevertheless, hetero-association enables us to consider this problem as a mapping of spatially organized inputs onto corresponding representations within a deeper processing layer. This formulation seems closer to the point of view of optimal encoding exposed in [15]. Last of all, neurons we have used in this paper are nonlinear and have binary outputs.

5. Conclusion

In this paper, we have focused on the capacity of a perceptron storing independent random patterns (ξ^μ, σ^μ) presenting internal correlations described by a matrix C. The generic term C_{ij} gives the correlation between the components i and j of ξ^μ. The matrix is symmetric, positive, and presents the property of rotational and translational invariance.

We have shown that the computation initiated by Gardner can be reproduced for any matrix C, and that the storage properties of the neural network depend only on its eigenvalues. In particular, the critical capacity at zero stability always equals $\alpha_c = 2$ regardless of the inner structure of the patterns. Numerical simulations have been performed with Gaussian and Ising patterns and are in excellent agreement with the theoretical predictions.

In order to obtain a geometrical interpretation of these analytical calculations, a second problem, conjugated to the initial one, has been introduced. It is indeed equivalent to considering self-correlated patterns stored by spherical synaptic vectors or patterns without internal correlations stored by synaptic vectors constrained to rely on an ellipsoid, the form of which depends only on the initial correlation matrix. Qualitative estimates are then possible for any correlation matrix.

Finally, the correlations between synaptic weights were investigated and general results valid for any matrix C derived. In the particular case of exponentially decreasing correlations inside one- or two-dimensional input patterns, the couplings profile looks like the so-called 'Mexican hat' distribution, including a excitatory centre and inhibitory flanks. Such an analogy with biological measures concerning centre-on cells in mammals' retina is outstanding but remains unexplained. We have also seen that a similar synaptic behaviour may be obtained from a simple modified Hebbian rule.

However, this paper deals only with hetero-associative mapping in feedfoward networks. It is of interest to see how the above properties are changed when we consider a fully connected neural network and try to store self-correlated patterns from an auto-associative standpoint. Preliminary results indicate that the critical capacity α_c is now an increasing function of the amount of spatial correlations and that the nature of couplings is deeply different since weights show ferromagnetic means. Detailed results will be reported elsewhere.

Acknowledgments

I am particularly grateful to I Kocher for her precious help concerning the analytical calculations and more generally this whole work. I thank M Mézard for numerous stimulating discussions and critical readings of the manuscript. I would also like to thank W Krauth and J P Nadal for useful remarks.

References

[1] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[2] Minsky M L and Papert S 1969 *Perceptrons* (Cambridge, MA: MIT Press)
[3] Mézard M and Paternello S 1989 *unpublished*
Barkai E, Hansel D and Kanter I 1990 *Phys. Rev. Lett.* **65** 2312
Barkai E and Kanter I 1990 *Europhys. Lett.* **14** 107
Mato G and Mourkazel C and N Parga 1991 *Bariloche Preprint*
Kocher I and Monasson R 1992 *J. Phys. A: Math. Gen.* **25** 367
[4] Krauth W and Mézard M 1989 *J. Physique* **50** 3057
Gutfreund H and Stein Y 1990 *J. Phys. A: Math. Gen.* **23** 2613
[5] Kanter I 1988 *Phys. Rev.* A **37** 2739
Nadal J P and Rau A 1991 *J. Physique* I **1** 1109
[6] Linsker R 1988 *Computer* **21** 105
Kay D J C Mac and Miller K D 1990 *Network* **1** 257

[7] Kepler T B and Abbott L F 1988 *J. Physique* **49** 1657
Forrest B M 1988 *J. Phys. A: Math. Gen.* **21** 245
Krauth W, Mézard M and Nadal J P 1988 *Complex Systems* **2** 387
[8] Cover T 1965 *IEEE Trans. Electron. Comput.* **14** 326
[9] Almeida J R de and Thouless D 1978 *J. Phys. A: Math. Gen.* **11** 983
[10] Krauth W and Mézard M 1987 *J. Phys. A: Math. Gen.* **20** L745
[11] Weisbuch G and Fogelman F 1985 *J. Physique Lett.* **46** L623
Gordon M B 1987 *J. Physique* **48** 2053
[12] Krauth W and Opper M 1989 *J. Phys. A: Math. Gen.* **22** L519
Derrida B, Griffiths R B and Prugel-Bennet A 1991 *J. Phys. A: Math. Gen.* **24** 4907
[13] Srinivasan M V, Laughlin S B and Dubs A 1982 *Proc. R. Soc.* B **216** 427
[14] Field D J 1987 *J. Opt. Soc. Am.* A 2379
[15] Atick J J and Redlich A N 1989 *Preprint* IASSNS-HEP-89/55; 1990 *Neural Comput.* **2** 308; 1990 *Preprint* IASSNS-HEP-90/51