Estimates of optimal storage conditions in neural network memories based on random matrix theory



1992 J. Phys. A: Math. Gen. 25 6251

(http://iopscience.iop.org/0305-4470/25/23/021)

J. Phys. A: Math. Gen. 25 (1992) 6251–6264. Printed in the UK

# Estimates of optimal storage conditions in neural network memories based on random matrix theory

W Tarkowski and M Lewenstein

Institute for Theoretical Physics, Polish Academy of Sciences, Al. Lotników 32/46, 02-668 Warsaw, Poland

Received 2 March 1992, in final form 29 June 1992

**Abstract.** We formulate a method for estimating the critical conditions for storage of sets of data in neural network memory. This variational method is based on random matrix theory and depends on calculating the average spectrum of a matrix, whose elements are given by overlaps of the stored patterns. Several generic cases of random overlap matrices are considered. We investigate the cases of simply uncorrelated random patterns, and 'spatially' and 'semantically' correlated ones. We obtain bounds of the critical curve in the control parameters space, which determine the stability of the stored data sets.

## 1. Introduction

One of the most important problems in the theory of attractor neural networks [1, 2, 3] is that of storage capacity. In the last years many works have been devoted to develop theories concerning optimal storage capacity of the perceptron and Hopfield-type [2] networks. The problem of optimal remembering of a set of random and statistically independent patterns was asked and solved by Gardner [4]. She formulated a so-called 'Gardner's program', within which one may obtain the critical conditions for storage of specified sets of patterns for arbitrary learning algorithms.

Originally, Gardner investigated independent and biased (but uncorrelated) patterns. In the case of purely independent and unbiased ones she obtained the well known result

$$\alpha_c = \left( \int_{-\kappa}^{+\infty} \mathrm{d}t \, \frac{\mathrm{e}^{-\frac{1}{2}t^2}}{\sqrt{2\pi}} (t+\kappa)^2 \right)^{-1} \tag{1}$$

where $\alpha_c$ is the optimal storage ratio [4] and $\kappa$ is the stability parameter. One may easily check that $\alpha_c(\kappa = 0) = 2$. For biased patterns maximal storage capacity may take larger values and in the case of sparse coding (e.g. when all units have the same values with the probability close to one) $\alpha_c$ tends to infinity for $\kappa \to 0$. Of course, one should stress that the remembering of information in network depends not only on the storage ratio but, generally, on the type of correlation of all considered patterns.

Gardner's method has been used by many authors to derive interesting results. It was used to calculate the optimal storage conditions in the presence of errors in recognition [5]. Many restrictions for the synaptic connection matrices $(J_{ij})$ were considered. Gardner's method was applied, for example, to binary couplings $(J_{ij} = \pm 1)$ [6], to other cost functions [7], and to multi-dimensional nets [8, 9].

0305-4470/92/236251+14$07.50  

We should add that Gardner's original idea was generalized to more realistic physical situations (i.e. the presence of noise in network dynamics [10, 11]). Gardner's program was investigated by Mezard in the frame of the so-called cavity method [12].

One should stress, however, that there are several open questions concerning Gardner's approach. It is very hard, for instance, to apply it to correlated patterns (when the average value of the correlation between different units is non-zero). This problem has been recently discussed by us [13] for the case of exponentially correlated patterns.

In Gardner's original paper the so-called replica method is used. This procedure, although very effective, is frequently criticized from purely mathematical point of view. It would, therefore, be of great importance to have some 'non-replica' methods of calculation and estimating the storage ratio as a function of parameters of the network. Such a variational method was proposed by Tarkowski *et al* [14]. We considered purely deterministic sets of patterns, which were invariant with respect to one-dimensional translations. The considered patterns had a 'pixel' shape. One of the most exciting results of our analysis concerned the shape of the stability curve (i.e. the dependence of the maximal stability parameter $\kappa$ versus the length of each 'pixel') for small storage ratios $\alpha$. The curve has quite an irregular, oscillating structure. This shape is an analytical reflection of the geometrical properties of the considered set of deterministic, highly ordered patterns.

In the present work we discuss the same method of estimating the critical conditions for storage of a set of random patterns in neural network memories. This method employs the properties of the matrix whose elements are overlaps between the patterns. For random ones, overlaps become stochastic variables, and thus our method employs random matrix theory [15]. The method is quite general and allows one to estimate storage conditions for various sets of patterns. This fact is of great importance, especially if exact results are hard to obtain. An important by-product of our paper is the presentation of more applications to the random matrix theory. We calculate average density of eigenvalues for a wide class of random overlap matrices. These results may be employed to study learning times for various learning algorithms [16–18].

The plan of the work is as follows: In section 2 we formulate our variational method of evaluating the fractional volume in the space of interactions [14]. We then apply it to stochastic sets of data and determine the lower and upper boundaries of $\alpha_c$. One obtains these bounds by investigation of a spectrum of the stochastic overlap matrix. Section 3 presents considerations concerning purely stochastic uncorrelated sets of data. Just to illustrate the method we derive bounds of Gardner's curve (1). In section 4 we investigate correlated patterns and obtain the limitations of the critical curve for such kinds of data. Two kinds of correlations are considered. Denoting the pattern by $\xi_i^\mu$, where $\mu$ enumerates the pattern and $i$ the site in the network, we introduce 'spatial' correlations when

$$\langle \xi_j^\mu \xi_{j'}^{\mu'} \rangle = \delta_{\mu\mu'} C(j, j') \tag{2}$$

and 'semantic' ones in the case of

$$\langle \xi_j^\mu \xi_{j'}^{\mu'} \rangle = C(\mu, \mu') \delta_{jj'}. \tag{3}$$

In the last part of this paper, the appendix, we present technical details of calculation of average spectrum of random overlap matrices. To this aim we use the

method of supersymmetry, and the Gaussian functional integration over commuting and anti-commuting (Grassman) variables [19, 20].

## 2. Fractional volume

In this section we present a method of evaluation of the fractional volume in the interaction space, which is defined as follows:

$$V_T = \frac{\prod_i [\int \prod_{j\neq i} \mathrm{d}J_{ij} \prod_\mu \Theta(\xi_i^\mu \sum_{j\neq i} \frac{J_{ij}}{\sqrt{N}} \xi_j^\mu - \kappa) \delta(\sum_{j\neq i} J_{ij}^2 - N)]}{\prod_i [\int \prod_{j\neq i} \mathrm{d}J_{ij} \delta(\sum_{j\neq i} J_{ij}^2 - N)]}. \tag{4}$$

Our approach does not make use of the replica method.

The fractional volume $V_T$ may be written in the form

$$V_T = \prod_{i=1}^{N} V_i = \exp\left\{ N \left( \frac{1}{N} \sum_i \ln V_i \right) \right\} \tag{5}$$

and

$$V_i = \frac{\int \prod_{j\neq i} \mathrm{d}J_{ij} \prod_\mu \Theta(\xi_i^\mu \sum_{j\neq i} \frac{J_{ij}}{\sqrt{N}} \xi_j^\mu - \kappa) \delta(\sum_{j\neq i} J_{ij}^2 - N)}{\int \prod_{j\neq i} \mathrm{d}J_{ij} \delta(\sum_{j\neq i} J_{ij}^2 - N)} =: \frac{\Phi_i}{\mathcal{N}_i}. \tag{6}$$

The normalization constant $\mathcal{N}$ is easy to obtain in the limit $N \to \infty$

$$\mathcal{N}_i = C e^{N(1+\ln(2\pi))/2} \tag{7}$$

where $C$ is a constant that for large $N$ behaves as $\ln(C/N) \to \infty$.

Usually, one proceeds by calculating the average $\langle \ln V_i \rangle$ with the help of the replica method [4]. Instead of performing the average, we integrate over $J_{ij}$, and rewrite the numerator of the expression (6) in the following form:

$$\Phi_i = \frac{1}{2\pi \mathrm{i}} \frac{1}{(2\pi)^{\alpha N}} \frac{(\sqrt{4\pi})^{\alpha N}}{(\det \hat{M}^i)^{1/2}} \int_C \mathrm{d}s \exp\left\{ N \left( s - \frac{1-\alpha}{2} \ln s + \frac{1}{2} \ln \pi \right) \right\}$$
$$\times \int_\kappa^\infty \prod_\mu \mathrm{d}\lambda_\mu \exp\left\{ -s \sum_{\mu,\mu'} \lambda_\mu (M^i)^{-1}_{\mu\mu'} \lambda_{\mu'} \right\} \tag{8}$$

where $C$ denotes the integration contour for $s$ going from $-\mathrm{i}\infty$ to $+\mathrm{i}\infty$. $(\hat{M}^i)^{-1}$ is the inverse of the overlap matrix $\hat{M}^i$, whose elements are given by

$$M^i_{\mu\mu'} = \frac{1}{N} \sum_{j\neq i} \xi_i^\mu \xi_j^\mu \xi_i^{\mu'} \xi_j^{\mu'}. \tag{9}$$

The dimension of the matrix $\hat{M}^i$ is $p = \alpha N$. Strictly speaking equation (8) is valid if and only if the matrix $\hat{M}^i$ has an inverse. Otherwise the integral over the $\lambda$s has to be restricted to the projection of the set of $\lambda \geqslant \kappa$ onto the subspace of $\lambda$s for which $(\hat{M}^i)^{-1}$ exists. The expression (8) as we shall see below, is the main ingredient of our approach. One should stress that the matrix $\hat{M}^i$ is in general $i$-dependent. Further proceeding consists of evaluating the expression (8) for large $N$ and $s$ limits,

and using the saddle point technique. It turns out that saddle point for $s$ is attained for real, positive $s$. Moreover, the value of $s$ at the saddle point tends to infinity, when $\alpha$ becomes critical. Only the minimum of the quadratic form in the exponent of the integrand of the expression (8) contributes to the integral over the $\lambda$s. Following [14] the final result reads

$$\left\langle \min_{\lambda_\mu \geqslant \kappa} \left( \sum_{\mu,\mu'} \lambda_\mu (M^i)^{-1}_{\mu\mu'} \lambda_{\mu'} \right) \right\rangle = N \tag{10}$$

where $\langle \cdot \rangle$ means the averaging over $\xi$s. The equation (10) determines the critical curve $\alpha_c(\kappa)$. Note that its left-hand side depends explicitly on $\kappa$ and implicitly on $\alpha$, through the dimension of $\hat{M}^i$. Obviously, the exact solution of this equation is very difficult. The minimum of the quadratic form in parenthesis has to be taken before averaging, and, in particular, depends on the realization of the random matrix $\hat{M}^i$. On the other hand, it is possible, however, to evaluate the minimum on the left-hand side of equation (10) for deterministic sets of patterns using a variational approach. The details of such applications of our method ware presented in [14].

The difficulties with exact accounting of (10) in the case of random patterns stimulated us to look for an approximate approach. Here we propose the approximate method of solving the equation (10), which is based on the theory of random matrices and allows us to determine upper and lower boundaries of the critical curve $\alpha_c$ as a function of the stability parameter $\kappa$. In order to do that one has to know the spectrum of the overlap matrix $\hat{M}^i$, defined in (10). If matrix $\hat{M}$ is bounded from above and below, we immediately obtain

$$\frac{\alpha\kappa^2 N}{\lambda_{\max}} \leqslant \left\langle \min_{\lambda_\mu \geqslant \kappa} \left( \sum_{\mu,\mu'} \lambda_\mu (M^i)^{-1}_{\mu\mu'} \lambda_{\mu'} \right) \right\rangle \leqslant \frac{\alpha\kappa^2 N}{\lambda_{\min}} \tag{11}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimal and the maximal value of the eigenvalues of the matrix $\hat{M}$, respectively. The upper and lower bounds of the critical curve have then the following form:

$$\kappa^2 = \lambda_{\max}/\alpha \qquad \kappa^2 = \lambda_{\min}/\alpha. \tag{12}$$

The advantage of this method is obvious: when evaluation of the exact critical curve is hard or impossible, one may easily obtain boundaries of this curve using the average density of the eigenvalue spectrum of the matrix $\hat{M}$. Of course, accounting of eigenvalues of the overlap matrix can also be, in particular cases, very difficult, but to this aim one may use standard methods, which were discovered and developed in the random matrix theory. In the next section we use such methods to solve a few interesting examples.

It is worth stressing that estimates of $\lambda_{\min}$ and $\lambda_{\max}$ from the eigenvalue spectrum are, strictly speaking, valid if and only if the eigenvalue spectrum is self-averaging. For standard random matrix ensembles it is true [15, 21]. The probability of finding an eigenvalue $\lambda$ outside of the interval $[\lambda_{\min}, \lambda_{\max}]$ is finite only then, when $|\lambda - \lambda_{\min}|$ or $|\lambda - \lambda_{\max}|$ are of the order of $O(N^{-1/6})$. We expect that a similar property holds for all of the examples of matrices considered in the following. All of these matrices are constructed for random unbiased patterns. For biased patterns correlation matrices have typically one additional eigenvalue of the order of $N$ [18]. This fact may lead to a significant modification of the upper bound for $\kappa$.

### 3. Spectrum of the overlap matrix for uncorrelated patterns

Let us investigate optimal storage conditions for the simplest case of unbiased and uncorrelated random patterns with the probability distribution

$$\Pr(\xi_i^\mu = \pm 1) = \tfrac{1}{2} \tag{13}$$

for all $\mu$ and $i$. Such a case of purely stochastic patterns was investigated by E Gardner [4]. Our treatment of this problem depends on of calculating the average eigenvalue density of the matrix $\hat{M}$. This is done by using the so-called supersymmetry method [20]. Leaving all technical details (which are contained in the appendix) we describe below the method of calculation of the average eigenvalue density, which becomes exact in the limit of large $N$.

In all of the cases considered the supersymmetry approach leads to the conclusion that the logarithm of the determinant of the matrix $\hat{M}$ is a self-averaging quantity (see appendix). In other words

$$\langle \det{}^{-1/2}(\hat{1}(\lambda - i\varepsilon) - \hat{M}) \rangle \cong \exp\left\{ -\frac{p}{2} \int d\lambda' \varrho(\lambda') \ln(\lambda - i\varepsilon - \lambda') \right\} \tag{14}$$

where $\varepsilon$ is a small positive real number (see [22]) and $\varrho(\lambda)$ is the average eigenvalue spectrum of the matrix $\hat{M}$

$$\varrho(\lambda) = \left\langle \frac{1}{p} \sum_{j=1}^{p} \delta(\lambda - \lambda_j) \right\rangle. \tag{15}$$

In this expression $\{\lambda_j\}$, $j = 1, \ldots, p$ denote the set of eigenvalues of the matrix $\hat{M}$.

If so, then we can calculate $\varrho(\lambda)$ as well as $\lambda_{\min}$ and $\lambda_{\max}$, simply by averaging (14). We introduce

$$\begin{aligned} Z(\lambda) &= \left\langle \int \mathcal{D}\left( \frac{x_\mu}{\sqrt{-\pi i}} \right) \exp\left\{ -i \sum_{\mu,\mu'} x_\mu ((\lambda - i\varepsilon)\delta_{\mu\mu'} - M_{\mu\mu'}) x_{\mu'} \right\} \right\rangle \\ &\equiv \langle \det{}^{-1/2}(\hat{1}(\lambda - i\varepsilon) - \hat{M}) \rangle \end{aligned} \tag{16}$$

where $\hat{M}$ is the overlap matrix defined in (9). Note that by changing the integration variables $x_\mu \to \xi_i^\mu x_\mu$ the matrix $\hat{M}$ from (9) can be substituted by

$$M_{\mu\mu'} = \frac{1}{N} \sum_{j \neq i} \xi_j^\mu \xi_j^{\mu'}. \tag{17}$$

From (14) (see also [22]) it follows that

$$\varrho(\lambda) = -\frac{2}{\alpha \pi N} \operatorname{Im} \frac{\partial}{\partial \lambda} \ln Z(\lambda). \tag{18}$$

On the other hand, it is easy to observe that

$$\begin{aligned} Z(\lambda) &= \left\langle \int \mathcal{D}x_\mu \mathcal{D}y_j \exp\{ i\lambda \sum_\mu x_\mu^2 + \frac{i}{4} \sum_j y_j^2 + \frac{i}{\sqrt{N}} \sum_{\mu,j} \xi_i^\mu \xi_j^\mu x_\mu y_j \} \right\rangle_\xi \\ &= \int \mathcal{D}x_\mu \mathcal{D}y_j \exp\left\{ i\lambda \sum_\mu x_\mu^2 + \frac{i}{4} \sum_j y_j^2 - \frac{i}{2N} \sum_{\mu,j} x_\mu^2 y_j^2 \right\}. \end{aligned} \tag{19}$$

The constant value in front of the all integrals, which expresses the quantity $Z(\lambda)$, is negligible because the ln $Z$ enters equation (18).

Introducing macro-variable $s = \sum_\mu x_\mu^2/N$ and its conjugate counterpart $\hat{s}$, the $Z(\lambda)$ reduces to

$$Z(\lambda) = \int ds\, d\hat{s}\, \exp\left\{-\frac{\alpha N}{2}\ln(\hat{s} - i\lambda) + Ns\hat{s} - \frac{N}{2}\ln\left(\frac{s}{2} - \frac{i}{4}\right)\right\}$$
$$=: \int ds\, d\hat{s}\, e^{-N\mathcal{F}(s,\hat{s})}. \tag{20}$$

The above integral is evaluated then by the saddle-point method. The final result has the following form (see also [16–18]):

$$\varrho(\lambda) = \frac{\sqrt{4\alpha - (1+\alpha-\lambda)^2}}{2\pi\alpha\lambda} \tag{21}$$

for $\lambda \in [(1-\sqrt{\alpha})^2, (1+\sqrt{\alpha})^2]$ where $\alpha < 1$, and

$$\varrho(\lambda) = \frac{\sqrt{4\alpha - (1+\alpha-\lambda)^2}}{2\pi\alpha^2\lambda} + \frac{\alpha-1}{\alpha}\delta(\lambda) \tag{22}$$

for $\lambda \in [(1-\sqrt{\alpha})^2, (1+\sqrt{\alpha})^2] \cup \{0\}$ while $\alpha \geqslant 1$. For other $\lambda$, the density $\varrho(\lambda) = 0$. As we can see, the density of the eigenvalues spectrum has quite a regular shape and takes the value zero on its extreme points. The presence of the delta function is quite easy to explain. The inverse of the matrix $\hat{M}$ may be written as a sum of $N$ projector operators in the $\alpha N$-dimensional space. When the dimension $p = \alpha N$ of the overlap matrix $\hat{M}$ becomes larger than $N$ some of the eigenvalues (exactly: $(\alpha - 1)N$) take the value zero. Now one may write investigated bounds for the critical curve. In this case

$$\lambda_{max} = (1+\sqrt{\alpha})^2 \tag{23}$$
$$\lambda_{min} = (1-\sqrt{\alpha})^2 \tag{24}$$

and the upper and lower boundaries respectively read

$$\kappa^2 = \frac{(1+\sqrt{\alpha})^2}{\alpha} \tag{25}$$
$$\kappa^2 = \frac{(1-\sqrt{\alpha})^2}{\alpha}. \tag{26}$$

Of course, we should take into consideration the $\delta$ function in the formula (22). That limits the validity of the lower bounds to the region of $0 < \alpha \leqslant 1$.

The results obtained in this section are plotted in figures 1 and 2 to allow comparison with the other results. The long-dashed lines in both figures are Gardner's exact curve, while the solid lines are the limitations obtained from equations (26). One should stress that the differences between all three lines for large enough $\kappa$ are very small. For small $\kappa$, the precision of our estimate is worse. Obviously, the lower bound has greater importance than the upper one. In the statistical sense (with probability one), the area below the lower line is the stability area. That means that for each point in the described $(\alpha, \kappa)$ area the conditions for storage of a set of $p = \alpha N$ random patterns with stability $\kappa$ are fulfilled. On the other hand, above the upper line, all points are unstabilized.

Summarizing this section we stress that results obtained by using the variational approach differ not much from the exact one at least for large $\kappa$. The relevance of our approximate method increases when it is impossible to obtain the exact result.

## 4. Estimations of the storage conditions for correlated sets of stochastic patterns

In this section we apply the variational method to calculate bounds of the critical curve for several specified sets of patterns. The exact expressions for the critical curves in this case are rather hard to obtain. In these cases our approximate method plays an essential role in understanding the mechanism of remembering and storing the sets of data. We stress that in principle the variational limitations can be obtained for arbitrary set of patterns. Of course, the exactness of these bounds can differ for each considered case.

As the first example of using of the variational approach we investigate the set of 'semantically' correlated patterns defined as follows:

$$\langle \xi_j^\mu \rangle = 0 \tag{27}$$

while

$$\langle \xi_j^\mu \xi_{j'}^{\mu'} \rangle = \delta_{jj'} C(\mu, \mu') \tag{28}$$

for all $j = 1, \ldots, N$ and $\mu = 1, \ldots, \alpha N$. Obviously

$$C(0) = 1. \tag{29}$$

The above expressions mean that the patterns are unbiased, i.e. the average over single unit for each pattern is zero ($\xi_i^\mu$ are statistically independent for different $i$). As a quite interesting example we can take the exponential shape of the function $C(|\mu - \mu'|) = \exp[-(1/L_c)|\mu - \mu'|]$. Random patterns with exponential correlations of the kind (28) are quite generic, and can be easily generated. For a given $i$ they correspond to thermal equilibrium states of the one-dimensional Ising model with the Hamiltonian $H_i = -\sum_\mu \xi_i^\mu \xi_i^{\mu+1}$. The temperature is related then to the correlation length $L_c$ through

$$L_c = -(\ln \tanh(1/T))^{-1}. \tag{30}$$

The special case of exponentially correlated patterns has recently been investigated by us [13]. For this particular case, we have been able to show that the storage capacity $\alpha$ tends to infinity as $L_c \to \infty$. This is a generalization of the classical Willshaw results for sparsely coded patterns ([23], see also [4]). We stress, however, that this exact result could only be obtained for the exponential form of $C(\mu, \mu')$.

Here we shall consider the general form of $C(\mu, \mu')$. In order to do this we again use the determinant method. After calculations similar to those in section 3 expression (16) becomes

$$Z(\lambda) = \int \mathcal{D}x_\mu \, \mathcal{D}y_j \, \exp\left\{ i\lambda \sum_\mu x_\mu^2 + \frac{i}{4} \sum_j y_j^2 - \frac{1}{2N} \sum_{\mu,\mu',j} x_\mu C_{\mu\mu'} x_{\mu'} y_j^2 \right\}. \tag{31}$$

Let $s = \sum_j y_j^2/N$, $\hat{s}$-conjugated variable and $C_k$ denotes eigenvalues of the correlation matrix $C(\mu, \mu')$. For $C(\mu, \mu') = C(|\mu - \mu'|)$ eigenvalues are given by the Fourier transform

$$C_k = \sum_{\mu=0}^{\alpha N - 1} C(\mu) e^{i\omega_k \mu} \tag{32}$$

where $\omega_k = 2\pi k/p$ and $p = \alpha N$.

In the limit of large $N$ and $p$ the quantity $Z(\lambda)$ may again be evaluated, using the saddle-point method. To this end we have to find a minimum of the 'free energy'

$$\mathcal{F} = s\hat{s} + \tfrac{1}{2}\ln\left(\hat{s} - \tfrac{1}{4}\right) + \tfrac{\alpha}{2}\left\langle \ln\left(\tfrac{s}{2}C - \lambda\right)\right\rangle_{\varrho(C)} \tag{33}$$

where $\langle\cdot\rangle_{\varrho(C)}$ denotes the average value over the spectrum of the correlation matrix

$$\langle f(C)\rangle = \int_{C_{\min}}^{C_{\max}} f(C')\varrho(C')\,\mathrm{d}C'. \tag{34}$$

For $C(\mu,\mu') = C(|\mu - \mu'|)$ we have

$$\varrho(C) = \frac{1}{|C(\omega)|} \tag{35}$$

for $\omega$ such that $C(\omega) = C$. Note that the requirement $C(0) = 1$ implies that

$$\int_{C_{\min}}^{C_{\max}} C'\varrho(C')\,\mathrm{d}C' = 1. \tag{36}$$

Using the saddle-point method to evaluate the integral (31) one may easily obtain

$$\frac{s}{2} + 2\alpha\lambda\left\langle \frac{1}{sC - 2\lambda}\right\rangle_{\varrho(C)} = 1 - \alpha \tag{37}$$

where $\langle\cdot\rangle_{\varrho(C)}$ is the average over the distribution of $C$ considered.

In this paper we find the bounds of the critical curve in the two generic cases: when the function $\varrho(C)$ is a uniform and semicircular distribution. From this point on we omit somewhat complicated, but rather elementary calculations and show only our results.

First we consider a uniform distribution of $C$

$$\varrho(C) = \frac{1}{C_{\max} - C_{\min}} \tag{38}$$

for $C_{\min} \leqslant C \leqslant C_{\max}$. Note that the condition (36) results in the equality $C_{\min} + C_{\max} = 2$. Secondly we consider the case where the distribution of $C$ has the semicircular form

$$\varrho(\lambda) = \frac{2}{\pi r^2}\sqrt{2r(C - a) - (C - a)^2} \tag{39}$$

for $a - r \leqslant C \leqslant a + r$, where $r$ is the radius of the circle and $a$ is the distance from the centre of it to the point $C = 0$. The constraint (36) implies that $r + a = 1$.

The accounting for the limits $\lambda_{\min}$, $\lambda_{\max}$ of the eigenvalue density consists in eliminating from the (37) the variable $s$. It can easily be seen that $\lambda_{\min}$, ($\lambda_{\max}$) are attained when a real solution of (37) ceases to exist. Simple graphical analysis of (37) shows that this happens at the point when the derivative of the left-hand side of (37) with respect to $s$ is zero

$$\frac{1}{2} - 2\alpha\lambda\left\langle \frac{C}{(sC - 2\lambda)^2}\right\rangle_{\varrho(C)} = 0. \tag{40}$$

Equation (40) means that at this point the real solution of (37) for $s$ becomes doubly degenerated and bifurcates into two complex ones. The two equations (37) and (40) have two unique real solutions $s(\alpha, \kappa)$ and $\lambda(\alpha, \kappa)$. The values of $\lambda(\alpha, \kappa)$ obtained in such a way are $\lambda_{\min}$ and $\lambda_{\max}$. In view of the complicated shape of the solution of (37) and (40) we solve them using a numerical method. The results (i.e. the dependence of the stability parameter $\kappa$ on the storage ratio $\alpha$) are plotted in figure 1. The main difference between the solutions obtained depends on the character of the edges of the distribution $\varrho(C)$. In the case of uniform distribution (38) the eigenvalue density $\varrho(C)$ has jumps at the edges of its support. For the semicircular law the eigenvalue density takes the value zero on each border of its carrier. One should stress that resulting bounds are practically identical for the two cases considered of distribution of $C$ with the same width, so we present only the results for uniform distribution $\varrho(C)$ in figure 1.

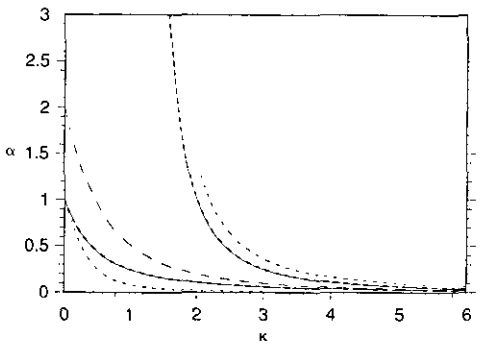


**Figure 1.** Estimates of the critical capacity $\alpha$ as a function of $\kappa$ for various types of stored patterns. The long-dashed line corresponds to Gardner's exact result for purely random patterns, whereas the solid lines correspond to our estimates for purely random ones. The thick dashed line presents estimates for 'semantically' correlated patterns with $C_{\max} - C_{\min} = 0.2$, whereas the short dashed line presents those for the same kind of patterns with $C_{\max} - C_{\min} = 1.8$.

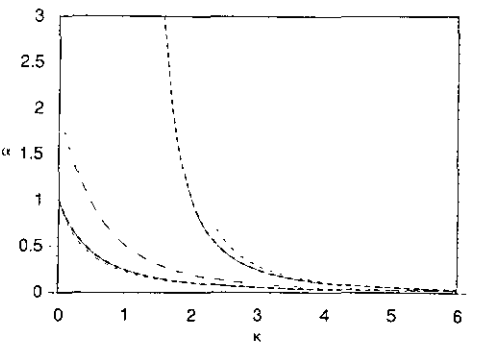


**Figure 2.** As figure 1, except that the thick dashed line now presents estimates for 'spatially' correlated patterns with $C_{\max} - C_{\min} = 0.2$, whereas the short dashed line presents those with $C_{\max} - C_{\min} = 1.8$.

Similarly we can calculate $\lambda_{\min}$, $\lambda_{\max}$ and $\varrho(\lambda)$ in the case of patterns correlated in a 'spatial' fashion. In this case $\xi_i^\mu$ are statistically independent for different $\mu$, $\mu'$, but

$$\langle \xi_j^\mu \rangle = 0 \tag{41}$$

whereas

$$\langle \xi_j^\mu \xi_{j'}^{\mu'} \rangle = \delta_{\mu\mu'} C(|j - j'|). \tag{42}$$

As before, we introduce macro-variables $s = \sum_\mu x_\mu^2 / N$, its conjugate counterpart $\hat{s}$ and after calculations similar to those in section 3 we obtain the 'free' energy in the form

$$\mathcal{F} = s\hat{s} + \tfrac{\alpha}{2} \ln(\hat{s} - \lambda) + \tfrac{1}{2} \left\langle \ln\left(\tfrac{s}{2} C - \tfrac{1}{4}\right) \right\rangle_{\varrho(C)} \tag{43}$$

where $\langle\cdot\rangle_{\varrho(C)}$ is the average over the spectrum of the correlation matrix $C(i-j)$ (cf expression (33), the only difference being that the spectral transform is 'spatial' rather than 'semantic'). The quantities $\lambda_{\min}$, $\lambda_{\max}$ are determined as before from the equation

$$2\lambda s + \left\langle \frac{1}{2sC-1} \right\rangle_{\varrho(C)} = \alpha - 1 \tag{44}$$

and its derivative

$$\lambda - \left\langle \frac{C}{(2sC-1)^2} \right\rangle_{\varrho(C)} = 0. \tag{45}$$

After performing the numerical evaluation we obtain the two limit values of $\lambda$. The results are plotted in figure 2 for the case of uniform distribution $\varrho(C)$. In both figures 1 and 2 the solid curves give estimations in the absence of any correlations. The thick dashed curves correspond to correlation width $C_{\max} - C_{\min} = 0.2$. Note that for such small widths correlations do not lead to significant modifications of the estimates. The bounds change for $(C_{\max} - C_{\min})$ of the order of 1 (see narrow dashed curves).

In summary, we stress the similarities and differences between the two forms of correlations of patterns. First of all the results for both uniform and semicircular distributions of $C$ (of the same width) are practically identical. This means that the shape of the spectrum of the correlation matrix does not play an important role in the limiting bounds. On the other hand for 'spatially' correlated patterns the bounds of the critical curves depend rather weakly on the width of the distribution $\varrho(C)$, and do not differ very much for the result for uncorrelated patterns. This suggests that the exact results for 'spatially' correlated patterns may also lie close to Gardner's original curve (1). The situation is quite different for 'semantically' correlated data. Here the dependence on the width of the distribution $\varrho(C)$ is rather strong. We may expect a similar strong dependence with the exact results. Indeed, analysis of exponentially correlated patterns shows that for 'semantically' correlated data the exact curve lies far above Gardner's one (1) when the correlation length is large [13].

## 5. Conclusions

In this paper we present a generic method of evaluating the critical conditions for the storage of information in neural networks. This variational method employs the analysis of the spectrum of overlap matrices and allows one to find bounds of the critical curve which determine the phases of existence and vanishing of stability of the stored sets of data. One should stress that with the help of our method these bounds may be obtained, practically, for arbitrary patterns, for any particular form of them, and their correlations, etc. In this paper several examples of such bounds have been considered. We studied a wide class of overlap matrices corresponding to various types of patterns. They are generic for many different physical situations. For example 'semantically' correlated data plays a big role in psychology and the theory of information processing. In fact, in the recent neurobiological experiments of Miyashita [24, 25] semantically correlated attractors were observed. Patterns in those experiments were learned in a specific sequence. Attractors that corresponded to

subsequent patterns exhibited correlations that decayed with the time delay between learning periods. 'Spatially' correlated patterns are used in optical information theory. One should also stress that our results may be directly used for evaluating the learning time of various learning algorithms [16–18].

Our method has many advantages, but is limited to the case of correlations that are bounded from below by positive $\lambda_{\min}$. In principle, however, it may be generalized to the case of $\lambda_{\min} = 0$. The minimum of the quadratic form in expression (8) should then be restricted to the set of $\lambda_\mu$s that are orthogonal to the kernel of the correlation matrix (i.e. a set of those $\lambda_\mu$s that are annihilated by the correlation matrix).

Finally we would like to mention the overlap matrices considered by us have the same properties as the ones that are sums of random projection operators. Such matrices have been recently discussed for applications in nuclear physics and quantum chaos [26].

**Acknowledgment**

The work reported in this paper was financed by the KBN under grant No 2-0207-91-01.

**Appendix. The supersymmetry method**

In this appendix we calculate in detail the eigenvalue spectrum of the matrix $\hat{M}$ defined in section 2 (see expression (9)). The method is easily generalized to all the other considered overlap matrices [19]. As the first step we define the average spectrum $\varrho(\lambda)$ of the matrix $\hat{M}$ in the same way as in section 3 (see expression (15)). Then one may write

$$\varrho(\lambda) = -\frac{1}{2\pi} \operatorname{Im} \frac{\partial}{\partial J} Z(J)|_{J=0} \tag{46}$$

where the generating function $Z(J)$ is defined as follows:

$$\begin{aligned} Z(J) \equiv \left\langle \int \mathcal{D}x_\mu \, \mathcal{D}dy_\mu \, \mathcal{D}\phi_\mu^* \, \mathcal{D}\phi_\mu \, \exp\Bigg[ 2\mathrm{i} \sum_{\mu,\mu'} x_\mu((\lambda - \mathrm{i}\varepsilon - J)\delta_{\mu\mu'} - M_{\mu\mu'})x_{\mu'} \right. \\ + 2\mathrm{i} \sum_{\mu,\mu'} y_\mu((\lambda - \mathrm{i}\varepsilon - J)\delta_{\mu\mu'} - M_{\mu\mu'})y_{\mu'} \\ \left. + 2\mathrm{i} \sum_{\mu,\mu'} \phi_\mu^*((\lambda - \mathrm{i}\varepsilon + J)\delta_{\mu\mu'} - M_{\mu\mu'})\phi_\mu \Bigg] \right\rangle \end{aligned} \tag{47}$$

where the matrix elements $M_{\mu,\mu'}$ are given by expression (17). In the above integral $x_\mu$ and $y_\mu$ are standard commuting variables, whereas $\phi_\mu$, $\phi_\mu^*$ are Grassman anti-commuting ones

$$\phi_\mu \phi_{\mu'} + \phi_{\mu'} \phi_\mu = 0 \tag{48}$$

$$\phi_\mu \phi_{\mu'}^* + \phi_{\mu'}^* \phi_\mu = 0. \tag{49}$$

The normalization of the Grassman variables is such that

$$\int \mathrm{d}\phi = 0 = \int \mathrm{d}\phi^* \tag{50}$$

$$\int \phi \,\mathrm{d}\phi = \frac{1}{\sqrt{2\pi}} = \int \phi^* \,\mathrm{d}\phi^* \tag{51}$$

so

$$\int \phi\phi^* \,\mathrm{d}\phi \,\mathrm{d}\phi^* = -\frac{1}{2\pi}. \tag{52}$$

One then defines the so-called graded matrix

$$\hat{S} = \begin{pmatrix} \hat{A} & \hat{\Psi} \\ \hat{\Phi} & \hat{B} \end{pmatrix} \tag{53}$$

where $\hat{A}$, $\hat{B}$, $\hat{\Phi}$ and $\hat{\Psi}$ are $N \times N$, $M \times M$, $M \times N$ and $N \times M$-dimensional matrices, respectively. $\hat{A}$, $\hat{B}$, have commuting matrix elements, whereas $\hat{\Phi}$ and $\hat{\Psi}$ haveanti-commuting ones. We may introduce now the symbol $\det_g$, which denotes the supersymmetry counterpart of a determinant, and for such graded matrices is defined as follows:

$$\det{}_g \hat{S} = \det(\hat{A} - \hat{\Psi}(\hat{B})^{-1}\hat{\Phi})(\det \hat{B})^{-1}. \tag{54}$$

After introducing the new 'bosonic' (commuting) variables $a_j$, $b_j$ and 'fermionic' (anti-commuting) ones $\alpha$, $\alpha^*$ the integral from expression (47) takes the form

$$\begin{aligned} Z(J) = \Big\langle \int \mathcal{D}x_\mu \,\mathcal{D}y_\mu \,\mathcal{D}\phi_\mu^* \,\mathcal{D}\phi_\mu \,\mathcal{D}a_j \,\mathcal{D}b_j \,\mathcal{D}\alpha_j^* \,\mathcal{D}\alpha_j \,\exp\Big\{ \mathrm{i}\sum_j a_j^2 + \mathrm{i}\sum_j b_j^2 \\ + 2\mathrm{i}\sum_\mu x_\mu^2(\lambda - \mathrm{i}\varepsilon - J) + 2\mathrm{i}\sum_\mu y_\mu^2(\lambda - \mathrm{i}\varepsilon - J) + 2\mathrm{i}\sum_\mu \phi_\mu^*\phi_\mu(\lambda - \mathrm{i}\varepsilon + J) \\ + 2\mathrm{i}\sqrt{\frac{2}{N}} \sum_{\mu,\mu',j} \xi_i^\mu \xi_j^\mu (a_j x_\mu + b_j x_\mu + \tfrac{1}{2}\phi_\mu \alpha_j + \tfrac{1}{2}\phi_\mu^* \alpha_j^*) \Big\} \Big\rangle. \end{aligned} \tag{55}$$

Now we can easily do the average over $\xi$s. Then we have to introduce the new macro-variables to disentangle the mixed terms in the resulting expression

$$\begin{aligned} X &= \frac{1}{N}\sum_\mu x_\mu^2 & Y &= \frac{1}{N}\sum_\mu y_\mu^2 \\ R &= \frac{1}{N}\sum_\mu x_\mu y_\mu & V &= \frac{1}{N}\sum_\mu \phi_\mu^*\phi_\mu \\ \Sigma &= \frac{1}{N}\sum_\mu x_\mu \phi_\mu & \Sigma^* &= \frac{1}{N}\sum_\mu x_\mu \phi_\mu^* \\ \Delta &= \frac{1}{N}\sum_\mu y_\mu \phi_\mu & \Delta^* &= \frac{1}{N}\sum_\mu y_\mu \phi_\mu^*. \end{aligned} \tag{56}$$

Variables $X$, $Y$, $R$ and $V$ are commuting, whereas $\Sigma$, $\Sigma^*$, $\Delta$ and $\Delta^*$ are anti-commuting. We also introduce conjugate counterparts for the eight above-defined

variables $\hat{X}$, $\hat{Y}$, $\hat{R}$, $\hat{V}$, $\hat{\Sigma}$, $\hat{\Sigma}^*$, $\hat{\Delta}$ and $\hat{\Delta}^*$. Integrals over the local variables $x_\mu$, $y_\mu$, $a_j$, $b_j$, $\alpha_j$, $\alpha_j^*$, $\phi_\mu$ and $\phi_\mu^*$ then become Gaussian and can be performed simply. The resulting formula is

$$\begin{aligned} Z(J) = \int \mathrm{d}X\,\mathrm{d}\hat{X}\,\mathrm{d}Y\,\mathrm{d}\hat{Y}\,\mathrm{d}R\,\mathrm{d}\hat{R}\,\mathrm{d}V\,\mathrm{d}\hat{V}\,\mathrm{d}\Sigma\,\mathrm{d}\hat{\Sigma}\,\mathrm{d}\Sigma^*\,\mathrm{d}\hat{\Sigma}^*\,\mathrm{d}\Delta\,\mathrm{d}\hat{\Delta}\,\mathrm{d}\Delta^*\,\mathrm{d}\hat{\Delta}^* \\ \times \exp\{N X\hat{X} + NY\hat{Y} + NR\hat{R} + NV\hat{V} + N\Sigma\hat{\Sigma} - N\Sigma^*\hat{\Sigma}^* \\ + N\Delta\hat{\Delta} - N\Delta^*\hat{\Delta}^* - \tfrac{1}{2}\ln\det{}_g\mathbf{Q}(X,Y,R,V,\Sigma,\Sigma^*,\Delta,\Delta^*) \\ - \tfrac{1}{2}\alpha\ln\det{}_g\mathbf{P}(\hat{X},\hat{Y},\hat{R},\hat{V},\hat{\Sigma},\hat{\Sigma}^*,\hat{\Delta},\hat{\Delta}^*)\}. \end{aligned} \tag{57}$$

The graded matrices $\mathbf{Q}$, $\mathbf{P}$ that enter expression (57) have the following form

$$\mathbf{Q} = \begin{pmatrix} \mathrm{i} - 4X & -4R & -4\Sigma \\ -4R & \mathrm{i} - 4Y & -4\Delta \\ -4\Sigma^* & -4\Delta^* & -\mathrm{i} + 2V \end{pmatrix} \tag{58}$$

$$\mathbf{P} = \begin{pmatrix} 2\mathrm{i}(\lambda - J) - \hat{X} & -\hat{R}/2 & -\hat{\Sigma} \\ -\hat{R}/2 & 2\mathrm{i}(\lambda - J) - \hat{Y} & -\hat{\Delta} \\ -\hat{\Sigma}^* & -\hat{\Delta}^* & 2\mathrm{i}(\lambda + J) - \hat{V} \end{pmatrix}. \tag{59}$$

The integral in the expression (57) may be evaluated using the saddle-point method. Generally the saddle-point equations give

$$\begin{aligned} R = 0 = \hat{R}. \\ \Sigma = \hat{\Sigma} = 0 = \hat{\Delta} = \Delta \end{aligned} \tag{60}$$

The 'bosonic' and 'fermionic' parts of graded matrices thus decorrelate and the final result is the same as (21) and (22). It indicates that indeed 'log det' is a self-averaging quantity. One may easily check by investigating the eigenvalues of the graded matrices $\mathbf{Q}$, $\mathbf{P}$ that the solution (60) is locally stable.

One may add that the above results can be also obtained within the replica method [27, 16–18, 22].

*Note added in proof.* The problem of storage of correlated patterns in neural network memory has been recently solved by Lewenstein and Tarkowski ('spatial' and 'semantical' correlations) [13, 28] and by Monasson ('spatial' correlations) [29].

## References


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

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

[3] Little W A 1974 The existence of persistent states in the brain *Math. Biosci.* **19** 101

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

[5] Gardner E and Derrida B 1988 Optimal storage properties of neural network models *J. Phys. A: Math. Gen.* **21** 271–84

[6] Krauth W and Opper M 1989 Critical storage capacity of the $J = \pm 1$ neural networks *J. Phys. A: Math. Gen.* **22** L519

[7] Griniasty M and Gutfreund H 1991 Learning and retrieval in attractor neural networks above saturation *J. Phys. A: Math. Gen.* **24** 715–34

[8] Griniasty M 1991 *Proceedings of the Workshop on Neural Networks: from Biology to High Energy Physics (Elba, 1991)*

[9] Parga N 1991 *Proceedings of the Workshop on Neural Networks: from Biology to High Energy Physics (Elba, 1991)*

[10] Amit D J, Evans M R, Horner H and Wong K Y M 1990 Retrieval phase diagrams for attractor neural networks with optimal interactions *J. Phys. A: Math. Gen.* **23** 3361–81

[11] Wong K Y M and Sherrington D 1990 Optimally adapted attractor neural networks in the presence of noise *J. Phys. A: Math. Gen.* **23** 4659

[12] Mezard M 1989 The space of interactions in neural networks: Gardner's computation with the cavity method *J. Phys. A: Math. Gen.* **22** 2181

[13] Lewenstein M and Tarkowski W 1992 Optimal storage of correlated patterns in neural network memories *Phys. Rev.* **46** 2139–42

[14] Tarkowski W, Komarnicki M and Lewenstein M 1991 Optimal storage of invariant sets of patterns in neural network memories *J. Phys. A: Math. Gen.* **24** 4197–217

[15] Mehta M L 1967 *Random Matrices and Statistical Theory of Energy Levels* (New York: Academic)

[16] Opper M 1989 Learning in neural networks: solvable dynamics *Europhys. Lett.* **8** 389

[17] Kinzel W and Opper M 1991 *Dynamics of Learning in Physics of Neural Networks* ed J L van Hemmen, E Domany and K Schulten (Berlin: Springer)

[18] Le Cun Y, Kanter I and Solla S A 1991 Eigenvalues of covariance matrices: application to neural-network learning *Phys. Rev. Lett.* **66** 2396–9

[19] Guhr T 1991 Dyson's correlation functions and graded symmetry *J. Math. Phys.* **32** 336–47

[20] Kuś M, Lewenstein M and Haake F 1991 Density of eigenvalues of random band matrices *Phys. Rev.* A **44** 2800–8

[21] Kosterlitz J M, Thouless D J and Jones R C 1976 Spherical model of a spin-glass *Phys. Rev. Lett.* **36** 1217–20

[22] Edwards S F and Jones R C 1976 The eigenvalue spectrum of a large symmetric random matrix *J. Phys. A: Math. Gen.* **9** 1595–1603

[23] Willshaw D J, Buneman O P and Longuet-Higgins H C 1969 Non-holographic associative memory *Nature* **222** 960

[24] Miyashita Y 1988 Neuronal correlate of visual associative long-term memory in the primate temporal cortex *Nature* **335** 817–9

[25] Griniasty M, Tsodyks M V and Amit D J 1992 Conversion of temporal correlations between stimuli to spatial correlations between attractors *Universitá di Roma* **856** Preprint

[26] Haake F, Izrailev F M, Lehman N, Saher D and Sommers H J 1991 Level density of random matrices for decaying systems *Preprint 91-98* Budker Institute of Nuclear Physics, Novosibirsk, Russia

[27] Crisanti A and Sompolinsky H 1987 Dynamics of spin systems with randomly asymmetric bonds: Langevin dynamics and a spherical model *Phys. Rev.* A **36** 4922

[28] Tarkowski W and Lewenstein M 1992 Storage of sets of correlated data in neural network memories *J. Phys. A: Math. Gen.* submitted

[29] Monasson R 1992 Properties of neural networks storing spatially correlated patterns *J. Phys. A: Math. Gen.* **25** 3701–20