Storage capacity of correlated perceptrons

D. Malzahn,[1] A. Engel,[1] and I. Kanter[2]
[1]*Institut für Theoretische Physik, Otto-von-Guericke-Universität, Universitätsplatz 2, Postfach 4120, D-39106 Magdeburg, Federal Republic of Germany*
[2]*Department of Physics, Bar Ilan University, Ramat Gan, 52100, Israel*
(Received 30 October 1996)

We consider an ensemble of K single-layer perceptrons exposed to random inputs and investigate the conditions under which the couplings of these perceptrons can be chosen such that prescribed correlations between the outputs occur. A general formalism is introduced using a multiperceptron cost function that allows one to determine the maximal number of random inputs as a function of the desired values of the correlations. Replica-symmetric results for $K=2$ and $K=3$ are compared with properties of two-layer networks of tree-structure and fixed Boolean function between hidden units and output. The results show which correlations in the hidden layer of multilayer neural networks are crucial for the value of the storage capacity. [S1063-651X(97)12605-8]

PACS number(s): 05.20.-y, 02.70.−c

I. INTRODUCTION

One of the central tasks in the field of statistical mechanics of neural networks is a deeper understanding of the information processing abilities of multilayer feed-forward networks (MLN) [1,2]. After a thorough analysis of the single-layer perceptron it soon became clear that the very properties that entail the larger computational power of MLN also make their theoretical description within the framework of statistical mechanics much harder. Even the simplest case with just one hidden layer containing many fewer units than the input layer and with a prewired Boolean function from the hidden layer to the output has proven to be rather complicated to analyze exactly [3–6]. It is therefore important to develop useful and reliable approximate methods to study these practically important systems. For the characterization of the generalization ability *bounds* for the performance parameters have been shown to yield useful orientations [7,8]. For the storage capacity, i.e., the typical maximal number of random input-output mappings that can be implemented by the network, only rather crude bounds exist so far, and these are independent of the hidden-to-output mapping [9].

Let us start the discussion with a number of general open questions regarding the capacity of MLN. These questions, although only partially answered in the present work, may serve as a call for further investigation by the community of the statistical mechanics of neural networks.

A. Correlations among the hidden units

The increased computational power of MLN stems from the possibility that the different subperceptrons between input and hidden layer can all operate in the region beyond their storage capacity. The typically occurring errors of this regime can be compensated by other subperceptrons. However, this ''division of labor'' only works appropriately if the errors do not occur for all subperceptrons in *the same* patterns. Hence, intricate correlations depending on the hidden-to-output mapping develop in the hidden layer when the number of input-output pairs increases [10]. This qualitative picture has already been used to propose and analyze a learning algorithm for a special MLN, the parity machine [11]. It has been observed for some time that the organization of internal representations described by these correlations is crucial for the understanding of the storage and generalization abilities of MLN [3,12–15].

The approximation suggested in this work is to replace ''*division of labor*'' by an ''*average division of labor.*'' An approximate treatment of a MLN becomes possible if one does not require a definite mapping from the hidden layer to the output but instead prescribes the values for the correlations, i.e., the *average* relation between the hidden units and the output and also among the different hidden units themselves. The task is then to determine how many random inputs can be implemented by a set of K perceptrons, such that the outputs show definite correlations.

B. Interplay between correlations and the capacity

This approach will highlight which type of correlation is easy to implement and which is difficult, i.e., reduce the storage capacity significantly. It is already known that increasing the average correlation between each one of the hidden units and the desired output decreases the capacity. This result can be exemplified by the following well-known limits. The lowest capacity is achieved for hidden units, which are fully correlated with the desired outputs. In this case there is no division of labor and the MLN shrinks to a simple perceptron. The other limit is the parity machine, in which the correlation between each hidden unit and the output is zero. In this case the upper bound for the capacity of MLN with one hidden layer is achieved. Nevertheless, the general framework of how the capacity depends on the correlations between the output and a *partial set* of the hidden units is still unknown. The main problem is that with increasing K there is a tradeoff between a more flexible division of labor and an increasing complexity of possible correlations [16,17].

C. Possible scaling for the capacity

Of particular interest is the limit of an infinite number K of hidden units for which only few analytical results are known. For the AND machine the capacity is of $O(1)$ [12], whereas for the committee machine and the parity machine the capacity is of order $(\ln K)^{\delta}$, with $\delta=1/2$ [15] and 1 [4], respectively. These results may suggest one of the following two possible scenarios: in the first scenario, the capacity varies continuously as a function of the hidden-output correlations. Any δ in the range $0\leqslant\delta\leqslant 1$ can be found, depending on the correlations. In the second possible scenario, $\delta=1$ holds for the parity machine only, and all other hidden-output correlations result in a δ with a finite distance from 1.

D. Space of possible correlations

The simultaneous prescription of correlations involving several hidden units has to take into account that not all combinations of correlations are possible since they all derive from a common probability distribution. The question of whether there are forbidden combinations of correlations and what is their measure will be partially answered in the following discussion.

This paper is organized as follows. Section II sets the task and fixes the notations. In Sec. III a formalism is presented that is a generalization of the canonical phase space method developed by Gardner and Derrida [18] for the single-layer perceptron. Section IV contains general results for an arbitrary number K of perceptrons with a special subset of fixed correlations. In Secs. V and VI we study in detail the situations of $K=2$ and $K=3$ perceptrons, respectively, and compare the results with those known for tree-structured MLN with the same number of hidden units. Finally, Sec. VII comprises our conclusions.

II. THE STORAGE PROBLEM FOR CORRELATED PERCEPTRONS

We consider K spherical perceptrons with N/K inputs, one output, and couplings $\mathbf{J}_k\in\mathbb{R}^{N/K}, \mathbf{J}_k\mathbf{J}_k=N/K$ with $k=1,\ldots,K$. Then we choose a set of $(\alpha N)K$ random inputs $\boldsymbol{\xi}_k^{\nu}\in\mathbb{R}^{N/K}$ and one overall random output $\sigma^{\nu}=\pm 1$ with $\nu=1,\ldots,\alpha N$. The total number of random input and output bits is hence $\alpha N(N+1)$ and the number of adjustable weights is N as for the standard perceptron and for multilayer networks with tree-structure and fixed Boolean function between hidden units and output.

The outputs of the K perceptrons are given by

$$\tau_k^{\nu}=\mathrm{sgn}\left(\sqrt{\frac{K}{N}}\mathbf{J}_k\boldsymbol{\xi}_k^{\nu}\right). \tag{1}$$

Our aim is to determine the critical number $\alpha_c N$ of patterns for which coupling vectors $\mathbf{J}_k$ exist such that the averages

$$c_1=\langle\tau_k\sigma\rangle=\frac{1}{\alpha N}\sum_{\nu}\tau_k^{\nu}\sigma^{\nu}, \tag{2}$$

$$c_2=\langle\tau_k\tau_l\sigma\rangle=\frac{1}{\alpha N}\sum_{\nu}\tau_k^{\nu}\tau_l^{\nu}\sigma^{\nu},$$

$$c_3=\langle\tau_k\tau_l\tau_m\sigma\rangle=\frac{1}{\alpha N}\sum_{\nu}\tau_k^{\nu}\tau_l^{\nu}\tau_m^{\nu}\sigma^{\nu},$$

$$\vdots$$

$$c_K=\langle\tau_1\cdots\tau_K\sigma\rangle=\frac{1}{\alpha N}\sum_{\nu}\tau_1^{\nu}\cdots\tau_K^{\nu}\sigma^{\nu} \tag{3}$$

have prescribed values $c_1,c_2,\ldots,c_K$. This can be seen as a generalization of the program of Gardner and Derrida (GD) [18] who considered only one perceptron, i.e., $K=1$, and determined α_c in dependence on the fraction of errors f_{GD} related to c_1 by $c_1=1-2f_{\mathrm{GD}}$. An important aspect of the present investigation is that not only the correlation of each individual output τ_k with σ but also the correlation between different τ_k is taken into account.

As usual we assume that the components of the input patterns $\boldsymbol{\xi}_k^{\nu}$ as well as the overall outputs σ^{ν} are independent random variables with zero mean and unit variance. The transformation $\boldsymbol{\xi}_k^{\nu}\rightarrow\sigma^{\nu}\boldsymbol{\xi}_k^{\nu}$ then preserves the statistical properties of the inputs. In the following we therefore take $\sigma^{\nu}=1$ for all $\nu=1,\ldots,\alpha N$ without loss of generality.

Note that due to the independence of the inputs at different perceptrons all outputs τ_k have identical statistical properties. Therefore the correlations c_m as defined in Eqs. (2) and (3) do not depend on the particular subset of hidden units for which they are calculated. This corresponds to the permutation symmetry between hidden units in MLN with appropriate decoder functions [4–6].

It is particularly interesting to enforce correlations c_m that are identical to those that develop spontaneously in MLN with special Boolean functions between hidden layer and output. It has recently been shown how these correlations can be calculated from the joint probability distribution of the stabilities at the hidden units [10]. For the parity machine with K hidden units one finds $c_m=0$ for $m<K$ and $c_K=1$. For the committee machine the expressions are more complicated, for $K=3$ one finds $c_1=5/12$, $c_2=-1/6$, and $c_3=-3/4$.

III. FORMALISM

To analyze the storage abilities of correlated perceptrons we use a generalization of the formalism introduced by Gardner and Derrida [18]. A well suited form for our purposes is the one proposed by Griniasty and Gutfreund [19]. We are hence led to introduce a *multiperceptron cost function* [20]:

$$E(\mathbf{J}_1,\cdots,\mathbf{J}_K)=\sum_{\nu}V(\tau_1^{\nu},\ldots,\tau_K^{\nu}) \tag{4}$$

$$=\sum_{\nu}\Bigg[-\sum_{k}\tau_k^{\nu}+\mu_2\sum_{(k,l)}\tau_k^{\nu}\tau_l^{\nu}+\mu_3\sum_{(k,l,m)}\tau_k^{\nu}\tau_l^{\nu}\tau_m^{\nu}+\cdots+\mu_K\tau_1^{\nu}\cdots\tau_K^{\nu}\Bigg]. \tag{5}$$

The parameters μ_m play the role of chemical potentials determining the costs for a violation of the constraints on the correlations c_m. Our aim is to characterize the coupling vectors $\mathbf{J}_k$ that minimize $E(\mathbf{J}_1, \ldots, \mathbf{J}_K)$ and to find the critical threshold α_c for the number of inputs for which no couplings $(\mathbf{J}_1, \ldots, \mathbf{J}_k)$ exist that realize the desired correlations. This can be done by calculating the free energy

$$f(\alpha,\beta,\mu_2,\ldots,\mu_K) = -\lim_{N\to\infty}\frac{1}{\beta N}\left\langle\left\langle \ln\int\prod_{k=1}^{K} d\mu(\mathbf{J}_k)\exp[-\beta E(\mathbf{J}_1,\ldots,\mathbf{J}_K)]\right\rangle\right\rangle, \tag{6}$$

where $\langle\langle\cdots\rangle\rangle$ denotes the quenched average over the inputs and $d\mu(\mathbf{J}) = (2\pi e)^{-N/2K}\prod_{i=1}^{N/K} dJ_i\,\delta(\sum_{i=1}^{N/K} J_i^2 - N/K)$ is the usual integration measure for spherical perceptrons. Then

$$g(\alpha_c,\mu_2,\ldots,\mu_K) = \lim_{\beta\to\infty} f(\alpha,\beta,\mu_2,\ldots,\mu_K) \tag{7}$$

gives the typical minimum of $E(\mathbf{J}_1, \ldots, \mathbf{J}_K)$. The limit $\beta\to\infty$ corresponds to the saturation limit $\alpha\to\alpha_c$. The values $c_m^{(s)}$ of the correlations c_m defined in Eq. (2) in this saturation limit are from Eqs. (4) and (6) given by

$$\frac{1}{\alpha_c} g(\alpha_c,\mu_2,\ldots,\mu_K) = -K c_1^{(s)} + \mu_2\binom{K}{2} c_2^{(s)} + \cdots + \mu_K c_K^{(s)}, \tag{8}$$

$$\frac{1}{\alpha_c}\frac{\partial g(\alpha_c,\mu_2,\ldots,\mu_K)}{\partial \mu_k} = \binom{K}{k} c_k^{(s)}, \quad k = 2,\ldots,K. \tag{9}$$

Inverting these equations, we find the saturation values α_c and $\mu_m^{(s)}$ as functions of $c_1, \ldots, c_K$, which is what we were looking for.

The calculation of $g(\alpha_c,\mu_2,\ldots,\mu_K)$ proceeds along similar lines as for the single perceptron case studied in [19]. Within replica symmetry one has to introduce an order parameter q characterizing the typical overlap between two coupling vectors that contribute significantly to the free energy (6). In the limit $\beta\to\infty$ it is convenient to replace this order parameter by $x = \beta(1-q)$. If the minimum of the cost function is not degenerated, we will find $q\to 1$ for $\beta\to\infty$ with x remaining of order 1. Qualitatively x describes the steepness of the minimum of the cost function. The smaller x, the fewer couplings contribute significantly to the free energy for large β, i.e., the steeper the minimum of the cost function. Accordingly $x=\infty$ corresponds to a degenerated minimum since $q\neq 1$ even for $\beta\to\infty$.

For all choices of the parameters μ_m there is a minimum $V_{\min} = \min_{\{\tau_k\}} V(\tau_1,\ldots,\tau_K)$ of $V(\tau_1,\ldots,\tau_K)$ and hence $\alpha N V_{\min}$ is a lower bound for the cost function $E(\mathbf{J}_1,\ldots,\mathbf{J}_K)$. Now consider the subset of $\{\tau_k\}$ configurations that realize $V_{\min}$ and calculate the correlations c_m for this subset. The resulting values for the c_m are special in two respects. First, the value of α_c corresponding to them will occur for $x=\infty$ since the minimum of E is degenerated for $\alpha<\alpha_c$. Second, exactly these values of c_m will occur in a MLN with that Boolean function between hidden layer and output that maps all the $\{\tau_k\}$ configurations that result in $V_{\min}$ on the output $+1$. Consequently MLN with K hidden units and a fixed Boolean function between hidden layer and output will show up as ''pure cases'' defined by $x=\infty$ at α_c in our analysis and all situations with $x<\infty$ can be interpreted as these pure cases above saturation. Changing the parameters μ_m or equivalently the prescribed values of the c_m will hence induce continuous transformations between the different possible MLN.

The main steps of the formal analysis are sketched in Appendix A. The final result reads as follows [cf. Eqs. (A10) and (A11)]:

$$g(\alpha_c,\mu_2,\ldots,\mu_k) = -\min_x\left[\frac{1}{2x} - \alpha_c\int\prod_k Dt_k F(x,t_k)\right], \tag{10}$$

where

$$F(x,t_k) = \min_{\lambda_1,\ldots,\lambda_K}\left[\frac{1}{2x}\sum_k (\lambda_k - t_k)^2 + V(\mathrm{sgn}(\lambda_1),\ldots,\mathrm{sgn}(\lambda_K))\right] \tag{11}$$

and $Dt = \exp(-t^2/2)\,dt/\sqrt{2\pi}$.

The minimization in Eq. (11) is nontrivial. The quadratic terms in Eq. (11) are smallest for $\lambda_k^0 = t_k$. They compete with the step functions in $V(\mathrm{sgn}(\lambda_1),\ldots,\mathrm{sgn}(\lambda_K))$, giving rise to discontinuous jumps in F whenever one λ_k crosses zero. Closer inspection shows that for the global minimum one has

$$\lambda_k^0 = t_k \quad \text{or} \quad \lambda_k^0 = \begin{cases} 0^+ & \text{if} \quad t_k<0 \\ 0^- & \text{if} \quad t_k>0. \end{cases} \tag{12}$$

The saddle point equation that determines x can be written in the form

$$\frac{1}{\alpha_c} = \int\prod_k Dt_k \sum_k (\lambda_k^0 - t_k)^2. \tag{13}$$

Note that in this equation only those regions in the Gaussian integrals for which $\lambda_k^0 \neq t_k$ contribute.

IV. GENERAL RESULTS FOR PRESCRIBED HIGHEST AND LOWEST CORRELATION

Of particular interest is the case in which only the values of c_1 and c_K are prescribed, i.e. $\mu_2 = \mu_3 = \cdots = \mu_{K-1} = 0$ in

TABLE I. Correlation coefficients and storage capacity for an ensemble of K perceptrons in the pure cases characterized by $x=\infty$ (see text).

	μ	c_1 $(x=\infty)$	c_K $(x=\infty)$	$1/\alpha_c$ $(x=\infty)$
I	$\mu<1$	1	1	$K/2$
II	$\mu=1$	$1-2/K+1/2^{(K-1)}K$	$-1+1/2^{(K-1)}$	$K/2-K\int_{-\infty}^{0}Dt\,t^2[H(t)]^{K-1}$
III	$\mu>1$	$1-2/K$	-1	$K/2-K\int_{0}^{\infty}Dt\,t^2([H(-t)]^{K-1}-[H(t)]^{K-1})$

the cost function (4). It describes the interpolation between individual perceptrons ($\mu_K=0$) and the parity machine ($\mu_K\to\pm\infty$), which is known to saturate the asymptotic upper bound $\alpha_c=\ln K/\ln 2$ for the storage capacity for large K [4]. This special case is also sufficient to discuss the relation with the most important tree-structured MLN for $K=2$ and $K=3$. Moreover, the necessary algebra simplifies somewhat.

Let us first note that the correlation coefficients c_1 and c_K are not independent of each other. It is hence not possible to prescribe arbitrary values for them. According to their definition (2),(3) we always have $c_1,c_K\in(-1,+1)$. Moreover, it is sufficient to consider positive values of c_1 only, which is guaranteed by the structure of the cost function (4). Finally the relation

$$c_K\geqslant Kc_1-(K-1) \tag{14}$$

must hold. It is a consequence of the obvious observation that the difference between c_1 and c_K is maximal if for every pattern at most one perceptron has negative output, which corresponds to the equality sign in Eq. (14).

To perform the detailed analysis we denote μ_K simply by μ to get

$$E(\mathbf{J}_1,\ldots,\mathbf{J}_K)=\sum_{\nu}\left[-\sum_{k}\tau_k^{\nu}+\mu\prod_{k}\tau_k^{\nu}\right]. \tag{15}$$

Accordingly Eq. (11) simplifies to

$$F(x,t_k)=\min_{\lambda_1,\ldots,\lambda_K}\left[\frac{1}{2x}\sum_{k}(\lambda_k-t_k)^2-\sum_{k}\mathrm{sgn}(\lambda_k)+\mu\,\mathrm{sgn}(\lambda_1\lambda_2\cdots\lambda_K)\right]. \tag{16}$$

In Appendix B the following expressions for the correlation coefficients c_1 and c_K are derived:

$$c_1=1-2\,H(2\sqrt{x})-[f_1(|\mu|,x,0)-f_2(|\mu|,x,0)], \tag{17}$$

$$c_K=2^K[1/2-H(2\sqrt{x})]^K-K\,\mathrm{sgn}(\mu)[f_1(|\mu|,x,0)+f_2(|\mu|,x,0)] \tag{18}$$

Moreover the saddlepoint equation fixing x can be transformed into

$$\frac{1}{K\alpha_c}=\frac{1}{2}-H(2\sqrt{x})-2\sqrt{x}\frac{e^{-2x}}{\sqrt{2\pi}}+\frac{1}{2}[f_1(|\mu|,x,1)+f_2(|\mu|,x,1)]. \tag{19}$$

As usual we have used the abbreviation $H(x)=\int_x^{\infty}Dt$. $f_1(|\mu|,x,L)$ and $f_2(|\mu|,x,L)$ (with $L=0,1$) are integrals over sums of products of error functions explicitly given in Appendix B. The final analysis of these equations has to be done numerically.

As discussed in the last section it is of particular interest to find the correlations c_1 and c_K for which $x=\infty$ at α_c. From Eqs. (17)–(19) and (B8)–(B11) we find the results that are listed in Table I.

Note that all three pairs $(c_1,c_K)|_{(x=\infty)}$ lie on the line given by Eq. (14), in fact (I) and (III) are the end points of this line.

It is at first sight surprising that the parity machine does not occur in Table I. However, from the structure of the cost function Eq. (15) it is clear that the internal representations of the parity function realize $V_{\min}$ only in the limit $\mu\to\pm\infty$. For finite $|\mu|$ the first term in Eq. (15) suppresses configurations with more than one negative output and gives rise to case (I) or (III).

V. $K=2$

The simplest case to apply the above concepts is provided by two perceptrons with $N/2$ inputs each corresponding to $K=2$. The only relevant correlations are c_1 and c_2 [see Eqs. (2) and (3)]. The relative importance of these in the cost function (15) is regulated by μ.

Solving Eq. (19) numerically for the case $K=2$ we find $c_1(\alpha_c,\mu)$ and $c_2(\alpha_c,\mu)$ from Eqs. (17) and (18) and inverting these dependencies we arrive at $\alpha_c(c_1,c_2)$.

In Fig. 1 (left) the dependence of α_c on c_2 for several values of c_1 is shown. Solutions exist only inside the shaded areas whose boundaries correspond to $c_1=0$ and $c_2=2c_1-1$, respectively [cf. Eq. (14)]. The maxima of $\alpha_c(c_2)$ at constant c_1 occur for the uncorrelated system $\mu=0$, implying $c_2=c_1^2$ as expected since an additional constraint on c_2 can only reduce α_c. The values of $\alpha_c(c_1,c_1^2)$ at these maxima are consistent with the results of Gardner and Derrida for the minimal fraction of errors $f_{\mathrm{GD}}=(1-c_1)/2$ [18].

As a complement, the dependence $\alpha_c(c_1)$ for fixed c_2 is shown in the right part of Fig. 1. Lines for c_2 and $-c_2$ start at the same point for $c_1=0$. This corresponds to $\mu=\pm\infty$,

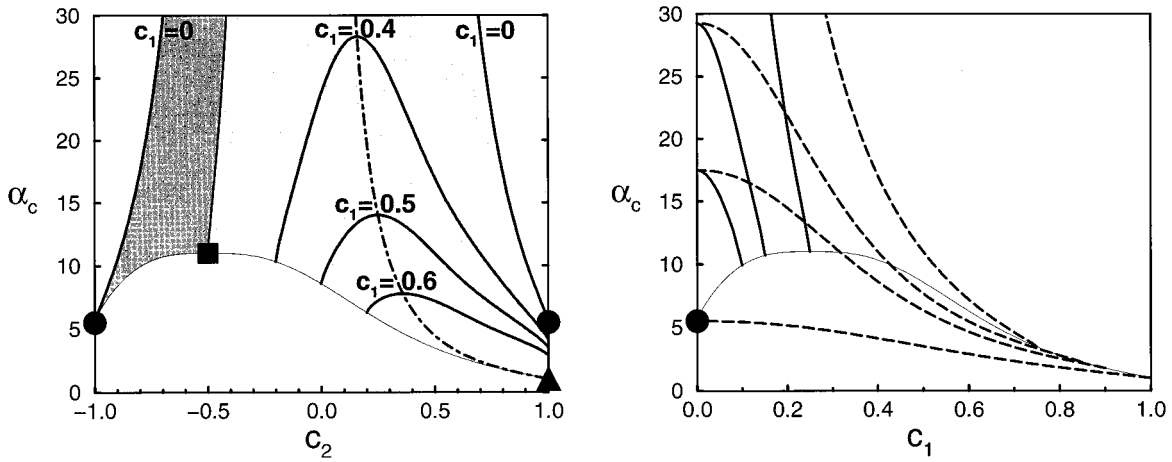

FIG. 1. Storage capacity $\alpha_c(c_1,c_2)$ for $K=2$ correlated perceptrons. Left: $\alpha_c(c_2)$ for $c_1=0$, 0.4, 0.5, and 0.6. Outside the shaded areas no solutions exist; dark shade corresponds to $\mu>1$ and light shade to $\mu<1$. The dashed-dotted line ($\mu=0$) gives the location of the maxima. The symbols denote the pure cases corresponding to the MLN summarized in Table II. Right: $\alpha_c(c_1)$ for (from bottom to top): $c_2=1$, 0.8, 0.7, and 0.5 (dashed) and $c_2=-0.8,-0.7$, and -0.5 (full). The lines end at the thin line given by $c_2=2c_1-1$. The symbol corresponds to the parity machine.

where the value of c_1 has negligible influence on the cost function (15). With increasing c_1 the value of α_c always decreases because additional constraints are to be satisfied. These new constraints give rise to $c_1>0$ and are hence harder to satisfy for negative values of c_2. Finally all lines end at the thin line given by $c_2=2c_1-1$.

The pure cases for $K=2$ defined by $x=\infty$ at α_c are indicated by symbols in Fig. 1. They correspond to two-layer networks with two hidden units and fixed Boolean functions between hidden layer and output and are summarized in Table II.

In our analysis the AND machine denotes the situation in which the two perceptrons have to give *simultaneously* the correct output $\sigma^\nu=+1$ for all patterns. The storage capacity is hence given by the Gardner result, i.e., $\alpha_c=1$ since each perceptron has $N/2$ couplings only. Note that the AND machine investigated in [12] has random outputs $\sigma^\nu=\pm 1$ and therefore the value for α_c is different. The XOR function defines the $K=2$ parity machine for which the replica-symmetric α_c was first obtained in [3,4]. The result for the OR machine is new; again it refers to the situation where random inputs all have to be mapped on $\sigma^\nu=+1$. Finally let us note that there is another rather trivial pure case given by $c_1=c_2=0$ with $\alpha_c=\infty$ corresponding to the Boolean function that gives output $+1$ on any input.

The results obtained for $K=2$ are summarized in Fig. 2, which shows the region of allowed values in the c_1-c_2 plane together with lines of constant α_c and constant μ. The arrows at the lines of constant μ point to smaller values of α_c. The above discussed hidden unit machines are again marked by the symbols of Table II. All other points can be interpreted as these machines above their storage capacity. Note that the same point could be associated with different machines beyond saturation since by prescribing the correlations appropriately we can induce continuous transitions between different machines.

TABLE II. Patterns of correlations for $K=2$ perceptrons equivalent to two-layer networks with fixed Boolean function between hidden units and output.

Symbol in Fig. 1	c_1	c_2	α_c $(x=\infty)$	μ	Boolean function
Triangle	1	1	1	<1	AND
Square	1/4	$-1/2$	11.01	$=1$	OR
Circle	0	-1	5.50	>1	XOR

VI. $K=3$

A similar analysis can be performed for $K=3$. As discussed in Sec. IV we set $\mu_2=0$ and denote μ_3 simply by μ. Similar to the last section we can then determine $\alpha_c(c_1,c_3)$ from a numerical analysis of Eqs. (17) and (18).

Figure 3 (left) shows the dependence of the critical storage capacity α_c on c_3 for fixed values of c_1. The dependen-

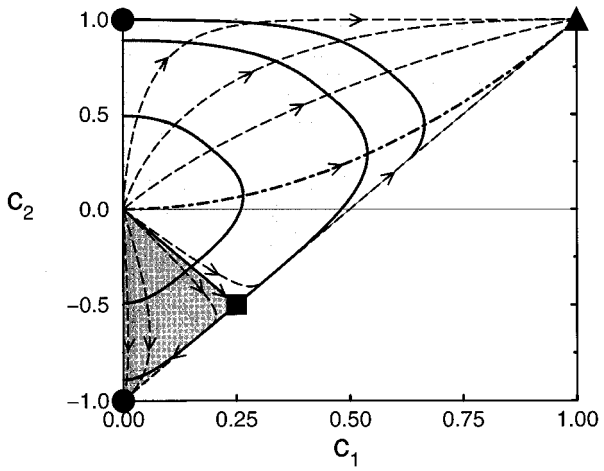

FIG. 2. Contour map of $\alpha_c(c_1,c_2)$ and $\mu(c_1,c_2)$ for $K=2$ correlated perceptrons. Full lines correspond to $\alpha_c=100$, $11.01=\alpha_c^{\mathrm{OR}}$, $5.50=\alpha_c^{\mathrm{XOR}}$ (from left to right), dashed lines to $\mu=-10$, -2, -1, 0, 0.99, 1.01, 2, and 10 (from top to bottom). Symbols denote the same MLN as in Table II.

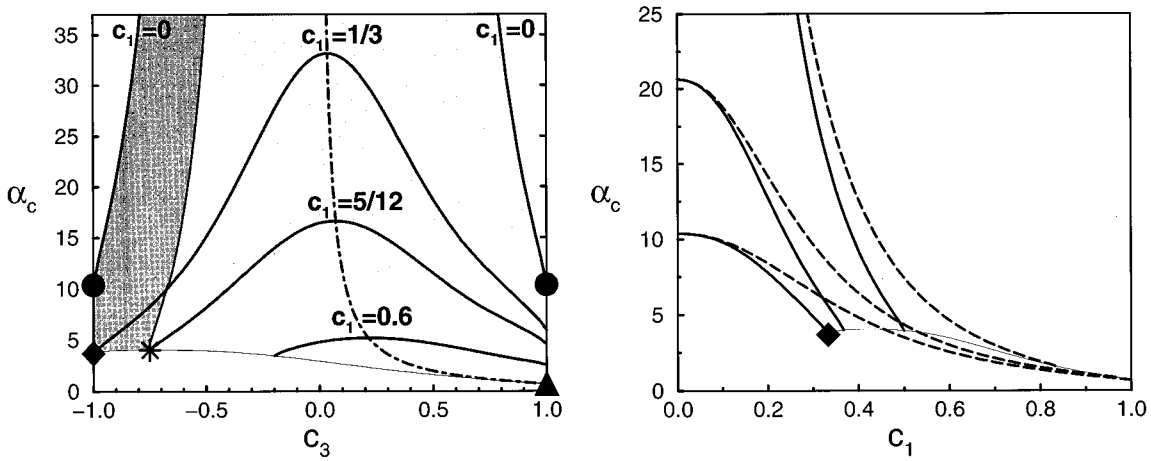

FIG. 3. Storage capacity $\alpha_c(c_1,c_3)$ for $K=3$ correlated perceptrons. Left: $\alpha_c(c_3)$ for $c_1=0,\ 1/3,\ 5/12$ and 3/5. Outside the shaded areas no solutions exist, dark shade corresponds to $\mu>1$, light shade to $\mu<1$. The dashed-dotted line ($\mu=0$) gives the location of the maxima. The symbols denote the pure cases corresponding to the MLN summarized in Table III. Right: $\alpha_c(c_1)$ for (from bottom to top): $c_3=1$, 0.9, and 0.5 (dashed) and $c_3=-1,-0.9$ and -0.5 (full). The lines end at the thin line given by $c_3=3c_1-2$. The symbol corresponds to the machine giving overall output $+1$ only if exactly one hidden unit is -1.

cies are rather similar to the case $K=2$ shown in the left part of Fig. 1. Again solutions $c_1(\alpha_c,c_3)$ exist only in shaded areas. The maxima of the $\alpha_c(c_3)$ curves lie on the dash-dotted line corresponding to independent perceptrons ($\mu=0$). They are hence characterized by $c_3=c_1^3$ and are again consistent with the Gardner-Derrida results on the minimal fraction of errors for perceptrons above saturation [18].

As a complement, the dependence $\alpha_c(c_1)$ for fixed values of c_3 is shown in the right part of Fig. 3. Again, similar to the case $K=2$, we find that α_c decreases with increasing c_1. In particular, the lines for $c_3=\pm 1$ show how the storage capacity decreases from the value of the $K=3$ parity machine at $c_1=0$ if additional constraints showing up in $c_1>0$ are included. All lines end at the thin line given by $c_3=3c_1-2$.

The symbols in Fig. 3 refer again to pure cases with $x=\infty$ at α_c corresponding to the MLN summarized in Table III. In addition to the AND and parity machine we now have the committe machine and a machine with the Boolean function for which the output is $+1$ if *exactly one* hidden unit is -1.

We can again summarize the results in a contour plot showing lines of constant α_c and μ in the c_1-c_3 plane (Fig. 4). Only combinations of c_1 and c_3 that belong to the shaded areas are possible: the light shade corresponds to $\mu<1$, dark shade to $\mu>1$. The arrows at the dashed lines of constant μ point again into regions of lower α_c; the symbols are those of Table III. Large values of c_1 imply a strong correlation of every perceptron with the common output and give therefore small α_c and a narrow interval of consistent values of c_3. Relaxing the constraint on c_1 allows a more efficient ''division of labor'' between the perceptrons and results in a broader spectrum of c_3 values and enhanced storage capacity. Accordingly the largest values of α_c are possible for $c_1=0$. Then α_c only depends on c_3 and starting from the value 10.37 for the parity machine at $c_3=\pm 1$ it increases without bound with decreasing $|c_3|$.

An important aspect of the case $K=3$ is that there is a correlation coefficient, c_2, that was not presribed (since we put $\mu_2=0$). It is nevertheless of interest to know the value of c_2 that corresponds to different choices of c_1 and c_3. The easiest way to obtain c_2 is via a maximum entropy argument. This is sketched in Appendix C. The result is

$$c_2=-\frac{1}{2}+\sqrt{\frac{1}{4}+c_1^2+c_1c_3}. \tag{20}$$

It is interesting to note that for the values $c_1=5/12$ and $c_3=-3/4$ characteristic for the committee machine this formula gives $c_2=-1/6$, which is in fact the correct result [10]. The committee function for $K=3$ hence does not imply constraints on c_2 and is already uniquely characterized by the values of c_1 and c_3.

TABLE III. Patterns of correlations for $K=3$ perceptrons equivalent to two-layer networks with fixed Boolean function between hidden units and output.

Symbol in Fig. 3	c_1	c_3	α_c $(x=\infty)$	μ	Boolean function
Triangle	1	1	2/3	$\mu<1$	AND
Star	5/12	$-3/4$	4.02	$\mu=1$	Committee
Diamond	1/3	-1	3.669	$\mu>1$	$(-++),(+-+),(++-)$
Circle	0	± 1	10.37	$\mu=\pm\infty$	Parity

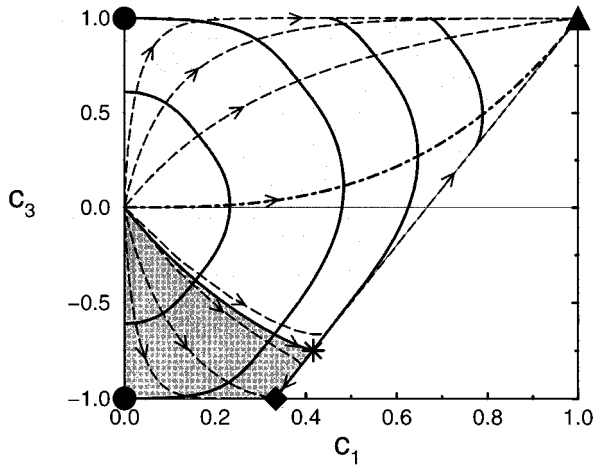

FIG. 4. Contour map of $\alpha_c(c_1,c_3)$ and $\mu(c_1,c_3)$ for $K=3$ correlated perceptrons. Full lines correspond to $\alpha_c=100$, $10.37=\alpha_c^{\mathrm{PAR}}$, $4.02=\alpha_c^{\mathrm{COM}}$ and 2 (from left to right), dashed lines to $\mu=-10$, $-2,-1,0$, 0.99, 1.01, 2, and 10 (from top to bottom). Symbols denote the same MLN as in Table (III).

VII. CONCLUSIONS

In the present paper we have considered ensembles of K perceptrons with random inputs and investigated the possibility to choose the couplings such that prescribed correlations c_m between the outputs of the perceptrons occur. For any combinatorically possible combination of c_m there is a critical value $\alpha_c(c_1,\ldots,c_K)$ and solutions for the couplings of the perceptrons exist if the number of inputs is less than $N\alpha_c$. These investigations establish a relation between the results for single perceptrons above their storage capacity and those for several MLN with tree structure and K hidden units and fixed Boolean function between hidden layer and output. Similar ideas were pursued in [11] and [13] where approximate expressions for the storage capacity of a parity machine and committee machine, respectively, were obtained from the results of Gardner and Derrida on the minimal fraction of errors of perceptrons beyond saturation and in [21] where analogies between a committee machine and noisy perceptrons were investigated. In the present paper the influence of higher correlations that are known to be important for the storage abilities was also taken into account. The results show which correlations are difficult to implement and are therefore important for the determination of the storage capacity and which are easy and therefore not very restrictive. A detailed analysis was carried out for $K=2$ and $K=3$.

The technique used is a generalization of the canonical phase space analysis introduced by Gardner and Derrida. The results were obtained within the replica-symmetric ansatz. They should hence be seen as a mere first orientation since it its well known that replica-symmetry breaking (RSB) is crucial for both the description of perceptrons above saturation [22] and the storage abilities of MLN [4–6]. An investigation of the problem within RSB though highly desirable seems technically rather involved. Also the extension of the analysis to asymptotic behavior for $K\to\infty$ would be very interesting and would hopefully shed some light on the still controversial problem of the storage capacity of MLN in this limit.

ACKNOWLEDGMENTS

We have benefited from discussions with Chris van den Broeck and John Hertz. A.E. is grateful to the Minerva Center for Neural Networks for hospitality during a stay at Bar Ilan University in Ramat Gan where the initial stages of this work were performed.

APPENDIX A

In this appendix we outline the calculation of the free energy Eq. (6) corresponding to the cost function Eq. (4) within replica symmetry. To this end we employ a generalization of the formalism of Griniasty and Gutfreund [19].

To perform the average over the random patterns we use the replica trick

$$f(\mu_2,\ldots,\mu_K,\beta)=-\frac{1}{\beta N}\langle\langle\ln\mathcal{Z}\rangle\rangle=-\frac{1}{\beta N}\lim_{n\to 0}\frac{\langle\langle\mathcal{Z}^n\rangle\rangle-1}{n} \tag{A1}$$

involving the partitition function $\mathcal{Z}$:

$$\mathcal{Z}=\int_{-\infty}^{\infty}\prod_k\frac{d\mathbf{J}_k}{\sqrt{2\pi e}}\,\delta(\mathbf{J}_k^2-N/K)\int_{-\infty}^{\infty}\prod_{k\nu}d\lambda_k^\nu\,\delta(\lambda_k^\nu-\mathbf{J}_k\boldsymbol{\xi}_k^\nu\sqrt{K/N})\,e^{-\beta\Sigma_\nu^{\alpha N}V(\lambda_1^\nu,\ldots,\lambda_K^\nu)}$$

$$V(\lambda_1^\nu,\ldots,\lambda_K^\nu)=-\sum_k\mathrm{sgn}(\lambda_k^\nu)+\mu_2\sum_{(k,l)}\mathrm{sgn}(\lambda_k^\nu\lambda_l^\nu)+\cdots+\mu_K\mathrm{sgn}(\lambda_1^\mu\cdots\lambda_K^\mu). \tag{A2}$$

Introducing integral representations for the δ functions and performing the average over the patterns we find

$$\langle\langle\mathcal{Z}^n\rangle\rangle=\int_{-\infty}^{\infty}\prod_{a<b;k}dq_k^{ab}\int_{-\infty}^{\infty}\prod_{a<b;k}dF_k^{ab}\frac{N}{2\pi K}\int_{-\infty}^{\infty}\prod_{a;k}\frac{dE_k^a}{4\pi K}\exp\left(\frac{N}{K}\left[\frac{1}{2}\sum_k\mathrm{tr}(Q_kA_k)+G_2(F_k^{ab},E_k^a)\right]+\alpha NG_1(Q_1\cdots Q_K)\right), \tag{A3}$$

where

$$G_2(F_k^{ab},E_k^a)=-\frac{1}{2}\sum_k\ [n+\mathrm{tr}(\ln A_k)] \tag{A4}$$

and

$$G_1(Q_1\cdots Q_K)=\ln\left[\int_{-\infty}^{\infty}\prod_k d\boldsymbol{\lambda}_k\int_{-\infty}^{\infty}\prod_k\frac{d\mathbf{y}_k}{(2\pi)^n}\exp\left(i\sum_k\ \mathbf{y}_k\boldsymbol{\lambda}_k-\frac{1}{2}\sum_k\ \mathbf{y}_kQ_k\mathbf{y}_k^T+\beta\sum_a\ V(\lambda_1^a,\ldots,\lambda_K^a)\right)\right]. \tag{A5}$$

Here $\boldsymbol{\lambda}_k=(\lambda_k^1,\ldots,\lambda_k^n)$ and $\mathbf{y}_k=(y_k^1,\ldots,y_k^n)$ and we have used the matrices Q_k and A_k:

$$Q_k=\begin{pmatrix}1 & q_k^{ab}\\ q_k^{ab} & 1\end{pmatrix},\quad A_k=\begin{pmatrix}iE_k^a & -iF_k^{ab}\\ -iF_k^{ab} & iE_k^a\end{pmatrix}, \tag{A6}$$

where as usual q_k^{ab} describes the overlap between two replicas a,b in the coupling space of perceptron k,

$$q_k^{ab}=\mathbf{J}_k^a\mathbf{J}_k^bK/N. \tag{A7}$$

The saddle point for E_k^a and F_k^{ab} is given by $Q_k^{-1}=A_k$, resulting in

$$\langle\langle\mathcal{Z}^n\rangle\rangle\simeq\int_{-\infty}^{\infty}\prod_{a<b;k}dq_k^{ab}\exp\left(\frac{N}{2K}\sum_k\ \ln(\det Q_k)+\alpha NG_1(Q_1\cdots Q_K)\right). \tag{A8}$$

To evaluate the remaining saddle point integral we use the replica-symmetric ansatz $q_k^{ab}=q_k$ for all $a\neq b$. Moreover, we expect permutation symmetry between the different perceptrons implying $q_k=q$ for all $k=1,\cdots,K$. Then $\ln[\det Q]=n[\ln(1-q)+q/(1-q)]$ and for $G_1(q_1,..,q_K)$ it follows that

$$\begin{aligned}G_1(q)&=\ln\left[\int_{-\infty}^{\infty}\prod_{a;k}d\lambda_k^a\int_{-\infty}^{\infty}\prod_{a;k}\frac{dy_k^a}{2\pi}\exp\left(i\sum_{a;k}\ y_k^a\lambda_k^a-\frac{1-q}{2}\sum_{a,k}\ (y_k^a)^2-\frac{q}{2}\sum_k\ \left(\sum_a\ y_k^a\right)^2-\beta\sum_a\ V(\lambda_1^a,\ldots,\lambda_K^a)\right)\right]\\&\simeq n\int_{-\infty}^{\infty}\prod_k\ Dt_k\ln\int_{-\infty}^{\infty}\prod_k\frac{d\lambda_k}{\sqrt{2\pi(1-q)}}\exp\left(-\sum_k\ \frac{(\lambda_k-t_k\sqrt{q})^2}{2(1-q)}-\beta V(\lambda_1,\ldots,\lambda_K)\right).\end{aligned} \tag{A9}$$

In order to calculate the function $g(\alpha_c,\mu_2,\ldots,\mu_K)$ [Eq. (7)] we have to consider the *saturation limit* $\beta\to\infty$. It is convenient then to use the rescaled saddle point variable $x=\beta(1-q)$ instead of q. In this way we obtain

$$g(\alpha_c,\mu_2,\ldots,\mu_K)=-\min_x\left[\frac{1}{2x}-\alpha_c\int_{-\infty}^{\infty}\prod_k\ Dt_kF(x,t_1,t_2,\ldots,t_K)\right], \tag{A10}$$

$$F(x,t_1,t_2,\ldots,t_K)=\min_{\lambda_1,\ldots,\lambda_K}\left(\sum_k\ \frac{(\lambda_k-t_k)^2}{2x}+V(\lambda_1,\ldots,\lambda_K)\right), \tag{A11}$$

which coincides with Eq. (10). The saddle point equation (13) determining x follows by explicit differentiation of Eq. (A10) with respect to x.

APPENDIX B

In this Appendix we sketch the main steps of the derivation of the saddle point equation (13) and of the free energy (A10) for the case where only c_1 and c_K are prescribed. We also give the explicit expressions for c_1 and c_K as a function of α_c and μ.

The calculation of g as given by Eqs. (A10) and (16) requires minimization of

$$F(x,t_1,t_2,\ldots,t_K)=\min_{\lambda_1,\lambda_2,\ldots,\lambda_K}\left[\sum_{k=1}^K\frac{(\lambda_k-t_k)^2}{2x}-\sum_{k=1}^K\ \mathrm{sgn}(\lambda_k)+\mu\prod_{k=1}^K\ \mathrm{sgn}(\lambda_k)\right]. \tag{B1}$$

From Eq. (12) we have

$$\mathrm{sgn}(\lambda_k^0)=\begin{cases}\mathrm{sgn}(t_k) & \text{if } \lambda_k^0=t_k\\ -\mathrm{sgn}(t_k) & \text{otherwise.}\end{cases} \tag{B2}$$

Equation (B1) then becomes

$$F(x,t_1,t_2,\ldots,t_K)=\left[\mu S(\boldsymbol{\lambda}^0)-\sum_{k=1}^{K}\operatorname{sgn}(t_k)+\sum_{\forall_j \lambda_j^0=0^{\pm}}\left(\frac{t_j^2}{2x}+2\operatorname{sgn}t_j\right)\right]. \tag{B3}$$

Here $S(\boldsymbol{\lambda}^0)=\Pi_{k=1}^{K}\operatorname{sgn}(\lambda_k^0)=(-1)^m\Pi_{k=1}^{K}\operatorname{sgn}(t_k)$ were m counts all $\lambda_k^0\equiv 0$. The last sum in Eq. (B3) has only contributions from those λ with $\lambda_j^0=0^{\pm}$.

To minimize F for given $\mathbf{t}=t_1,\ldots,t_K$ we have to find which of the 2^K configurations $\{\lambda_1^0,\ldots,\lambda_K^0\}$, $\lambda_k^0=\{0^{\pm},t_k\}$, minimizes Eq. (B3). A suitable procedure to do this is as follows. We first make the last term in Eq. (B3) as small as possible. That is for all t_j with $t_j\in(-2\sqrt{x},0)$ we choose for a first try $\lambda_j^0=0^{\pm}$. We denote the resulting value for $S(\boldsymbol{\lambda}^0)$ by S^*. [$S^*=(-1)^{\eta}$ where η is the number of all $t_k<-2\sqrt{x}$.] If $\mu S^*(\mathbf{t})<0$, the optimal configuration has already been found because the first summand is at its miminum as well. If, on the other hand, $\mu S^*(\mathbf{t})>0$ there is competition between the first and the last terms in Eq. (B3). One may then change the sign of $S(\boldsymbol{\lambda}^0)$ in order to lower $F(x,t_1,t_2,\ldots,t_K)$ by $2|\mu|$ by either setting a single $\lambda_l^0=0^{\pm}$ although $t_l\notin(-2\sqrt{x},0)$ or setting a single $\lambda_l^0=t_l$ for one $t_l\in(-2\sqrt{x},0)$. The corresponding changes in F are $2[w(t_l)-|\mu|]$, where

$$w(t)=\begin{cases} t^2/4x-1 & \text{if } t\in(-\infty,-2\sqrt{x}) \\ -t^2/4x+1 & \text{if } t\in(-2\sqrt{x},0) \\ t^2/4x+1 & \text{if } t\in(0,\infty). \end{cases} \tag{B4}$$

In the saddle point equation (13) only regions in the integral contribute for which $\lambda_j^0\neq t_j$ for at least one j. Formalizing the above consideration we find

$$\frac{1}{\alpha_c}=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}Dt_1\cdots Dt_K\left\{\sum_{k=1}^{K}t_{k=1}^2\Theta_I(t_k)+K\Theta(\mu S^*)\Theta(|\mu|-w(t_1))t_1^2(-1)^{\Theta_I(t_1)}\prod_{k=2}^{K}\Theta(w(t_k)-w(t_1))\right\}, \tag{B5}$$

$$\Theta_I(t)=\begin{cases} 1 & \text{if } t\in(-2\sqrt{x},0) \\ 0 & \text{otherwise.} \end{cases} \tag{B6}$$

The first term of Eq. (B5) stems from our first guess minimizing the last term of Eq. (B3) only. The various Θ functions in the term that contributes only for $\mu S^*>0$ implement the different cases discussed in context with Eq. (B4). The integration variables can be defined such that $t_1<t_2<\cdots<t_K$ always with no restriction of generality.

The integration over $t_2,\ldots,t_K$ yields a product of sums of two error functions. Finally the saddle point equation reads

$$\frac{1}{K\alpha_c}=\frac{1}{2}-H(2\sqrt{x})-2\sqrt{x}\frac{e^{-2x}}{\sqrt{2\pi}}+\frac{1}{2}[f_1(|\mu|,x,1)+f_2(|\mu|,x,1)], \tag{B7}$$

$$f_1(|\mu|<1,x,L)=[-1]^L\int_{-2\sqrt{x}}^{-2\sqrt{x(1-|\mu|)}}Dt_1t_1^{2L}([H(t_1)+H_m(t_1)]^{K-1}+\operatorname{sgn}\mu[H(t_1)-H_m(t_1)]^{K-1}), \tag{B8}$$

$$f_1(|\mu|>1,x,L)=\left\{\int_0^{2\sqrt{x(|\mu|-1)}}Dt_1t_1^{2L}([H(t_1)+H_p(t_1)]^{K-1}+\operatorname{sgn}\mu[H(t_1)-H_p(t_1)]^{K-1})\right.$$
$$\left.+[-1]^L\int_{-2\sqrt{x}}^{0}Dt_1t_1^{2L}([H(t_1)+H_m(t_1)]^{K-1}+\operatorname{sgn}\mu[H(t_1)-H_m(t_1)]^{K-1})\right\}, \tag{B9}$$

where we introduced the abbreviation $H_p(t_1)=H(\sqrt{8x+t_1^2})$, $H_m(t_1)=H(\sqrt{|8x-t_1^2|})$, and $H_m^-(t_1)=H(-\sqrt{|8x-t_1^2|})$. As usual $H(t)=\int_t^{\infty}Dt$. Similarly

$$f_2(|\mu|<1,x,L)=\int_{-2\sqrt{x(1+|\mu|)}}^{-2\sqrt{x}}Dt_1t_1^{2L}([H_m^-(t_1)+H(-t_1)]^{K-1}-\operatorname{sgn}\mu[H_m^-(t_1)-H(-t_1)]^{K-1}), \tag{B10}$$

$$f_2(|\mu|>1,x,L)=\left\{\int_{-2\sqrt{2x}}^{-2\sqrt{x}}Dt_1t_1^{2L}([H_m^-(t_1)+H(-t_1)]^{K-1}-\operatorname{sgn}\mu[H_m^-(t_1)-H(-t_1)]^{K-1})\right.$$
$$\left.+\int_{-2\sqrt{x(1+|\mu|)}}^{-2\sqrt{2x}}Dt_1t_1^{2L}([H_m(t_1)+H(-t_1)]^{K-1}-\operatorname{sgn}\mu[H_m(t_1)-H(-t_1)]^{K-1})\right\} \tag{B11}$$

A common feature of Eqs. (B8)–(B11) is that in the binomial expression those terms cancel, which correspond to regions with $\mu S^* < 0$.

The calculation of g proceeds along similar lines:

$$g/\alpha_c = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} Dt_1 \cdots Dt_K \left\{ \mu S^* - \sum_{k=1}^{K} \mathrm{sgn}(t_k) + 2 \sum_{\forall_j \lambda_j^0 = 0} \mathrm{sgn} t_j \right\}$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} Dt_1 \cdots Dt_K \left\{ \mu S^* - 2 \sum_{k=1}^{K} \Theta_I(t_k) + 2K\Theta(\mu S^*)\Theta(|\mu| - w(t_1))[-\mu S^* + (-1)^{\Theta_I(t_1)} \mathrm{sgn}(t_1)] \right.$$

$$\left. \times \prod_{k=2}^{K} \Theta(w(t_k) - w(t_1)) \right\}. \tag{B12}$$

We find

$$g/\alpha_c = -2KE(2\sqrt{x}) + 2^K \mu E^K(2\sqrt{x}) + K\{f_1(|\mu|,x,0) - f_2(|\mu|,x,0) - |\mu|[f_1(|\mu|,x,0) + f_2(|\mu|,x,0)]\}. \tag{B13}$$

Performing the derivative of g/α_c with respect to μ one realizes that there is no contribution from the μ dependence of the integration limits in Eqs. (B8)–(B11). Hence the expression (B13) for g/α_c is already of the form $g/\alpha_c = -Kc_1 + \mu c_K$ and we arrive at Eqs. (17) and (18) for the correlation coefficients c_1 and c_K.

APPENDIX C

To determine c_2 for given values of c_1 and c_3 we look for the probability distribution $P(\tau_1, \tau_2, \tau_3)$ that for the given values of c_1 and c_3 realizes the maximal entropy. Because of the permutation symmetry between the perceptrons we have only to determine the probabilities p_k of output configurations with k negative outputs where $k = 0, \ldots, 3$. Hence we have to maximize

$$S = -p_0 \ln p_0 - 3p_1 \ln p_1 - 3p_2 \ln p_2 - p_3 \ln p_3 + \lambda_0(p_0 + 3p_1 + 3p_2 + p_3 - 1) + \lambda_1(p_0 + p_1 - p_2 - p_3 - c_1) + \lambda_3(p_0 - 3p_1 + 3p_2 - p_3 - c_3), \tag{C1}$$

where the λ_k are the Lagrange multipliers incorporating the constraints. Performing the derivatives with respect to the p_k yields

$$p_0 p_3 = p_1 p_2. \tag{C2}$$

Using the constraints to solve for the p_k gives

$$c_2 = -\frac{1}{2} \pm \sqrt{\frac{1}{4} + c_1^2 + c_1 c_3}, \tag{C3}$$

where only the upper sign gives rise to positive values for all p_k.

[1] J. A. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, CA, 1991).

[2] *Parallel Distributed Processing*, edited by D. E. Rummelhart and J. E. McClelland (MIT Press, Cambridge, MA, 1986).

[3] M. Mezard and S. Patarnello (unpublished).

[4] E. Barkai, D. Hansel, and I. Kanter, Phys. Rev. Lett. **65**, 2312 (1990).

[5] E. Barkai, D. Hansel, and H. Sompolinsky, Phys. Rev. A **45**, 4146 (1992).

[6] A. Engel, H. M. Koehler, F. Tschepke, H. Vollmayr, and A. Zippelius, Phys. Rev. A **45**, 7590 (1992).

[7] D. Haussler, M. Kearns, and R. Schapire, in *IVth Annual Workshop on Computational Learning Theory (COLT 91), Santa Cruz, 1991* (Morgan Kaufmann, San Mateo, CA, 1991), pp. 61–74.

[8] M. Opper, Phys. Rev. E **51**, 3613 (1995).

[9] G. J. Mitchison and R. M. Durbin, Biol. Cybern. **60**, 345 (1989).

[10] A. Engel, J. Phys. A **29**, L323 (1996).

[11] M Biehl and M. Opper, Phys. Rev. A **44**, 6888 (1991).

[12] M. Griniasty and T. Grossman, Phys. Rev. A **45**, 8924 (1992).

[13] A. Priel, M. Blatt, T. Grossman, E. Domany, and I. Kanter, Phys. Rev. E **50**, 577 (1994).

[14] B. Schottky, J. Phys. A **28**, 4515 (1995).

[15] R. Monasson and R. Zecchina, Phys. Rev. Lett. **75**, 2432 (1995).

[16] G. Cybenko, Math. Control Signals Systems **2**, 303 (1989).

[17] D. Saad and S. Solla, Phys. Rev. Lett. **74**, 4337 (1995).

[18] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988)

[19] M. Griniasty and H. Gutfreund, J. Phys. A **24**, 715 (1991).

[20] The application of this technique to a single perceptron has been extensively investigated in M. Bouten, J. Schietse, and C. van den Broeck, Phys. Rev. E **52**, 1958 (1995).

[21] M. Copelli, O. Kinouchi, and N. Caticha, Phys. Rev. E **53**, 6341 (1996).

[22] M. Bouten, J. Phys. A **27**, 6021 (1994).