Broken symmetries in multilayered perceptrons

E. Barkai
Department of Physics, Bar Ilan University, 52100 Ramat Gan, Israel

D. Hansel* and H. Sompolinsky
Racah Institute of Physics, Hebrew University, 91904 Jerusalem, Israel

(Received 19 June 1991)

The statistical mechanics of two-layered perceptrons with N input units, K hidden units, and a single output unit that makes a decision based on a majority rule (Committee Machine), is studied. Two architectures are considered. In the nonoverlapping case the hidden units do not share common inputs. In the fully connected case each hidden unit is connected to the entire input layer. In both cases the network realizes a random dichotomy of P inputs. The statistical properties of the space of solutions as a function of P is studied, using the replica method, and by numerical simulations, in the regime where $N \gg K$. In the nonoverlapping architecture with *continuously varying* weights the capacity, defined as the maximal number of P per weight, (α_c), is calculated under a *replica-symmetric* (RS) *ansatz*. At large K, α_c diverges as $K^{1/2}$ in contradiction with the rigorous upper bound, $\alpha_c < C \ln K$, where C is a proportionality constant, derived by Mitchison and Durbin [Biol. Cybern. **60**, 345 (1989)]. This suggests a strong replica-symmetry-breaking effect. The instability of the RS solution is shown to occur at a value of α which remains finite in the large-K limit. A one-step replica-symmetry-breaking (RSB) ansatz is studied for $K = 3$ and in the limit K goes to infinity. The results indicate that $\alpha_c(K)$ diverges with K, probably logarithmically. The occurrence of RSB far below the capacity limit is confirmed by comparison of the theoretical results with numerical simulations for $K = 3$. This symmetry breaking implies that unlike the single-layer perceptron case, the space of solutions of the two-layer perceptron breaks, beyond a critical value of α, into many disjoint subregions. The entropies of the connected subregions are almost degenerate, their relative difference being of order $1/N$. In the case of a nonoverlapping Committee Machine with *binary*, i.e., ± 1 weights, $\alpha_c \leq 1$ is an upper bound for all K. The RS theory predicts $\alpha_c = 0.92$ for $K = 3$ and $\alpha_c = 0.95$ for the large-K limit. The theoretical prediction (for $K = 3$) is in excellent agreement with the numerical estimate based on an exhaustive search in the space of solutions for small N. These results indicate that in the binary case there is no RSB in the space of solutions below the maximal capacity. In the fully connected architecture, the solution's phase space has a global permutation symmetry (PS) reflecting the invariance under permuting the hidden units. The order parameters that signal the spontaneous breaking of this symmetry are defined. The replica-symmetry theory shows that for small α the PS is maintained. For larger values of $\alpha < \alpha_c$ the symmetry is broken, implying the breaking of the solution space into disjoint regions. These regions are related by permutation symmetry, hence they are fully degenerate with respect to their entropies and statistical properties. This prediction has been tested by simulations of the $K = 3$ case, calculating the order parameters by random walks in the space of solutions. They yield good evidence for existence of a phase with broken permutation symmetry at values of $\alpha \geq 2$. Finally, both theory and simulations show that for a typical fully connected network the connections joining the same input to a pair of hidden units are negatively correlated.

PACS number(s): 87.10.+e, 05.50.+q, 64.60.Cn

I. INTRODUCTION

In her pioneering work, Gardner [1] has demonstrated that a statistical-mechanics approach can be helpful for studying properties of perceptrons. This approach has been applied successfully in different problems [2]. In most of them, the networks considered have the simplest architecture: one input layer of N units and one output unit.

However, as it is well known, the computational power of such a one-layer network is limited. Nonseparable problems can be implemented only if additional layers of hidden units are added. It is therefore of significant interest to investigate the statistical properties of multilayer perceptrons. Single-layer perceptrons are simple in that the space of solutions, when they exist, is convex. This is in general not true for multilayer networks. Thus the space of solution may be of complex shape, and in particular may consist of disjoint subregions in the network space. The connectedness of the solution space is investigated here by studying the occurrence of spontaneous symmetry breaking. Such a symmetry breaking signals the breaking of the solution space into disjoint regions, each one with reduced symmetry relative to the symmetry of the entire solution space.

We study symmetry-breaking phenomena in two-layer networks performing random dichotomies with two different architectures. In one architecture the connections

from the inputs to the hidden layer form nonoverlapping receptive fields, i.e., different hidden units do not share the same inputs, Fig. 1(a). The second architecture is one in which the hidden layer is fully connected, Fig. 1(b). In both the layer of connections from the hidden units to the output is fixed to be 1.

One type of symmetry breaking that can occur in the present problem is replica symmetry breaking (RSB)[3]. In spin glasses it is associated with the near degeneracy of the ground state due to the combination of randomness and frustration. In the case of the single-layer perceptron, RSB occurs only at or above the maximal capacity. This holds both for continuously varying weights [1] and with binary weights [4]. Below the capacity the symmetry is unbroken, which is in agreement with the convexity of the solution space. We will study the occurrence of RSB in the case of the nonoverlapping Committee Machine, below the maximal capacity.

A second symmetry studied in this work is permutation symmetry. It is relevant only for the fully connected architecture, where it reflects the invariance under permuting the hidden units. Of course this symmetry has no analog in the single-layer case.

Another aspect of the performance of the multilayer network is the *storage capacity*. As in the single-layer perceptron, it is defined here as the number of random dichotomies that the system can realize, per weight. In particular, it would be interesting to know the dependence of this capacity on the number of hidden units.

The problem of the information capacity of multilayered networks has been addressed, using geometrical methods, by Baum [5] and by Mitchison and Durbin [6]. Baum has obtained bounds on the smallest size of a multilayer network able to implement an *arbitrary dichotomy*. Using arguments based on a counting theorem of Cover [7], Baum shows that to implement an arbitrary dichotomy of P vectors in general position one needs a network of at least $P/\log_2 P$ weights.

Mitchison and Durbin have derived upper bounds of the capacity per synapse for *random* dichotomy of binary inputs for two networks: a fully connected two-layer Committee Machine and a fully connected Parity Machine. The output of the parity machine is the product of the outputs of the hidden units, whereas in the Committee Machine the output is computing the majority rule of the hidden-unit values. Mitchison and Durbin have

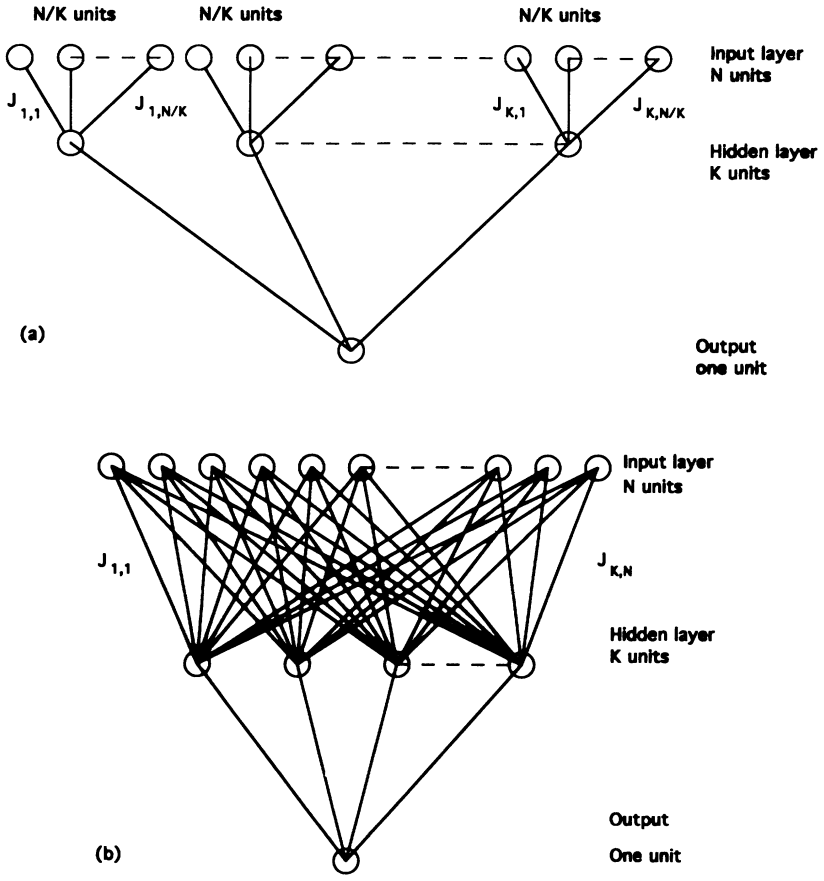

FIG. 1. The architectures of the two particular networks studied in this paper. In both networks the weights connecting the hidden units to the output are fixed and equal to one. (a) A Committee Machine with nonoverlapping receptive fields: the input layer has N units. Each of the K hidden units is connected to N/K inputs. There is no overlap between the receptive fields. (b) A fully connected Committee Machine: the input layer has N units. Each of these units is connected to the K hidden units.

shown that in both networks the maximal capacity per synapse is bounded by a function of the number of hidden units which behaves logarithmically as the number of hidden units goes to infinity.

Recently, Barkai, Hansel, and Kanter [8] have studied the parity machine with K hidden units using Gardner's statistical mechanical method. Their main results were that: (i) in the RS theory the capacity per synapse is proportional to K^2 at large K, violating the bound of Mitchison and Durbin, (ii) calculations based on a one-step RSB yield a capacity per synapse which behaves as $\log K/\log 2$ saturating the upper bound of Mitchison and Durbin for this network and suggesting that in the large-K limit the one-step RSB is actually exact as it is in the random-energy model [9] or in the so-called simplest-spin-glass model [10]. It should be noted that already for $K = 2$ the two-step RSB corrects only slightly, by less than 5%, the maximal-capacity estimates [11]. It is important that unlike the single-layer perceptron, in the parity machine the RSB occurs far below the capacity limit, indicating the complex form of the solution space. One aim of the present work is to study the analogous situation for the more interesting Committee Machine.

In Sec. II we introduce and study the nonoverlapping Committee Machine: Section II A introduces the architecture and the task of the machine. Sections II B and II C deal with networks of continuous synaptic weights having three and large number of hidden units, respectively. In Sec. II D we study the nonoverlapping network with binary weights. In Sec. III we study the fully connected network: Section III A introduces the model and the basic symmetries. In Sec. III B we study the occurrence of permutation symmetry breaking in the case of three hidden units. The results are summarized and discussed in Sec. IV.

II. COMMITTEE MACHINE WITH NONOVERLAPPING RECEPTIVE FIELDS

A. The model: architecture and task

We consider a two-layer feedforward neural network consisting of N binary input units [12], one hidden layer with K binary neurons and one single-output neuron. The input units are divided into K disjoint sets of N/K elements. All inputs of a set feed into the same single hidden neuron. Each hidden neuron receives input from only one set. The output of the network is obtained by a majority rule of the configurations in the hidden layer. This network ("Committee Machine [13] with nonoverlapping receptive fields") is shown in Fig. 1(a).

A configuration of the input layer will be denoted by $[S_{li}]$, $l = 1, ..., K$, $i = 1, ..., N/K$ with $S_{li}=\pm 1$. The local field h_l on the lth hidden unit is defined by

$$h_l = \sum_{i=1}^{N/K} J_{li} S_{li}, \tag{2.1}$$

where J_{li} is the value of the connection between the input unit (l, i), $l = 1, ..., K$, $i = 1, ..., N/K$, that belongs to the lth receptive field, and the hidden unit l. The configuration of the second layer is denoted as $[\sigma_l]$, $l = 1, ..., K$ where

$$\sigma_l = \text{sgn}\ (h_l) \tag{2.2}$$

[sgn(x) denotes the sign of x]. The connections between the hidden layer and the output are fixed and equal to one, and the output of the perceptron is simply given by

$$\sigma = \text{sgn}\left(\sum_{l=1}^{K} \sigma_l\right) . \tag{2.3}$$

We study the performance of the network in a task consisting of mapping a set of P input patterns $([\xi_{li}^{\mu}])$, $l = 1, ..., K$, $i = 1, ..., N/K$, $\mu = 1, ..., P$, $\xi_{li}^{\mu} = \pm 1$ onto a set of P respective outputs σ^{μ}. The mapping is assumed to be random, i.e., each of the input variables $[\xi_{li}^{\mu}]$, $l = 1, ..., K$, $i = 1, ..., N/K$, $\mu = 1, ..., P$, $\xi_{li}^{\mu} = \pm 1$ and the desired outputs $\sigma^{\mu} = \pm 1$ are chosen at random with equal probability of ± 1.

Our goal is the following:

(1) Calculate the capacity of the network, i.e., the maximum number of input-output pairs, P_c that can be stored in the system.

(2) Study the statistical properties of the space weights that store these mappings and the changes in these properties as the number of stored mappings increases.

Following Gardner's method [1] one formulates the problem in a statistical mechanics framework as follows. For a given realization of the P patterns we compute the volume of the subspace of the networks which realize the desired mapping. For continuously varying weights this volume is

$$V = \int_{-\infty}^{\infty} \prod_{i,l} dJ_{li} \prod_{l} \delta\left(\sum_{i} J_{li}^2 - N/K\right) \times \prod_{\mu} \Theta\left(\sigma^{\mu} \sum_{l} \sigma_l^{\mu}\right), \tag{2.4}$$

where $\sigma_l^{\mu} = \text{sgn}(h_l^{\mu})$, $h_l^{\mu} = \sum_{i=1}^{N/K} J_{li}\xi_{li}^{\mu}$. We imposed K normalization constraints

$$\sum_{i=1}^{N/K} J_{li}^2 = N/K, \quad l = 1, ..., K . \tag{2.5}$$

Note that since the hidden neurons are threshold elements, multiplying each set of the N/K connections J_{li} by an arbitrary positive constant does not change the output of the network. We will also consider (Sec. II D) the case where the weights are constrained to the values ± 1. In this case the integral over the J_{li} has to be replaced by a sum over all the 2^N configurations of weights.

The average over the patterns must be performed on $\ln V$. For that end one uses the replica trick, the details of which are given in Appendix A. The study of the statistical mechanics of networks with general K is difficult as the equations for the order parameters contain K-multiple integrals. In the following we concentrate on two simple cases: the case of $K = 3$ and the case of

large K ($K \to \infty$) where the K multiple integrals can be reduced to a single Gaussian integral.

B. Three hidden units and continuous weights

The smallest nontrivial Committee Machine has three hidden units. In this section, the results of the RS theory for this case are presented, the RSB instability point is located and the one-step RSB solution is studied.

1. Replica-symmetric theory

Using the results obtained in Appendix A within the RS ansatz for $[\ln V]$ ([] denotes average over the patterns), one obtains for network of continuous synaptic weights with three hidden units:

$$\frac{1}{N}[\ln V] = G_0 + \alpha G_1, \tag{2.6}$$

where

$$\alpha \equiv P/N \tag{2.7}$$

is the number of patterns per weight, which is assumed to be constant as $N \to \infty$. The functions G_0 and G_1 are

$$G_0 = \frac{1}{2}\left(E + q\hat{q} + \frac{\hat{q}}{E} - \ln E - \hat{q} + \ln(2\pi)\right) \tag{2.8}$$

and

$$G_1 = \int \prod_{l=1}^{3} Dv_l \ln\left(\Sigma_{(3)}\right), \tag{2.9}$$

$$\Sigma_{(3)} = H_1H_2 + H_1H_3 + H_3H_2 - 2H_1H_2H_3. \tag{2.10}$$

We have used the notation

$$H_l = H\left(\left(\frac{q}{1-q}\right)^{1/2} v_l\right), \tag{2.11}$$

where

$$H(x) = \int_x^\infty Dy. \tag{2.12}$$

The symbol Dy denotes a Gaussian measure:

$$Dy = \frac{dy}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right).$$

The quantities G_0 and G_1 depend on the order parameters q, $\hat{q}$, and E. The order parameter q is the analog of the Edwards-Anderson (EA) order parameter [3]

$$q = \left[\frac{K}{N}\sum_{i=1}^{N/K} \langle J_{li}\rangle^2\right], \tag{2.13}$$

where the brackets $< >$ denote average over all networks that realize the random mapping on the given set of patterns. The order parameters $\hat{q}$ and E are the conjugate parameters of q and of the normalization condition, respectively.

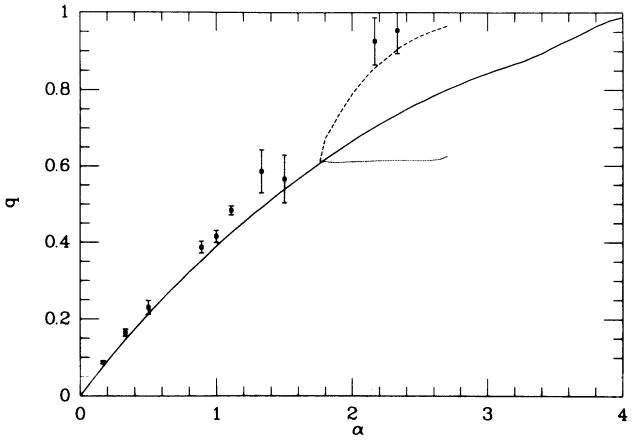

FIG. 2. The nonoverlapping architecture with three hidden units: the order parameter q as a function of α. The solid line is the prediction of the replica-symmetric theory. The dotted and the dashed lines are the first-step RSB calculation of q_0 and q_1, respectively. The points are the results of the simulations. The error bars represent the fluctuations in the measured q in the different samples.

Eliminating $\hat{q}$ from the saddle point equation [(A19) and (A20)] one obtains for the order parameter q:

$$\frac{q}{1-q} = \alpha \int Dv_l \left(\frac{H_1'}{H_1}\right)^2 \left(1 - \frac{H_2H_3}{\Sigma_{(3)}}\right)^2 \tag{2.14}$$

where

$$H_l' = -\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{q}{2(1-q)} v_l^2\right).$$

The graph of q as a function of α is plotted in Fig. 2 (solid line). Taking the limit $q \to 1$ of Eq. (2.14) yields the maximal capacity

$$\alpha_c = \frac{1}{\frac{5}{8} - \frac{3+\sqrt{3}}{4\pi}} \simeq 4.02. \tag{2.15}$$

An interesting quantity is the probability $\epsilon(\alpha)$ that presenting one of the αN patterns, one of the hidden units, say $l = 1$, is in a state $-\sigma$ (opposite to the state of the output unit). Thus, $\epsilon(\alpha)$ reflects the correlation between the state of each hidden unit and the output unit. In the case of $K = 3$, computing the field distribution $P(h_1, h_2, h_3)$ in the hidden layer and integrating it over h_2, h_3 and over $h_1\sigma \geq 0$ one obtains

$$\epsilon = \int Dv_1 Dv_2 Dv_3 (1 - H_1) \frac{H_2H_3}{\Sigma_{(3)}}. \tag{2.16}$$

Before training, i.e., for random J_{li}, $\epsilon = 0.25$. This is because for a given network output, say $\sigma = +1$, there are four equally probable configurations (+++,++−,+−+,−++). Indeed in Eq. (2.16), ϵ starts from

$$\epsilon(\alpha = 0) = 0.25 \tag{2.17}$$

increases monotonically with α. It reaches the value

$$\epsilon_c = \frac{7}{24} \tag{2.18}$$

at the critical capacity. The difference $\epsilon(\alpha) - 0.25$ measures the enhanced correlation between the state of the hidden unit and the output unit, reflecting the effect of learning.

2. *One-step replica-symmetry-breaking theory*

Stability analysis of the replica-symmetric solution for $K = 3$ is given in Appendix B. The RS solution is stable only for

$$\alpha \leq \alpha_{\mathrm{RSB}} \sim 1.76, \tag{2.19}$$

which corresponds to $q_{\mathrm{RSB}} \simeq 0.61$. Hence the RSB occurs for a number of patterns per synapse significantly smaller than the replica-symmetric capacity and the replica symmetry has to be broken.

In this subsection we evaluate the maximal capacity predicted by a one-step replica-symmetry-breaking scheme in the manner of Parisi [3]. In a general RSB ansatz, the ergodicity is broken leading to a decomposition of the Gibbs state into many pure states. In our case, which deals with zero-temperature statistical mechanics, breaking of replica symmetry means that the space of solutions breaks into disjoint subspaces of networks. Each subspace is a pure state. It enters in the expectation value of an observable O with a relative weight $P_a : < O >= \sum_a P_a < O >_a$, where a is a pure state index and P_a is

$$P_a = \frac{V_a}{V}. \tag{2.20}$$

V_a is the volume defined in Eq. (2.4) restricted to the a's pure state and V is the total volume of solutions. Here $<>_a$ corresponds to the averaging restricted to pure state a. Furthermore, RSB implies that the logarithm of the volumes V_a of the disjoint subspaces differ by quantities of order unity, hence their separate contributions must be taken into account even as $N \to \infty$.

Within the first-step RSB ansatz (detailed in Appendix B), the distribution function of the pure states is characterized by the following order parameters:

$$q_0 = \left[\sum_{i=1}^{N/K} \frac{K}{N} \langle J_{li} \rangle_a \langle J_{li} \rangle_b \right], \tag{2.21}$$

$$q_1 = \left[\sum_{i=1}^{N/K} \frac{K}{N} \langle J_{li} \rangle_a \langle J_{li} \rangle_a \right], \tag{2.22}$$

$$m = 1 - \sum_a [P_a]^2 \tag{2.23}$$

and the conjugates $\hat{q}_0, \hat{q}_1$, and E.

In term of these order parameters (and after elimination of the conjugate order parameters E_l^α and $\hat{q}_l^{\alpha,\beta}$) one finds

$$G_0 = \frac{1}{2}\left[\frac{1 + (m-1)\Delta q_1}{1 - q_1 + m\Delta q_1} + \ln 2\pi + \left(1 - \frac{1}{m}\right) \ln(1 - q_1) + \frac{1}{m} \ln(1 - q_1 + m\Delta q_1) \right], \tag{2.24}$$

$$G_1 = \int \prod_{l=1}^{3} Dv_l \ln \left(\int \prod_{l=1}^{3} Du_l (\tilde{\Sigma}_{(3)})^m \right), \tag{2.25}$$

where $\tilde{\Sigma}_{(3)} = \tilde{H}_1\tilde{H}_2 + \tilde{H}_1\tilde{H}_3 + \tilde{H}_2\tilde{H}_3 - 2\tilde{H}_1\tilde{H}_2\tilde{H}_3$. Here we have defined

$$\tilde{H}_l = H\left[\left(\frac{\Delta q_1}{1 - q_1} \right)^{1/2} u_l + \left(\frac{q_0}{1 - q_1} \right)^{1/2} v_l \right] \tag{2.26}$$

and $\Delta q_1 = q_1 - q_0$.

The three order parameters are determined by the three equations:

$$\frac{\partial G}{\partial q_0} = \frac{\partial G}{\partial q_1} = \frac{\partial G}{\partial m} = 0. \tag{2.27}$$

The numerical analysis of this ansatz is not easy in particular due to the multiple integrals of G_1. We have computed q_0, q_1, and m by minimizing G with respect to them. To this end we have used a standard minimization routine. The obtained values were then checked directly on the saddle point equations. The uniqueness of the RSB solution was also checked. The results for q_0 and q_1 are shown in Fig. 2 (dotted and dashed lines, respectively). As can be seen from this figure, the RSB occurs through a second-order phase transition, at α which coincides with the RS instability point given in Eq. (2.19). Above the transition, q_0 is almost constant and m decreases from 1. As $\alpha \to \alpha_c$, q_1 approaches 1, m decreases to 0, while q_0 remains at $q_0 = 0.63$. According to Eqs. (2.21), (2.22), and (2.23), at the RSB phase each of the pure states shrinks to a single solution at the criticality. It is found that

$$\alpha_c \simeq 3.0. \tag{2.28}$$

Our evaluation of the one-step correction to α_c is not very accurate due to numerical uncertainties. Nevertheless it does provide a rough estimate of α_c. Our conclusion is that for $K = 3$ the correction with respect to the RS estimate of the maximal capacity is of the order of 25%.

3. *Numerical simulations*

As there is no learning algorithm which is proved to converge for multilayered perceptrons, simulations can only give some empiric insight. We have used a Least Action Learning (LAL) algorithm of the type described by Nilsson [13, 6]. All the patterns are sequentially presented to the network. Presenting ξ^μ one updates the coupling constants as follows.

(1) If the Committee Machine gives the right answer (i.e., σ^μ) all the coupling constants remain unchanged

and the next pattern is presented.

(2) If the answer is wrong, the list of all the local fields h_l^μ $(l = 1, \ldots, K)$ in the hidden units is computed. The N/K coupling constants J_{li}, which correspond to the hidden unit l which has the local field easiest to improve (i.e., such that $\sigma^\mu h_l^\mu$ is the less negative) are updated with the standard perceptron algorithm [1] considering these N/K coupling constants between the input layer and the l hidden unit as a one-layer perceptron.

(3) One returns to step (1)

One presentation of the *whole set* of the patterns is named a session. The algorithm finishes when either all the patterns of the set are known or the number of sessions reachs some fixed $n_{\rm ses}$.

We have simulated networks of different sizes N: $N = 300$, $N = 450$, and $N = 600$. The results were averaged over a sample of, respectively, 50, 50, and 10 sets of patterns. The runs were performed over a maximum of $n_{\rm ses} = 3000$ sessions and some of the runs were allowed to keep over 6000 sessions.

The results of these simulations are summarized in Fig. 3. The fraction of successfully learned realizations in the sample is given. Taking as an empiric criterion the perfect learning of half of the sample one concludes that this Least Action Learning algorithm is efficient only up to

$$\alpha_{c(\rm LAL)} \simeq 2.42 \pm 0.02. \tag{2.29}$$

We also checked that approaching this range of α the learning time increases sharply.

An insight on the validity of the replica-symmetric solution can be obtained from computation of the order parameter by simulations. This can be performed by averaging the relevant quantities over a random walk in the space of solutions as follows. For a given network (N fixed) a realization of P patterns is first learned by the system using the Least Action Learning algorithm. A random walk starts at this point: at each time step the (old) set of weights $J_{li}^{\rm old}$ is changed by δJ_{li} chosen at random in $[-\delta, \delta]$ provided that $J_{li}^{\rm new} = J_{li}^{\rm old} + \delta J_{li}$ remains in the space of solutions for the given realization. Most of our simulations were performed on networks of sizes $N = 90$. Because of the finite size in a given realization of the patterns $q_{l=1}, q_{l=2}$, and $q_{l=3}$ are different. Hence, the quantity q to be compared with the theory was computed by averaging q_l over the three hidden units and also over different realizations of the patterns (5 to 10 in our simulations).

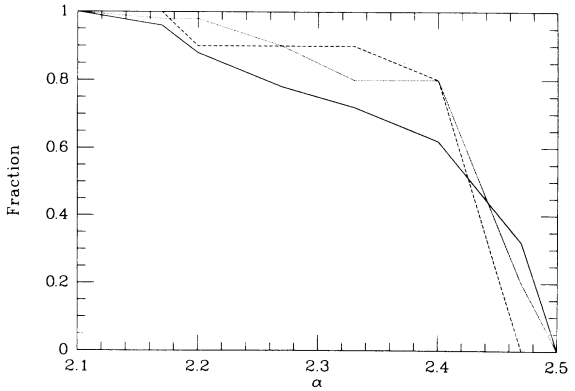

FIG. 3. The nonoverlapping architecture: results from numerical simulations of the Least Action Learning algorithm for different sizes N of the input layer and $K = 3$ hidden units. The lines give the ratio of the number of successfully learned realizations to the total size of the sample ($N = 300$ solid line, $N = 450$ dotted line and $N = 600$ dashed line). The runs were stopped after 3000 or 6000 sessions.

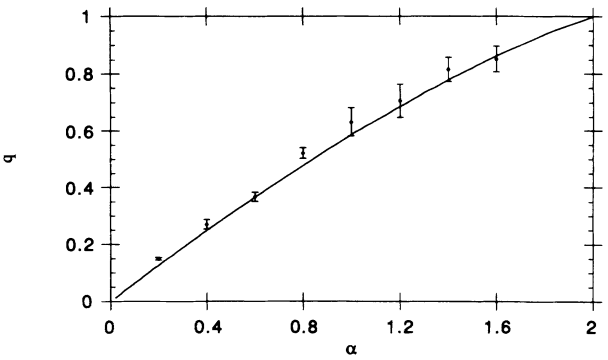

FIG. 4. The EA order parameter of the single-layer perceptron as a function of α, measured in zero-temperature Monte Carlo simulations.

The simulations' results are shown in Fig. 2. We have found that for $\alpha < 1.1$, δ can be quite big and the q calculated from the simulations agrees with the theoretical RS solution, even for small averaging time (not many random walk steps). As α increases, even not so close to the RSB transition, the phase space appears to be complex. In particular, very narrow corridors connect different regions of the phase space and a small δ must be chosen to ensure a sufficiently good sampling of the whole solutions space. Such a small δ imposes large averaging time to explore a significantly large part of the solutions space. The need for the long averaging time is a signature of the vicinity of the RSB transition, where the corridors connecting different regions of the solutions space shrink to zero, and disconnected regions in solutions space appear. As can be seen from Fig. 2, for $\alpha > \alpha_{\rm RSB}$, the EA order parameter q is close to the theoretical order parameter q_1 calculated with the assumption of the one-step RSB ansatz.

For comparison the results of similar numerical experiments for a one-layer perceptron, where RSB does not occur and the solutions space is connected and convex, are presented in Fig. 4.

C. Network with continuous synapses and large K

1. Replica-symmetric theory

We now turn to the case where $K \to \infty$. In that limit the RS result for G_1 is

$$G_1 = \int_{-\infty}^{\infty} Dx \ln H \left[\left(\frac{q_{\rm eff}}{1 - q_{\rm eff}} \right)^{1/2} x \right], \tag{2.30}$$

where

$$q_{\rm eff} = 1 - \frac{2}{\pi} \arccos q \tag{2.31}$$

(see Appendix A). It should be noticed that G_1 in the limit of large K is similar to its expression for $K = 1$ [1] provided q is replaced by $q_{\rm eff}$ which takes into account the interaction between different hidden units. As far as G_1 is concerned the network is equivalent to a one-layer perceptron with effective order parameters. This property remains true for a one-step replica-symmetry-breaking ansatz. From this expression one obtains the relation between q and the number of patterns per weight α:

$$\alpha = \frac{\pi}{2} \left(\frac{1+q}{1-q} \right)^{1/2} \left(\frac{q}{1-q} \right) (1 - q_{\rm eff}) \langle W^2 \rangle^{-1}, \tag{2.32}$$

where $< >$ denotes Gaussian average over x, and where

$$W = -\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{q_{\rm eff}}{2(1-q_{\rm eff})} x^2 \right) \times \left\{ H\left[\left(\frac{q_{\rm eff}}{1-q_{\rm eff}} \right)^{1/2} x \right] \right\}^{-1}. \tag{2.33}$$

The graph of q as function of α is shown in Fig. 5 (solid line). For $q \to 1$ one finds that α goes to infinity. A change of the order of the limits (taking first the limit $q \to 1$ and then the limit $K \to \infty$), leads to the replica-symmetric capacity

$$\alpha_c \sim \left(\frac{72}{\pi} \right)^{1/2} K^{1/2}. \tag{2.34}$$

This expression obtained by estimation of $\alpha_c(K)$ calculated in Appendix C in the case of large K. Numerical evaluation of $\alpha_c(K)$ for finite K shows that this asymptotic behavior is reached rapidly (already for $K \simeq 15$). This power law for α_c contradicts the rigorous bound

$$\alpha_c < c \ln K / \ln 2 \tag{2.35}$$

(c is some constant independent of N and K), that was obtained by extending the arguments of Mitchison and Durbin [6] to the nonoverlapping receptive field's architecture. The details of the derivation of this bound are given in Appendix D.

The field distribution in the hidden layer can be evaluated in the large-K limit. In particular one finds that at the first nontrivial order, the probability ϵ that presenting one of the αN patterns a hidden unit is in a state of opposite sign than the output unit, is given by

$$\epsilon = \frac{1}{2} - \frac{\eta}{\sqrt{2\pi K}}, \tag{2.36}$$

where η is going from 1 to $\frac{1}{2}$ while q is going from 0 to 1. This implies that correlations built between the hidden units and the output unit, are seen in changes of order $1/\sqrt{K}$ in ϵ, starting from

$$\epsilon = \frac{1}{2} - \frac{1}{\sqrt{2\pi K}}$$

for a random network.

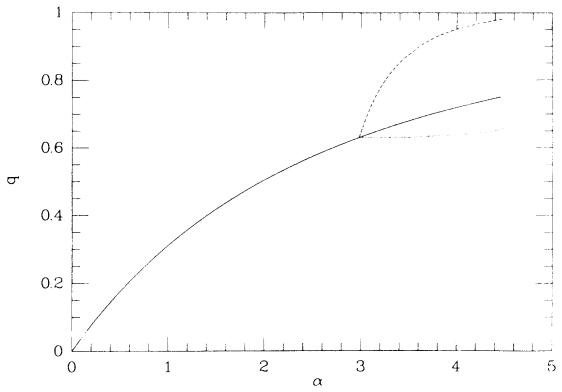

FIG. 5. The nonoverlapping architecture with $K \to \infty$: the order parameter q as a function of α. The solid line is the prediction of the replica-symmetric theory. The dotted and the dashed lines show q_0 and q_1 calculated within the first-step RSB ansatz.

2. *One-step replica-symmetry-breaking theory*

The discrepancy between the power-law divergence of α_c at large K and the bound of Mitchison and Durbin suggests a strong replica-symmetry breaking. Indeed, the RSB instability occurs at *a finite number of patterns per synapse*:

$$\alpha \leq \alpha_{\rm RSB} \simeq 2.95 \tag{2.37}$$

which corresponds to $q_{\rm RSB} \simeq 0.62$.

Within the first step RSB ansatz, and for large K [for $K(1-q)$ large], G_1 can be written in the form

$$G_1^{\rm RSB} = \frac{1}{m} \int_{-\infty}^{\infty} DR \ln \left\{ \int_{-\infty}^{\infty} DS \left[H\left(\left(\frac{\Delta q_{1\rm eff}}{1-q_{1\rm eff}} \right)^{1/2} S + \left(\frac{q_{0\rm eff}}{1-q_{1\rm eff}} \right)^{1/2} R \right) \right]^m \right\}. \tag{2.38}$$

$q_{0\rm eff}$ and $q_{1\rm eff}$ are effective order parameters given by

$$q_{0\rm eff} = 1 - \frac{2}{\pi} \arccos q_0, \tag{2.39}$$

$$q_{1\rm eff} = 1 - \frac{2}{\pi} \arccos q_1 \tag{2.40}$$

with

$$\Delta q_{1\rm eff} = q_{1\rm eff} - q_{0\rm eff}. \tag{2.41}$$

Like in the RS case $G_1^{\rm RSB}$ is similar to the expression one has for $K = 1$ provided one replaces the order parameters

by the effective ones. In the limit $q \to 1$, α diverges showing that also at the one-step RSB the capacity *per synapse* is infinite when $K \to \infty$. Numerical solution of the saddle point equations is shown in Fig. 5 (q_0 dotted line and q_1 dashed line). It shows that the RSB transition is of second order and occurs at the point predicted by the instability analysis keeping the dominant contribution in $1/K$: $q_1 = q_0 = 0.62, \alpha_{\rm RSB} = 2.95$. As in the $K = 3$ case, here also q_0 varies very slowly above the RSB transition and is equal roughly to 0.63, which implies shrinking of each of the pure states to zero volume near the criticality.

The determination of the asymptotic behavior of α_c would necessitate resumming the most divergent terms in $G_1^{\rm RSB}$. This would be very interesting but unfortunately this seems a very difficult task.

D. Binary synapses: $K = 3$ and large K

Capabilities of networks built with discrete weights is of importance in particular for hardware realizations. It has been shown recently that the one-layer perceptron with binary synaptic weights ($J_{lj} = \pm 1$) [4], displays properties that recall the random-energy model and in particular a phase exists where the system is frozen. This section is devoted to the generalization of this result to nonoverlapping Committee Machine architecture.

Unlike the case of continuous synaptic weights, for discrete synapses the maximal capacity is the α_c at which the entropy vanishes [4, 14]. This value is bounded from above by $\alpha_{\rm ann}$ at which the entropy of the corresponding annealed network vanishes. It is easy to see, that for the nonoverlapping Committee Machine this value is

$$\alpha_{\rm ann} = 1. \tag{2.42}$$

This value does not depend on the internal structure of the committe machine, i.e., the number of the hidden units.

In the calculation of the capacity itself, assuming a replica-symmetric ansatz, G_0 given in Eq. (A6) reduces to

$$G_{0\rm binary} = -\hat{q}(1-q)/2 + \int_{-\infty}^{\infty} Dt \ln[2\cos(\sqrt{\hat{q}}t)]. \tag{2.43}$$

In a network with three hidden units ($K = 3$) one finds that the replica-symmetric entropy vanishes at

$$\alpha_c \simeq 0.92 \quad \text{and} \quad q_c \simeq 0.38. \tag{2.44}$$

The replica-symmetric saddle point remains stable at α_c and this value is a good candidate for an estimate of the maximal capacity. Unlike the continuous synaptic weights case, where the solutions volume shrinks to zero when $q \to 1$, here the solutions volume shrinks to zero for finite q ($q < 1$). As α approaches α_c the volume of solutions decreases and the structure of the space becomes complex, though it remains connected. At $\alpha = \alpha_c$, this volume is zero leaving a nonextensive number of disconnected solutions, which causes a complete freezing of the synapses (when looking on the synapses as dynamic variables). In order to confirm this identification, we have analyzed the first-step RSB solution of the problem, which yields the vanishing of the RS entropy as a condition for the criticality.

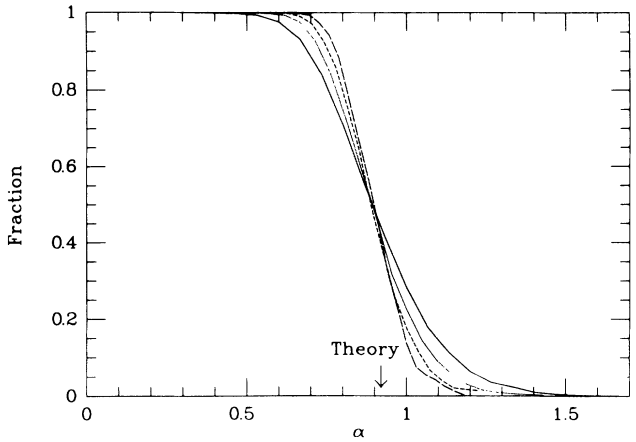

FIG. 6. The fraction of success in exhaustive search of solutions for a $K = 3$ nonoverlapping Committee Machine with binary type weights. The different curves correspond to networks of sizes: $N = 15$ solid line, $N = 21$ dotted line, $N = 27$ short dashed line and $N = 33$ long dashed line. The search was done with 100 to 10000 samples for each point.

To confirm the theoretical results, we have performed an exhaustive search on binary networks with three hidden units. Random samples of P binary input vectors and outputs were chosen. For each sample all solutions were found by an exhaustive search in the space of binary networks. We have performed the search on networks of sizes $N = 15, 21, 27$, and 33. The results are displayed in Fig. 6. We define the average capacity as the value of α for which half of the samples have solutions. As can be seen from the figure, the average capacity found for these small networks is almost independent of N (displaying only a minor finite size effect), and is close to the theoretical capacity found for the infinite network ($N \to \infty$).

For K approaching infinity, the replica-symmetric entropy vanishes at

$$\alpha_c \simeq 0.95 \quad \text{and} \quad q_c \simeq 0.31\ . \tag{2.45}$$

The replica-symmetric solution remains stable at α_c, and the scenario at the criticality is the same as in the three hidden units network. Comparing these results with the capacity for $K = 1$ [4, 14] ($\alpha_c \simeq 0.83$) one concludes that the improvement of the performance of this multilayer network is very small for binary weights. This is not so surprising as the annealed approximation put a strong upper bound for α_c, namely, $\alpha_c \leq 1$.

III. FULLY CONNECTED COMMITTEE MACHINE

A. The model and the basic symmetries

We consider the fully connected network with N binary inputs and one binary output [shown in Fig. 1(b)]. The

hidden layer consists of K binary units, each connected to all the input units. The weights connecting the hidden layer with the output are fixed and equal to one. The solution space has a global permutation symmetry. If the $K \times N$ matrix $[J_{li}], l = 1,...,K; i = 1,...,N$, is a solution for a given realization of the patterns, so are the matrices $[J'_{li}]$, $J'_{li} \equiv J_{\Sigma(l)i}, l = 1,...,K; i = 1,...,N$, for all Σ in S_K, where S_K is the permutation group of the hidden units.

The present system is described by the following $K \times K$ order parameter matrix:

$$Q_{l,l'} = \left[\frac{1}{N}\sum_i < J_{li} >< J_{l'i} >\right]. \tag{3.1}$$

Since on the average there is no statistical difference between the different hidden units, the matrix has the form

$$Q_{l,l'} = q_2\delta_{l,l'} + q_0(1 - \delta_{l,l'}). \tag{3.2}$$

An additional important order parameter is

$$q_1 = \left[\frac{1}{N}\sum_i < J_{li}J_{l'i} >\right], \quad l \neq l' \tag{3.3}$$

this order parameter measures the average correlations between a pair of connections that share the same input in a given solution. Note that in contrast to q_2, which is necessarily positive, the order parameters q_0 or q_1 can be either positive or negative. In the latter case the ordering between the hidden units is of antiferromagnetic type.

1. *Permutation symmetric phase*

In the permutation symmetric phase the solutions that are related by permutation symmetry are part of a single connected space of solutions, and are therefore included in the averaging denoted by $< >$. Therefore,

$$< J_{li} >= \frac{1}{K!}\sum_{S_K} < J_{\Sigma(l)i} >$$

is independent of the hidden unit index (l). Thus, in this phase

$$q_0 = q_2 = \left[\frac{1}{N}\sum_i \left\langle \frac{1}{K}\sum_{l=1}^{K} J_{li}\right\rangle^2\right]. \tag{3.4}$$

At a (permutation symmetric) critical capacity, α_c, the solution is unique up to a permutation of the weights between the K hidden units. Thus, as $\alpha \to \alpha_c$, $< J_{li} >$ $< J_{l'i} > \to < J_{li}J_{l'i} >$. Hence, by Eq. (3.4)

$$q_0 \to \frac{1}{K}\left[1 + (K-1)q_1\right].$$

This establishes the relation at the critical capacity

$$1 - q_1 + K\Delta q_1 = 0, \quad \alpha \to \alpha_c, \tag{3.5}$$

where $\Delta q_1 = (q_1 - q_0)$.

2. *Phase with broken permutation symmetry*

In the phase with permutation symmetry breaking (PSB), solutions which are related by a permutation of the hidden units belong to different disconnected parts of the solution space and are averaged separately. In this phase q_2 is not equal to q_0 and their difference

$$q_2 - q_0 = \left[\frac{1}{N}\sum_i < J_{li} >^2\right] - \left[\frac{1}{N}\sum_i < J_{li} >< J_{l'i} >\right], \quad l \neq l' \tag{3.6}$$

measures the degree of breaking of the PS. When the maximal capacity is reached, the solution is unique, hence

$$q_1 = q_0, \quad q_2 = 1, \quad \alpha \to \alpha_c. \tag{3.7}$$

B. Three hidden units

1. *Symmetric solution*

Assuming a symmetric phase with the two order parameters q_0 and q_1, one obtains for a network with three hidden units the following saddle point equations, derived by replica methods similar to that of Appendix A:

$$\frac{q_0}{(1 - q_1 + 3\Delta q_1)^2} = \frac{12\alpha}{1 - q_1}\int_{-\infty}^{\infty} Du\left(\frac{\langle HH'(1-H)\rangle}{\langle 3H^2 - 2H^3\rangle}\right)^2, \tag{3.8}$$

$$\frac{3\Delta q_1^2 + q_1(1-q_1)}{(1 - q_1 + 3\Delta q_1)^2} = 3\alpha\int_{-\infty}^{\infty} Du\frac{\langle H'^2(1-2H)\rangle}{\langle 3H^2 - 2H^3\rangle}. \tag{3.9}$$

This determines the order parameters as function of the capacity per synapse, $\alpha \equiv P/(NK)$. We have introduced the following notation:

$$H = H(au + bt),$$

$$H' = H'(au + bt) = -\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(au+bt)^2}{2}\right)$$

and the average of $f(t)$ with respect to the Gaussian measure Dt is denoted by $< f(t) >$. The coefficients a and b are $a = \sqrt{q_0/(1-q_1)}$, $b = \sqrt{(q_1 - q_0)/(1-q_1)}$.

Expanding these equations at small α one finds the following behavior:

$$q_0 = \frac{3\alpha}{2\pi} + O(\alpha^2), \tag{3.10}$$

$$q_1 = -9\frac{\alpha}{4\pi^2}\left(1 - \frac{2}{\pi}\right) + O(\alpha^3). \tag{3.11}$$

The complete numerical solution of Eqs. (3.8) and (3.9) is plotted in Fig. 7. The order parameter q_1 is always neg-

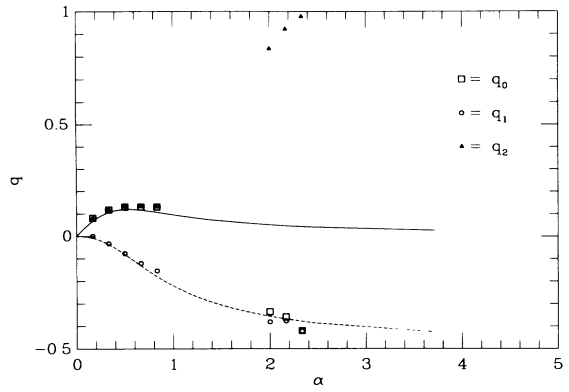

FIG. 7. The fully connected architecture ($K = 3$): q_0 (solid line) and q_1 (dashed line) as a function of α, within the RS and PS ansatz ($q_0 = q_2$). The squares, the circles, and the triangles represent simulation results for q_0, q_1, and q_2, respectively, indicating existence of PSB phase.

ative and decreases monotonicly to -0.5 as α increases, while q_0 reaches a maximum value $q_0 \simeq 0.12$ for $\alpha \simeq 0.57$. Note that q_1 is negative showing that the interaction between the different hidden units is of antiferromagnetic type. This can be understood qualitatively: because of the majority rule of the hidden-unit states performed by the output unit, most of the internal configurations that satisfy the task are such that two hidden units are antiparallel; as the receptive fields of all the hidden units are the same, this results in the anticorrelation of the coupling constants in most of the networks that realize the task.

Another important feature of the permutation and replica-symmetric solution is that *no critical capacity exists for this solution.* Application of the criticality criterium for a PS phase, namely, $1 - q_1 + 3\Delta q_1 = 0$, gives $\alpha_c \to \infty$, with $q_0 \to 0$ and $q_1 \to -\frac{1}{2}$. This *absence of a maximal capacity for the ergodic solution* implies that this solution is not valid everywhere, especially for large α.

2. *Numerical evidence for PSB*

In order to learn about the ergodicity in the network's solutions space, and possible breaking of the PS, we have calculated from simulations (Monte Carlo) the EA order parameters q_0, q_1, and q_2. To this end, we have used the zero temperature dynamics described in Sec. II B 3 for the nonoverlapping receptive fields networks.

Our simulations were performed on networks of size $N = 20$, and the results are shown in Fig. 7. It was found, that for small α ($\alpha < 0.8$) the system is PS, ($q_0 = q_2$), with small deviations from the PS and RS theoretical predictions due to finite-size effects. For $0.8 < \alpha < 2.0$ we were not able to determine whether the phase is PS or not, because of the diverging relaxation times. For $\alpha > 2.0$, the network is in a PSB phase, with q_0 negative a little bit larger than q_1 and with q_2 close to 1. Finally, as α is increased, $q_2 \to 1$ while $q_0 \to q_1$, which corresponds to the vicinity of the criticality.

To emphasize the distinction between the two phases, we show a typical time evolution of the averaging of q_0, q_1, and q_2, in those two phases (Fig. 8). By definition, at the beginning of the simulation, $q_2 = 1$ and $q_0 = q_1$, and in most of the runs q_0 and q_1 both started from negative value. As the number of averaging steps is increased, q_2 decreased from 1 and q_0 diverges from q_1 having a larger value than q_1. For small α ($\alpha = 0.5$), q_0 increases to positive values, and becomes equal to q_2, which indicate a PS phase [Fig. 8(a)]. For large α ($\alpha = 1.83$), q_0 increases, but remains negative near q_1, and q_2 stays in values near 1. This indicates a PSB phase [Fig. 8(b)].

To estimate the capacity of the fully connected networks, we have used the same algorithm as in the nonoverlapping receptive fields networks. We have simulated networks of different sizes N: $N = 100$, $N = 150$, and $N = 200$ averaging the results over samples of 10 or 50 realizations. The runs were performed over a maximum of $n_{\text{ses}} = 3000$ sessions and some of the runs were allowed to keep over 6000 sessions.

The results of these simulations are summarized in Fig. 9. As α grows and approaches $\alpha \simeq 2.4$ the learning time is rapidly increasing. This indicates the vicinity of the maximal capacity (of this algorithm). From the empiric criterion for the maximal capacity (the perfect learning of half of the sample) one concludes that this LAL algorithm

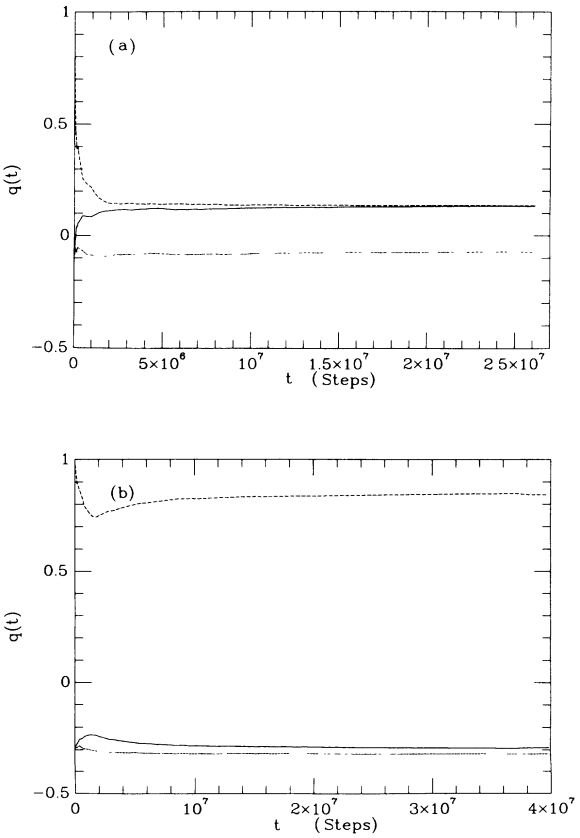

FIG. 8. The fully connected architecture: Typical relaxation of the order parameters q_0 (solid line), q_1 (dotted line), and q_2 (dashed line) as function of the averaging time for (a) $\alpha = 0.5$ (PS phase); (b) $\alpha = 1.83$ (PSB phase).

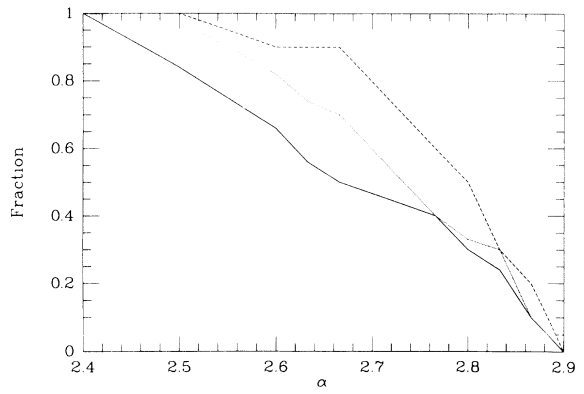

FIG. 9. The fully connected architecture: results from numerical simulations of the Least Action Learning algorithm for different sizes N of the input layer and $K = 3$ hidden units. The lines give the ratio of the number of successfully learned realizations to the total size of the sample ($N = 100$ solid line, $N = 150$ dotted line, and $N = 200$ dashed line). The runs were stopped after 3000 or 6000 sessions.

is no more efficient for a number of patterns per synapse larger than

$$\alpha_{c(\mathrm{LAL})} \simeq 2.82 \pm 0.02, \tag{3.12}$$

which is slightly larger than the corresponding quantity for the nonoverlapping architecture $\alpha_{c(\mathrm{LAL})} \simeq 2.42$.

IV. DISCUSSION AND CONCLUSIONS

In this work we have studied the properties of the space of solutions of two-layered neural networks that perform random dichotomy. We have focused on the occurrence of spontaneous breaking of two symmetries: replica symmetry (RS) and permutation symmetry (PS). The occurrence of symmetry breaking is an important probe of the connectedness of the space of solutions.

In order to study the RSB we have investigated a two-layer network with an architecture of a Committee Machine with nonoverlapping receptive fields [shown in Fig. 1(a)]. In this architecture the solution space does not posses any global symmetry. Thus, the only relevant symmetry in this case is the RS. Our main finding is that in the two-layer case (with continuously varying weights) replica symmetry breaking occurs at values of α far below the capacity. This phenomenon is particularly strong in the case of large K, i.e., a large number of hidden units. Our analysis of the limit $K \to \infty$ (keeping N much bigger than K) suggests that the maximal capacity diverges with K and is bounded by $\ln K$. However we have found that RSB occurs already at finite α. Figures 2 and 5 show the order parameters that signal the breaking of RS in networks with three and many hidden units, respectively.

RSB has been found in the case of a single-layer perceptron [1]. In that case RSB occurs only for α above the maximal capacity, α_c, i.e., in the regime where the networks perform the dichotomy with a nonzero amount of error. (Note that α is defined as the number of patterns per weight.) Thus the RSB is similar to that occurring in spinglasses, in that it is related to the degeneracy and fluctuations of the frustrated ground state. In the present case, the RSB is related to the breakdown of the solution space, which breaks into many disconnected regions in the space of networks. Since our study is strictly at zero temperature the RSB is associated not with near-energy degeneracy but rather with near-entropy degeneracy. Specifically the RSB implies in our case that the entropies of the different "valleys" differ by an amount of order 1, whereas the total entropy is of the order of the number of connections in the network (N).

We have also studied the case of binary weights. In this case, for general K the annealed approximation provides a bound of $\alpha_c = 1$ on the maximal capacity. Our analysis shows that RSB occurs only above α_c. Below it the space of solutions is connected. However as α approaches α_c the solution space becomes increasingly more ramified. Thus the entropy of the solution space approaches zero as $\alpha \to \alpha_c$ but the average overlap between a pair of solutions remains smaller than 1. Our simulations calculating the maximal capacity of networks with three hidden units are given in Fig. 6. They are in excellent agreement with the prediction based on the replica-symmetric theory.

It is interesting to note that in the case of binary weights the properties of the two-layer system discussed above are qualitatively similar to those of a single-layer binary perceptron, as found by Krauth and Mézard [4]. In both cases the RSB occurs only at the capacity limit.

To probe the effect of PS and its breaking in multilayer networks we have studied a two-layer Committee Machine with fully connected architecture, Fig. 1(b). In this network each of the hidden units has a full receptive field of the input pattern. Due to this structure, the solutions space has a global permutation symmetry: Permutation of the hidden units in a given solution yields another, completely equivalent solution. In a PS state the solutions that are related by permutation transformation belong to the same connected region of solutions. Breaking of this symmetry implies that these equivalent solutions reside in different, disjoint regions. This symmetry breaking is measured by the order parameter $q_2 - q_0$, defined in Eq. (3.6).

To investigate PSB we have calculated numerically the order parameters q_0, q_1, and q_2, Eqs. (3.2) and Eq. (3.3), by averaging over solutions that were generated by a random walk in a single "valley" of the solution space. We have found that for small α the PS is unbroken. This is shown in the results of Fig. 7, where $q_2 - q_0 = 0$ for small values of α. For large values of α, a strong PSB occurs with q_2 being close to 1 and q_0 approaches $q_1 < 0$. The evolution of the order parameters with the averaging time is shown in Fig. 8, where the different behavior in the PS and the PSB regimes is clearly demonstrated.

The fully connected architecture also provides an opportunity to study the correlations between the values of different weights of the network, in a given solution. In general, correlations between the different parts of the network may arise from the underlying structure of the task, e.g., spatial correlations in the inputs. In our case, of random dichotomies correlations exist only between weights (connecting the input and hidden layers) that

share the same inputs. Both the theoretical analysis (for small α) and the numerical results, Figs. 7 and 8, show that these average correlations are negative for all values of α. These antiferromagnetic correlations are a property of the *typical* solutions. It does not exclude the existence of solutions with positive overlaps of these weights. Although our study of the fully connected network has focused on the breaking of permutation symmetry, we expect that the system also exhibits replica-symmetry breaking for large values of α. A systematic theoretical study of the properties of this system in the broken symmetry states remains a difficult challenge.

ACKNOWLEDGMENTS

E.B. acknowledges the warm hospitality of C.P.T. Ecole Polytechnique (Palaiseau, France) where part of this work was done. We thank S. Amari for a useful discussion and for drawing our attention to the work of Mitchison and Durbin, Ref. [6]. Helpful discussions with E. Domany, D. S. Fisher, M. Griniasty, T. Grossman, H. Gutfreund, and D. Huse are acknowledged.

APPENDIX A: THE REPLICA THEORY OF THE NONOVERLAPPING COMMITTEE MACHINE

1. Replica theory

In order to calculate the average of $\ln V$ [defined in Eq. (2.4)] over the patterns, one uses the replica trick based on the identity:

$$[\ln V] = \lim_{n\to 0} \frac{[V^n]-1}{n} \tag{A1}$$

and a performance of a continuation from positive integer n, where n is the number of replicas of the network, to $n \to 0$. The [] denotes the average over the disorder (i.e., over different realizations of the set of P patterns). In the large-N limit and for finite K, the average of these n replicas of the network, has the form

$$[V^n] = \int \prod_{\substack{\alpha,\beta,l \\ \alpha<\beta}} dq_l^{\alpha,\beta} d\hat{q}_l^{\alpha,\beta} \prod_{\alpha,l} dE_l^\alpha \exp[nNG(\{q_l^{\alpha,\beta}, \hat{q}_l^{\alpha,\beta}, E_l^\alpha\})]. \tag{A2}$$

The function G is the sum of two contributions:

$$G(\{q_l^{\alpha,\beta}, \hat{q}_l^{\alpha,\beta}, E_l^\alpha\}) = G_0(\{\hat{q}_l^{\alpha,\beta}, E_l^\alpha\}) + \alpha G_1(\{q_l^{\alpha,\beta}\}), \tag{A3}$$

where

$$\alpha = \frac{P}{N} \tag{A4}$$

is the number of patterns per weight. The function G_0 depends on the possible constraints imposed on the weights variables. For continuous weights:

$$G_0 = -\frac{1}{nK} \sum_{\substack{\alpha,\beta,l \\ \alpha<\beta}} \hat{q}_l^{\alpha,\beta} q_l^{\alpha,\beta} + \frac{1}{nK} \sum_l \ln \left[\int \prod_\alpha dJ_l^\alpha \exp\left(-\sum_\alpha E_l^\alpha (J_l^{\alpha 2} - 1) + \sum_{\substack{\alpha,\beta \\ \alpha<\beta}} \hat{q}_l^{\alpha,\beta} J_l^\alpha J_l^\beta \right) \right] \tag{A5}$$

and for binary weights, $J_{lj} = \pm 1$:

$$G_{0\text{binary}} = -\frac{1}{nK} \sum_{\substack{\alpha,\beta,l \\ \alpha<\beta}} \hat{q}_l^{\alpha,\beta} q_l^{\alpha,\beta} + \frac{1}{nK} \sum_l \ln \left[\prod_\alpha \left(\sum_{J_l^\alpha = \pm 1} \right) \exp\left(\sum_{\substack{\alpha,\beta \\ \alpha<\beta}} \hat{q}_l^{\alpha,\beta} J_l^\alpha J_l^\beta \right) \right] \tag{A6}$$

G_1 is given by

$$G_1 = \frac{1}{n} \ln \int \prod_{\alpha,l} d\lambda_{\alpha,l} \frac{dx_{\alpha,l}}{2\pi} \prod_\alpha \Theta\left(\sum_l \mathrm{sgn}(\lambda_{\alpha,l}) \right) \exp\left(-\frac{1}{2} \sum_{\alpha,l} (x_{\alpha,l}^2 - 2ix_{\alpha,l}\lambda_{\alpha,l}) - \sum_{\substack{\alpha,\beta,l \\ \alpha<\beta}} q_l^{\alpha,\beta} x_{\alpha,l} x_{\beta,l} \right), \tag{A7}$$

which is independent of the nature of the weights. The order parameters $q_l^{\alpha,\beta}$ are related to the weights by

$$q_l^{\alpha,\beta} = \frac{K}{N} \sum_i J_{li}^\alpha J_{li}^\beta. \tag{A8}$$

The parameters $\hat{q}_l^{\alpha,\beta}$ $(\alpha,\beta = 1,...,n; l = 1,...,K)$ are the conjugate variables to the order parameters $q_l^{\alpha,\beta}$, while E_l^α are the conjugate variables that impose the constraints (2.5). In the thermodynamic limit $[V^n]$ is computed by the saddle point method. Performing the limit $n \to 0$ leads to

$$\frac{1}{N}[\ln V] = \mathrm{extr}_{q_l^{\alpha,\beta}, \hat{q}_l^{\alpha,\beta}, E_l^\alpha} G\left(\{q_l^{\alpha,\beta}, \hat{q}_l^{\alpha,\beta}, E_l^\alpha\} \right). \tag{A9}$$

The extremum is taken over the $q_l^{\alpha,\beta}, \hat{q}_l^{\alpha,\beta}$, and E_l^α. In order to solve the saddle-point equations a more specific ansatz has to be done on the symmetries of the order-

parameter matrices.

We first address the dependence of the order parameters in the saddle point on the hidden unit index l. Since the distribution of the input pattern is the same in each of the hidden units, averages are to be independent of l,

$$q_l^{\alpha,\beta} = q^{\alpha,\beta}, \quad \hat{q}_l^{\alpha,\beta} = \hat{q}^{\alpha,\beta}, \quad E_l^{\alpha} = E^{\alpha}. \tag{A10}$$

In the next paragraph we discuss the dependence of the order parameters on the replica indices.

In the next sections of this appendix and in the following appendices we discuss only the case of continuous synaptic weights.

2. Replica-symmetric theory for general K

In this part one assumes the replica symmetry of the saddle point, i.e., that at the saddle point for all l:

$$q_l^{\alpha,\beta} = q, \quad \hat{q}_l^{\alpha,\beta} = \hat{q} \quad \text{for all } \alpha \neq \beta,$$
$$E_l^{\alpha} = E \quad \text{for all } \alpha. \tag{A11}$$

The order parameter q is the analog of the EA order parameter:

$$q = \left[\frac{K}{N} \sum_{i=1}^{N/K} \langle J_{li} \rangle^2 \right], \tag{A12}$$

where the brackets $< >$ denote average over all networks that realize the given set of patterns.

Under the replica symmetry assumption G_0 and G_1 are

$$G_0 = \frac{1}{2} \left(E + q\hat{q} + \frac{\hat{q}}{E} - \ln E - \hat{q} + \ln(2\pi) \right), \tag{A13}$$

$$G_1 = \int \prod_{l=1}^{K} Dv_l \ln \Sigma(\{v_l\}), \tag{A14}$$

where

$$\Sigma(v_l) = \sum_{l=1}^{(K+1)/2} \sum_{\sigma(\in S_K)} \prod_{j=1}^{l} H_{\sigma(j)} \prod_{j=l+1}^{K} (1 - H_{\sigma(j)}). \tag{A15}$$

σ is a permutation element of the group of permutation of the K indices, denoted S_K. For instance for $K = 3$,

$$\Sigma_{(3)}(\{v_l\}) = H_1H_2 + H_1H_3 + H_3H_2 - 2H_1H_2H_3. \tag{A16}$$

We have used the notation

$$H_l = H \left[\left(\frac{q}{1-q} \right)^{1/2} v_l \right], \tag{A17}$$

where $\mathrm{H(x)}=\int_x^{\infty} Dy$. Dy is a Gaussian measure:

$$Dy = \frac{dy}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2} \right).$$

Differentiating with respect to q, to $\hat{q}$ and to E leads to the saddle-point equations:

$$E = \frac{1}{1-q}, \tag{A18}$$

$$\frac{\hat{q}}{2} = \frac{q}{(1-q)^2}, \tag{A19}$$

$$\hat{q} = \alpha \frac{\partial G_1}{\partial q}. \tag{A20}$$

The third equation is cumbersome, and has been studied in the text for specific cases. At small value of α, q is small. At criticality, when the volume of available solutions shrinks to zero, q reaches 1.

For the case where the number of hidden units is large, $K \to \infty$, the expression of G_1 can have a simpler form. Using the integral representation of the Heaviside function, it is easy to show that

$$G_1 = \int_{-\infty}^{\infty} \prod_l Dv_l \ln \left(\int_{-\infty}^{\infty} dx \frac{1}{2\pi} \int_0^{\infty} d\lambda \exp(i\lambda x) \prod_l (\cos x + iF_l \sin x) \right), \tag{A21}$$

where $F_l = 1 - 2H_l$. In the limit of large K the contributions to the integral on x are coming from $x = O\left(1/\sqrt{K}\right)$. Expanding the integrand in powers of x up to order x^2 and performing the integral on x one obtains

$$G_1 = \int_{-\infty}^{\infty} \prod_l Dv_l \ln H \left(\frac{V}{\sqrt{1-R}} \right), \tag{A22}$$

where

$$V = \frac{1}{\sqrt{K}} \sum_l F_l$$

and

$$R = \frac{1}{K} \sum_l F_l^2.$$

The quantities V and R are a sum of a large number of terms each of which depending on a Gaussian variable v_l. These Gaussian variables are *uncorrelated*. As a consequence a "central limit theorem" applies to V and R. It is then possible to integrate over the v_l leading to the simple expression:

$$G_1 = \int_{-\infty}^{\infty} Dx \ln H \left[\left(\frac{q_{\mathrm{eff}}}{1 - q_{\mathrm{eff}}} \right)^{1/2} x \right], \tag{A23}$$

where

$$q_{\rm eff} = 4\left\langle H^2\left[\left(\frac{q}{1-q}\right)^{1/2} t\right]\right\rangle - 1 = 1 - \frac{2}{\pi}\arccos q \tag{A24}$$

(here the brackets $<>$ denote Gaussian average over t). Note that this expression holds only when $K(1-q)$ is large.

APPENDIX B: INSTABILITY OF THE REPLICA-SYMMETRIC SOLUTION

The Hessian matrix computed at the replica-symmetric saddle point characterizes the fluctuations in the order parameters $q_l^{\alpha,\beta}$, $\hat{q}_l^{\alpha,\beta}$, and E_l^{α} around the RS saddle point. An instability of the RS solution is signaled by a change of sign of at least one of the eigenvalues of this matrix which is of dimension $Kn^2 \times Kn^2$ and can be represented in a block form:

$$\hat{M}_{l,l'} = \hat{U}\delta_{l,l'} + \hat{V}(1-\delta_{l,l'}), \tag{B1}$$

where $\hat{M}_{l,l'}$, $\hat{U}$, and $\hat{V}$ are matrices with dimension $n^2 \times n^2$. $\hat{U}$ and $\hat{V}$ contain the quadratic fluctuations of the order parameters in the same and different hidden unit, respectively. Because of the block form of $\hat{M}$, the eigenproblem splits into the uncoupled diagonalization of the two matrices: $\hat{U}-\hat{V}$ and $\hat{U}+(K-1)\hat{V}$. The eigenvectors of $\hat{U}-\hat{V}$ correspond to fluctuations in the directions that break the permutation symmetry. The eigenvectors of $\hat{U}+(K-1)\hat{V}$ represent fluctuations that do not break this symmetry. The most unstable mode corresponds to an eigenvector of $\hat{U}+(K-1)\hat{V}$ that breaks the replica symmetry. Using the method of Ref. [1], we find that the RS stability criterion is

$$K\alpha(1-q)^2\gamma < 1, \tag{B2}$$

where $\gamma = \gamma_0 + (K-1)\gamma_1$, with

$$\gamma_0 = <(\overline{x_1^2})^2>_v - 2<\overline{x_1^2}(\overline{x_1})^2>_v + <(\overline{x_1})^4>_v \tag{B3}$$

and

$$\gamma_1 = <(\overline{x_1x_2})^2>_v - 2<(\overline{x_1x_2})(\overline{x_1})(\overline{x_2})>_v + <(\overline{x_1})^2(\overline{x_2})^2>_v . \tag{B4}$$

One has defined

$$\overline{x_1^n x_2^m} = \frac{\int\prod_l d\lambda_l \frac{dx_l}{2\pi}\Theta\left(\sum_l \mathrm{sgn}(\lambda_l)\right)\exp\left(-\sum_l\left[\frac{1}{2}(1-q)x_l^2 + ix_l(\lambda_l - \sqrt{q}v_l)\right]\right)x_1^n x_2^m}{\int\prod_l d\lambda_l \frac{dx_l}{2\pi}\Theta\left(\sum_l \mathrm{sgn}(\lambda_l)\right)\exp\left(-\sum_l\left[\frac{1}{2}(1-q)x_l^2 + ix_l(\lambda_l - \sqrt{q}v_l)\right]\right)} \tag{B5}$$

($<>_v$ denotes Gaussian average with measure $\prod_l Dv_l$). All the other eigenvalues, including those of the matrix $\hat{U}-\hat{V}$ change sign only at larger values of α where the RS solution is already not stable.

In the $K=3$ case, γ can be written in the form

$$\gamma = \frac{1}{(1-q)^2}\left(\frac{q}{(1+q)} <(1-X)^2W_1^2>_v + \frac{(1-2q)(1-q)}{(1+2q)(1+q)} <(1-X)^4W_1^4>_v - 2<W_1^2W_2^2X^2Y^2>_v\right), \tag{B6}$$

where we have used notation: $W_l = H_l'/H_l$ with

$$H_l' = -\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{q}{2(1-q)}v_l^2\right)$$

and $X = H_2H_3/\Sigma_{(3)}$, $Y = H_1H_3/\Sigma_{(3)}$ with $\Sigma_{(3)} = H_1H_2 + H_1H_3 + H_3H_2 - 2H_1H_2H_3$.

In the large-K limit ($K \to \infty$), the calculation of the relevant eigenvalue of the Hessian matrix (keeping the dominant contribution in the large-K limit) gives the following instability criterion:

$$1 - \frac{2\alpha\gamma}{\pi}\frac{1-q}{(1-q_{\rm eff})(1+q)} > 0, \tag{B7}$$

where

$$\gamma = \frac{q}{\sqrt{1-q^2}} <W^2> + \frac{2}{\pi(1-q_{\rm eff})}\left[\frac{q_{\rm eff}}{1-q_{\rm eff}} <x^2W^2> + 2\left(\frac{q_{\rm eff}}{1-q_{\rm eff}}\right)^{1/2} <xW^3> + <W^4>\right], \tag{B8}$$

where $<>$ denotes Gaussian average of x and W is defined in Eq. (2.33). The above instability indicates that one has to consider solutions which do not satisfy the symmetry of Eq. (A11). Breaking this symmetry, using the one-step block scheme introduced by Parisi [3] implies the following parametrization of the $q_l^{\alpha,\beta}$ matrix for all l:

$$q_l^{\alpha,\beta} = q_1, \qquad \hat{q}_l^{\alpha,\beta} = \hat{q}_1 \ \text{ for all } I\left(\frac{\alpha}{m}\right) = I\left(\frac{\beta}{m}\right), \tag{B9}$$

$$q_l^{\alpha,\beta} = q_0,$$
$$\hat{q}_l^{\alpha,\beta} = \hat{q}_0 \quad \text{for all } I\left(\frac{\alpha}{m}\right) \neq I\left(\frac{\beta}{m}\right), \tag{B10}$$
$$E_l^{\alpha} = E \quad \text{for all } \alpha,$$

where m is the number of replicas in each block and $I(x)$ is an integer valued function: its value is the smallest integer which is greater than or equal to x.

APPENDIX C: CAPACITY OF THE NONOVERLAPPING COMMITTEE MACHINE

In this appendix the resummation of the most diverging contributions to G_1 [given in Eq. (A14)] in the limit $q \to 1$ is sketched, in order to obtain the maximal capacity of the Nonoverlapping Committee Machine within the RS ansatz.

In the limit $q \to 1$, straightforward combinatorics leads to the following leading behavior:

$$G_1 = \sum_{s=0}^{(K-1)/2} \frac{1}{2^s} \frac{K!}{s!\left(\frac{K+1}{2}-s\right)!\left(\frac{K-3}{2}\right)!} \int_0^\infty Dv_1 \prod_{l=2}^{(K+3)/2-s} \left(\int_0^{v_1} Dv_l\right) \ln\left(H_2 H_3 \cdots H_{(K+3)/2-s}\right) [H(v_1)]^{(K-3)/2}, \tag{C1}$$

where in the discussed limit $\ln(H_l) \to -qv_l^2/2(1-q)$. Performing the sum and all the integrals but the one on v_1 one finds

$$G_1 = -\frac{q}{2(1-q)} \frac{\Gamma(K+1)}{\Gamma\left(\frac{K-1}{2}\right)\Gamma\left(\frac{K+1}{2}\right)} \int_0^\infty Dz[1-H(z)]^{(K-1)/2}[H(z)]^{(K-3)/2}\left[\tfrac{1}{2} - H(z) + zH'(z)\right]. \tag{C2}$$

Using the saddle-point equation

$$-\frac{q}{2(1-q)^2} = \alpha \frac{\partial G_1}{\partial q} \tag{C3}$$

the replica-symmetric maximum capacity is found to be

$$\frac{1}{\alpha_c} = \frac{K-1}{2} \frac{\Gamma(K)}{\left[\Gamma\left(\frac{K+1}{2}\right)\right]^2} \int_0^\infty DzH(z)^{(K-3)/2}[1-H(z)]^{(K-1)/2}\left[\tfrac{1}{2} - H(z) + zH'(z)\right] \tag{C4}$$

where $T(x) = (x-1)!$.

For $K = 1$, one obtains

$$\alpha_c = 2 \tag{C5}$$

which is the capacity of the one-layer perceptron obtained by Cover and Gardner [1].

When K is large a saddle-point estimate of the integral leads to the asymptotic behavior:

$$\alpha_c \sim \left(\frac{72}{\pi}\right)^{1/2} K^{1/2}. \tag{C6}$$

APPENDIX D: BOUND ON THE CAPACITY OF THE COMMITTEE MACHINE

An upper bound for the capacity of the nonoverlapping receptive fields Committee Machine with continuous weights can be obtained by calculating the number of possible partitions of the P random input vectors. The maximal capacity of the network, P_c, will be given by

$$\frac{I_{\text{par}}(N,K,P_c)}{2^{P_c}} = \frac{1}{2}, \tag{D1}$$

where $I_{\text{par}}(N,K,P_c)$ is the average number of partitions of network with N input units, which is subdivided to K nonoverlapping receptive fields. Exact calculation of I_{par} is a hard problem. However, one can find upper bounds for the maximal capacity of the network by approximating I_{par}.

Any set of K hyperplanes (obtained by particular realization of the hidden units synapses), each inducing a dichotomy, induces a partition. But, not all the partitions so obtained are distinct, since it is possible that different sets of dichotomies give rise to the same partition. Thus the number of partitions can be approximated from above by

$$I_{\text{par}}(N,K,P_c) = \left[C\left(P_c, \frac{N}{K}\right)\right]^K, \tag{D2}$$

where $C(P_c, \frac{N}{K})$ is the number of dichotomies of P_c vectors by hyperplane in N/K dimensional space:

$$C\left(P_c, \frac{N}{K}\right) = 2 \sum_{i=0}^{(N/K)-1} \binom{P_c - 1}{i}. \tag{D3}$$

$C(P_c, N/K)$ can be written in an integral form

$$C\left(P_c, \frac{N}{K}\right) = \frac{2^{P_c}(N\alpha - 1)!}{(N\alpha - \frac{N}{K} - 1)!\,(\frac{N}{K} - 1)!} \times \int_0^{\frac{1}{2}} dt\, t^{N[\alpha - (1/K)]-1}(1-t)^{(N/K)-1} \tag{D4}$$

when $\alpha \equiv P_c/N$.

In the thermodynamic limit, $N \to \infty$, one can use the saddle-point method, obtaining

$$C\left(P_c, \frac{N}{K}\right) = \frac{2}{\sqrt{2\pi(N/K)e}}(\alpha K)^{P_c - 1/2} \times(\alpha K - 1)^{-[P_c-(N/K)+(1/2)]} \quad \text{(D5)}$$

Substituting $C(P_c, N/K)$ in the equation defining P_c, using the fact that $P_c \to \infty$ while α is finite, one gets

$$\alpha_* K = (\alpha_* K - 1)^{[1-1/(\alpha_* K)]} 2^{(1/K)}. \quad \text{(D6)}$$

Solving this equation for $K = 3$, gives the upper bound of the capacity to be $\alpha_* = 5.43$. In the $K \to \infty$ case, the previous equation reduces to

$$1 + \ln(\alpha_* K) = \alpha_* \ln 2 \quad \text{(D7)}$$

which gives $\alpha_* = O(\ln K)$.

The same method can be applied to derive an upper bound of the capacity in the fully connected network. In this case the dimension of the hyperplan spanned by the hidden units is N (and not N/K). However, the upper bound for the capacity for $K \to \infty$ behaves also like $\alpha_* = O(\ln K)$.

---

* Permanent address: Centre de Physique Théorique, Ecole Polytechnique, 91128 Palaiseau, France.

[1] E. Gardner, Europhys. Lett. **4**, 481 (1987); J. Phys. A **21**, 257 (1988); E. Gardner and B. Derrida, *ibid.* **21**, 271 (1988).

[2] See, for instance, the memorial volume of Elisabeth Gardner, J. Phys. A **22** (1989); D. Hansel and H. Sompolinsky, Europhys. Lett. **11**, 687 (1990); S. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A (to be published).

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

[4] W. Krauth and M. Mézard, J. Phys. **30**, 3057 (1989).

[5] E. Baum, J. Complex. **4**, 193 (1988).

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

[7] T. Cover, IEEE Trans. Electron. Comput. **14**, 326 (1965).

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

[9] B. Derrida, Phys. Rev. B **24**, 2613 (1981).

[10] D.J. Gross and M. Mezard, Nucl. Phys. **B240**, 431 (1984).

[11] E. Barkai, D. Hansel, and I. Kanter (unpublished).

[12] For concreteness we study in this paper the case of binary inputs. However, all the results for large N apply also to other cases, e.g., real valued inputs with Gaussian distribution.

[13] N.J. Nilsson, *Learning Machines* (McGraw-Hill, New York, 1965).

[14] H. Gutfreund and Y. Stein, J. Phys. A **24**, 2613 (1990).